NUMERICAL LINEAR ALGEBRA

Rolf Rannacher

Institute of Applied Mathematics
Heidelberg University

Lecture Notes WS 2013/2014

February 7, 2014

Address of Author:

Institute of Applied Mathematics
Heidelberg University
Im Neuenheimer Feld 293/294
D-69120 Heidelberg, Germany

rannacher@iwr.uni-heidelberg.de
http://www.numerik.uni-hd.de

Contents

0 Introduction **1**

0.1 Basic notation of Linear Algebra and Analysis 1

0.2 Linear algebraic systems and eigenvalue problems 2

0.3 Numerical approaches 3

0.4 Applications and origin of problems 4

0.4.1 Gaussian equalization calculus 4

0.4.2 Discretization of elliptic PDEs 5

0.4.3 Hydrodynamic stability analysis 9

1 Matrix Analysis: Linear Systems and Eigenvalue Problems **13**

1.1 The normed Euclidean space $\mathbb{K}^n$ 13

1.1.1 Vector norms and scalar products 13

1.1.2 Linear mappings and matrices 21

1.1.3 Non-quadratic linear systems 25

1.1.4 Eigenvalues and eigenvectors 27

1.1.5 Similarity transformations 29

1.1.6 Matrix analysis 31

1.2 Spectra and pseudo-spectra of matrices 35

1.2.1 Stability of dynamical systems 35

1.2.2 Pseudospectrum of a matrix 38

1.3 Perturbation theory and conditioning 43

1.3.1 Conditioning of linear algebraic systems 43

1.3.2 Conditioning of eigenvalue problems 45

1.4 Exercises 48

2 Direct Solution Methods **53**

2.1 Gaussian elimination, LR and Cholesky decomposition 53

2.1.1 Gaussian elimination and LR decomposition 53

2.1.2 Accuracy improvement by final iteration 61

2.1.3 Inverse computation and the Gauß-Jordan algorithm 63

2.2 Special matrices 67

2.2.1 Band matrices 67

2.2.2 Diagonally dominant matrices . . . 69
2.2.3 Positive definite matrices . . . 70
2.3 Irregular linear systems and QR decomposition . . . 72
2.3.1 Householder algorithm . . . 74
2.4 Singular value decomposition . . . 78
2.5 "Direct" determination of eigenvalues . . . 83
2.5.1 Reduction methods . . . 83
2.5.2 Hyman's method . . . 87
2.5.3 Sturm's method . . . 89
2.6 Exercises . . . 91

3 Iterative Methods for Linear Algebraic Systems 93
3.1 Fixed-point iteration and defect correction . . . 93
3.1.1 Stopping criteria . . . 97
3.1.2 Construction of iterative methods . . . 98
3.1.3 Jacobi- and Gauß-Seidel methods . . . 101
3.2 Acceleration methods . . . 104
3.2.1 SOR method . . . 104
3.2.2 Chebyshev acceleration . . . 110
3.3 Descent methods . . . 116
3.3.1 Gradient method . . . 117
3.3.2 Conjugate gradient method (CG method) . . . 121
3.3.3 Generalized CG methods and Krylov space methods . . . 126
3.3.4 Preconditioning (PCG methods) . . . 128
3.4 A model problem . . . 130
3.5 Exercises . . . 134

4 Iterative Methods for Eigenvalue Problems 141
4.1 Methods for the partial eigenvalue problem . . . 141
4.1.1 The "Power Method" . . . 141
4.1.2 The "Inverse Iteration" . . . 143
4.2 Methods for the full eigenvalue problem . . . 146
4.2.1 The LR and QR method . . . 146
4.2.2 Computation of the singular value decomposition . . . 151

4.3 Krylov space methods 152
4.3.1 Lanczos and Arnoldi method 154
4.3.2 Computation of the pseudospectrum 159
4.4 Exercises 169

5 Multigrid Methods **173**
5.1 Multigrid methods for linear systems 173
5.1.1 Multigrid methods in the finite element context 174
5.1.2 Convergence analysis 181
5.2 Multigrid methods for eigenvalue problems (a short review) 187
5.2.1 Direct multigrid approach 188
5.2.2 Accelerated Arnoldi and Lanczos method 189
5.3 Exercises 191

Bibliography **194**

Index **198**

0 Introduction

Subject of this course are numerical algorithms for solving problems in Linear Algebra, such as linear algebraic systems and corresponding matrix eigenvalue problems. The emphasis is on iterative methods suitable for large-scale problems arising, e. g., in the discretization of partial differential equations and in network problems.

0.1 Basic notation of Linear Algebra and Analysis

At first, we introduce some standard notation in the context of (finite dimensional) vector spaces of functions and their derivatives. Let $\mathbb{K}$ denote the field of real or complex numbers $\mathbb{R}$ or $\mathbb{C}$, respectively. Accordingly, for $n \in \mathbb{N}$, let $\mathbb{K}^n$ denote the n-dimensional vector space of n-tuples $x = (x_1, \ldots, x_n)$ with components $x_i \in \mathbb{K}$, $i = 1, \ldots, n$. For these addition and scalar multiplication are defined by:

$$x + y := (x_1 + y_1, \ldots, x_n + y_n), \qquad \alpha x := (\alpha x_1, \ldots, \alpha x_n), \quad \alpha \in \mathbb{K}.$$

The elements $x \in \mathbb{K}^n$ are, depending on the suitable interpretation, addressed as "points" or "vectors" . Here, one may imagine x as the end point of a vector attached at the origin of the chosen cartesian[1] coordinate system and the components x_i as its "coordinates" with respect to this coordinate system. In general, we consider vectors as "column vectors". Within the "vector calculus" its row version is written as $(x_1, \ldots, x_n)^T$. The zero vector $(0, \ldots, 0)$ may also briefly be written as 0. Usually, we prefer this coordinate-oriented notation over a coordinate-free notation because of its greater clearness. A set of vectors $\{a^1, \ldots, a^k\}$ in $\mathbb{K}^n$ is called "linearly independent" if

$$\sum_{i=1}^{k} c_i a^i = 0, \;\; c_i \in \mathbb{K} \quad \Rightarrow \quad c_i = 0, \;\; i = 1, \ldots, k.$$

Such a set of $k = n$ linearly independent vectors is called a "basis" of $\mathbb{K}^n$, which spans all of $\mathbb{K}^n$, i. e., each element $x \in \mathbb{K}^n$ can (uniquely) be written as a linear combination of the form

$$x = \sum_{i=1}^{n} c_i a^i, \;\; c_i \in \mathbb{K}.$$

Each (finite dimensional) vector space, such as $\mathbb{K}^n$, possesses a basis. The special "cartesian basis" $\{e^1, \ldots, e^n\}$ is formed by the "cartesian unit vectors" $e^i := (\delta_{1i}, \ldots, \delta_{ni})$, $\delta_{ii} = 1$ and $\delta_{ij} = 0$, for $i \neq j$, being the usual Kronecker symbol. The elements of this basis are mutually orthonormal, i. e., with respect to the euclidian scalar product, there holds $(e^i, e^j)_2 := \sum_{k=1}^{n} e^i_k e^j_k = \delta_{ij}$. "Matrices" $A \in \mathbb{K}^{n \times n}$ are two-dimensional square arrays of numbers from $\mathbb{K}$ written in the form $A = (a_{ij})_{i,j=1}^{n}$, where the first index, i, refers to the row and the second one, j, to the column (counted from the left upper corner of the array) at which the element a_{ij} is positioned. Usually, matrices are *square* arrays, but in some situations also rectangular matrices may occur. The set of (square) matrices forms a vector space with addition

[1] René Descartes (1596-1650): French mathematician and philosopher ("cogito ergo sum"); worked in the Netherlands and later in Stockholm; first to recognize the close relation between geometry and arithmetic and founded analytic geometry.

and scalar multiplication defined in the natural elementwise sense,

$$A = (a_{ij})_{i,j=1}^n,\ B = (b_{ij})_{i,j=1}^n,\ c \in \mathbb{K} \quad \Rightarrow \quad cA + B = (ca_{ij} + b_{ij})_{i,j=1}^n.$$

For matrices and vectors natural multiplications are defined by

$$Ax = \Big(\sum_{k=1}^d a_{ik} x_k \Big)_{i=1}^n \in \mathbb{K}^n, \quad AB = \Big(\sum_{k=1}^d a_{ik} b_{kj} \Big)_{i,j=1}^n \in \mathbb{K}^{n \times n}.$$

Matrices are used to represent linear mappings in $\mathbb{K}^d$ with respect to a given basis, mostly a cartesian basis, $\varphi(x) = Ax$. By $\bar{A}^T = (a_{ij}^T)_{i,j=1}^n$, we denote the conjugate "transpose" of a matrix $A = (a_{ij})_{i,j=1}^n \in \mathbb{K}^{n \times n}$ with the elements $a_{ij}^T = \bar{a}_{ji}$. For matrices $A, B \in \mathbb{K}^{n \times n}$ there holds $(AB)^T = B^T A^T$. Matrices for which $A = \bar{A}^T$ are called "symmetric" in the case $\mathbb{K} = \mathbb{R}$ and "hermitian" in the case $\mathbb{K} = \mathbb{C}$.

Further, we have to deal with scalar or vector-valued functions $u = u(x) \in \mathbb{K}^n$ for arguments $x \in \mathbb{K}^n$. For derivatives of differentiable functions, we use the notation

$$\partial_x u := \frac{\partial u}{\partial x},\ \partial_x^2 u := \frac{\partial^2 u}{\partial^2 x},\ \dots\ , \quad \partial_i u := \frac{\partial u}{\partial x_i},\ \partial_{ij}^2 u := \frac{\partial^2 u}{\partial x_i \partial x_j},\ \dots\ ,$$

and analogously also for higher-order derivatives. With the nabla operator ∇ the "gradient" of a scalar function and the "divergence" of a vector function are written as $\operatorname{grad} u = \nabla u := (\partial_1 u, ..., \partial_d u)$ and $\operatorname{div} u = \nabla \cdot u := \partial_1 u_1 + ... + \partial_d u_d$, respectively. For a vector $\beta \in \mathbb{R}^d$ the derivative in direction β is written as $\partial_\beta u := \beta \cdot \nabla u$. Combination of gradient and divergence yields the "Laplacian[2] -Operator"

$$\nabla \cdot \nabla u = \Delta u = \partial_1^2 u + ... + \partial_d^2 u.$$

The symbol $\nabla^m u$ denotes the "tensor" of all partial derivatives of order m of u, i. e., in two dimensions $u = u(x_1, x_2)$, $\nabla^2 u = (\partial_1^i \partial_2^j u)_{i+j=2}$.

0.2 Linear algebraic systems and eigenvalue problems

Let A be an $m \times n$-matrix and b an m-vector,

$$A = (a_{jk})_{j,k=1}^{m,n} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}, \qquad b = (b_j)_{j=1}^m = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

[2] Pierre Simon Marquis de Laplace (1749-1827): French mathematician and astronomer; prof. in Paris; founded among other fields probability calculus.

We seek an n-vector $x = (x_k)_{k=1,\dots,n}$ such that

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{0.2.1}$$

or written in short as $Ax = b$. This is called a "linear system" (of equations). It is called "underdetermined" for $m < n$, "quadratic" for $m = n$, and "overdetermined" for $m > n$. The linear system is solvable if and only if $\operatorname{rank}(A) = \operatorname{rank}[A, b]$ (rank(A) = number of linearly independent columns of A) with the composed matrix

$$[A, b] = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right].$$

In the "quadratic" case the solvability of the system (0.2.1) is equivalent to any one of the following properties of the coefficient matrix $A \in \mathbb{K}^{n\times n}$:

- $Ax = 0$ implies $x = 0$.
- $\operatorname{rank}(A) = n$.
- $\det(A) \neq 0$.
- All eigenvalues of A are nonzero.

A number $\lambda \in \mathbb{C}$ is called "eigenvalue" of the (quadratic) matrix $A \in \mathbb{K}^{n\times n}$ if there exists a corresponding vector $w \in \mathbb{K}^n \setminus \{0\}$, called "eigenvector", such that

$$Aw = \lambda w. \tag{0.2.2}$$

Eigenvalues are just the zeros of the characteristic polynomial $\chi_A(z) := \det(A - zI)$ of A, so that by the fundamental theorem of Algebra each $n \times n$-matrix has exactly n eigenvalues counted accordingly to their (algebraic) multiplicities. The corresponding eigenvectors span linear subspaces of $\mathbb{K}^n$ called "eigenspaces". Eigenvalue problems play an important role in many problems from science and engineering, e. g., they represent energy levels in physical models (e. g., Schrödinger equation in Quantum Mechanics) or determine the stability or instability of dynamical systems (e. g., Navier-Stokes equations in hydrodynamics).

0.3 Numerical approaches

We will mainly consider numerical methods for solving *quadratic* linear systems and associated eigenvalue problems. The emphasis will be on medium- and large-scale problems, i. e., problems of dimension $n \approx 10^4 - 10^9$, which at the upper end pose particularly hard requirements on the algorithms with respect to storage and work efficiency. Problems of that size usually involve matrices with special structure such as "band structure" and/or extreme "sparsity", i. e., only very few matrix elements in each row are non-zero. Most of the classical methods, which have originally been designed for "full" but smaller matrices, cannot realistically be applied to such large problems. Therefore, modern methods extensively exploit the particular sparsity structure of the matrices. These methods split into two classes, "direct methods" and "iterative methods".

Definition 0.1: *A "direct" method for the solution of a linear system $Ax = b$ is an algorithm, which (neglecting round-off errors) delivers the exact solution x in finitely many arithmetic operations. "Gaussian elimination" is a typical example of such a "direct method". In contrast to that an "iterative method" constructs a sequence of approximate solutions $\{x^t\}_{t\in\mathbb{N}}$, which only in the limit $t \to \infty$ converge to the exact solution, i. e., $\lim_{t\to\infty} x^t = x$. "Richardson iteration" or more general fixed-point methods of similar kind are typical example of such "iterative methods". In analyzing a direct method, we are mainly interested in the work count, i. e., the asymptotic number of arithmetic operations needed for achieving the final result depending on the problem size, e. g., $\mathcal{O}(n^3)$, while in an iterative method, we look at the work count needed for one iteration step and the number of iteration steps for reducing the initial error by a certain fixed factor, e. g., 10^{-1}, or the asymptotic speed of convergence ("linear", "quadratic, etc.).*

However, there is no sharp separation between the two classes of "direct" or "iterative" methods as many theoretically "direct" methods are actually used in "iterative" form in practice. A typical method of this type is the classical "conjugate gradient (CG) method" of Hestenes and Stiefel, which in principle is a direct method (after n iteration steps) but is usually terminated like an iterative methods already after $m \ll n$ steps.

0.4 Applications and origin of problems

We present some applications, from which large linear algebra problems originate. This illustrates how the various possible structures of matrices may look like.

0.4.1 Gaussian equalization calculus

A classical application in Astronomy is the Gaussian equalization calculus (method of least error-squares): For given functions $u^1, \ldots, u^n$ and points $(x_j, y_j) \in \mathbb{R}^2$, $j = 1, \ldots, m$, $m > n$, a linear combination

$$u(x) = \sum_{k=1}^{n} c_k u^k(x)$$

is to be determined such that the "mean deviation"

$$\Delta_2 := \Big(\sum_{j=1}^{m} |u(x_j) - y_j|^2 \Big)^{1/2}$$

becomes minimal. (The so-called "Chebyshev[3] equalization problem" in which the "maximal deviation" $\Delta_\infty := \max_{j=1,\ldots,m} |u(x_j) - y_j|$ is minimized poses much more severe difficulties and is therefore used only for smaller n.) For the solution of the Gaussian equalization problem, we set

$$\begin{aligned} y &:= (y_1, \ldots, y_m), \qquad c \equiv (c_1, \ldots, c_n), \\ a^k &:= (u^k(x_1), \ldots, u^k(x_m)), \ k = 1, \ldots, n, \qquad A \equiv [a^1, \ldots, a^n]. \end{aligned}$$

[3] Pafnuty Lvovich Chebyshev (1821-1894): Russian mathematician; prof. in St. Petersburg; contributions to number theory, probability theory and especially to approximation theory; developed the general theory of orthogonal polynomials.

Using this notation, now the quadratic functional

$$F(c) = \Big(\sum_{j=1}^{m} |(Ac - y)_j|^2 \Big)^{1/2}$$

is to be minimized with respect to $c \in \mathbb{R}^n$. This is equivalent to solving the overdetermined linear system $Ac = y$ in the sense of finding a vector c with minimal mean error-squares, i. e., with minimal "defect". In case that $\operatorname{rank}(A) = n$ this "minimal-defect solution" c is determined by the so-called "normal equation"

$$A^T A c = A^T y, \tag{0.4.3}$$

a quadratic linear $n \times n$-system with a positive definite (and hence regular) coefficient matrix $A^T A$. In the particular case of polynomial fitting, i. e., $u^k(x) = x^{k-1}$, the "optimal" solution

$$u(x) = \sum_{k=1}^{n} c_k x^{k-1}$$

is called "Gaussian equalization parabola" for the points (x_j, y_j), $j = 1, \ldots, m$. Because of the regularity of the "Vandermondian[4] determinant"

$$\det \begin{bmatrix} 1 & x_1 & \cdots & x_1^{n-1} \\ 1 & x_2 & \cdots & x_2^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^{n-1} \end{bmatrix} = \prod_{j,k=1, j<k}^{n} (x_k - x_j) \neq 0,$$

for mutually distinct points x_j there holds $\operatorname{rank}(A) = n$, i. e., the equalization parabola is uniquely determined.

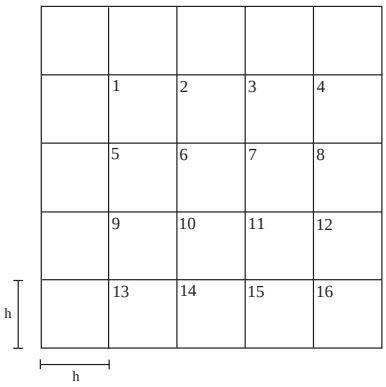
0.4.2 Discretization of elliptic PDEs

The numerical solution of partial differential equations requires first an appropriate "discretization" of the differential operator, e. g., by a "difference approximation" of the derivatives. Consider, for example, the so-called "first boundary value problem of the Laplacian operator",

$$Lu := -\Delta u = f \quad \text{in } \Omega, \quad u = g \text{ on } \partial\Omega, \tag{0.4.4}$$

posed on a domain $\Omega \subset \mathbb{R}^n$ with boundary $\partial\Omega$. Here, for a given right-hand side function $f = f(x_1, x_2)$ and boundary function $g = g(x_1, x_2)$, assumed to be continuous, an on Ω twice differentiable and on $\bar{\Omega}$ continuous solution function $u = u(x_1, x_2)$ is to be determined such that (0.4.4) holds. The region $\bar{\Omega}$, e. g., the unit square, is covered by an equidistant cartesian mesh Ω_h with "mesh boundary" $\partial\Omega_h$ and the mesh points $P \in \Omega_h$ may be numbered row-wise.

[4] Alexandre-Thophile Vandermonde (1735-1796): French mathematician; gifted musician, came late to mathematics and published here only four papers (nevertheless member of the Academy of Sciences in Paris); contributions to theory of determinants and combinatorial problem (curiously enough the determinant called after him does not appear explicitly in his papers).



$h = \frac{1}{m+1}$ mesh width

$n = m^2$ "inner" mesh points

Figure 1: Finite difference mesh

At "interior" mesh points $P \in \Omega_h$ the differential operators in x_1- and x_2-direction are approximated by second-order central difference quotients, which act on mesh functions $u_h(P)$. This results in "difference equations" of the form

$$L_h u_h(P) := \sum_{Q \in N(P)} \sigma(P,Q) u_h(Q) = f_h(P), \quad P \in \Omega_h, \tag{0.4.5}$$

$$u_h(P) = g_h(P), \quad P \in \partial\Omega_h, \tag{0.4.6}$$

with certain mesh neighborhoods $N(P) \subset \Omega \cup \partial\Omega_h$ of points $P \in \Omega_h$ and approximations $f_h(\cdot)$ to f and $g_h(\cdot)$ to g. We set the coefficients $\sigma(P,Q) := 0$ for points $Q \notin N(P)$. The considered difference operator based on second-order central difference quotients for approximating second derivatives is called "5-point difference operator" since it uses 5 points (Accordingly, its three-dimensional analogue is called "7-point difference operator"). Then, for $P \in \Omega_h$ there holds

$$\sum_{Q \in \Omega_h} \sigma(P,Q) u_h(Q) = f_h(P) - \sum_{Q \in \partial\Omega_h} \sigma(P,Q) g_h(Q). \tag{0.4.7}$$

For any numbering of the mesh points in Ω_h and $\partial\Omega_h$, $\Omega_h = \{P_i,\, i = 1, ..., n\}$, $\partial\Omega_h = \{P_i,\, i = n+1, ..., n+m\}$, we obtain a quadratic linear system for the vector of approximate mesh values $U = (U_i)_{i=1}^N$, $U_i := u_h(P_i)$.

$$AU = F, \tag{0.4.8}$$

with $A = (a_{ij})_{i,j=1}^n$, $F = (b_j)_{j=1}^n$, where

$$a_{ij} := \sigma(P_i, P_j), \quad b_j := f_h(P_j) - \sum_{i=n+1}^{n+m} \sigma(P_j, P_i) g_h(P_i).$$

In the considered special case of the unit square and row-wise numbering of the interior mesh points Ω_h the 5-point difference approximation of the Laplacian yields the following sparse matrix of dimension $n = m^2$:

$$A = \frac{1}{h^2}\begin{bmatrix} B_m & -I_m & & \\ -I_m & B_m & -I_m & \\ & -I_m & B_m & \ddots \\ & & \ddots & \ddots \end{bmatrix}\Bigg\}n \qquad B_m = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\Bigg\}m\,,$$

where I_m is the $m\times m$-unit matrix. The matrix A is a very sparse band matrix with half-band width m, symmetric and (irreducibly) diagonally dominant. This implies that it is regular and positive definite. In three dimensions the corresponding matrix has dimension $n = m^3$ and half-band width m^2 and shares all the other mentioned properties of its two-dimensional analogue. In practice, $n \gg 10^4$ up to $n \approx 10^7$ in three dimensions. If problem (0.4.4) is only part of a larger mathematical model involving complex domains and several physical quantities such as (in chemically reacting flow models) velocity, pressure, density, temperature and chemical species, the dimension of the complete system may quickly reach up to $n \approx 10^7 - 10^9$.

To estimate a realistic size of the algebraic problem oriented by the needs of a practical application, we consider the above model problem (Poisson equation on the unit square) with an adjusted right-hand side and boundary function such that the exact solution is given by $u(x, y) = \sin(\pi x)\sin(\pi y)$,

$$-\Delta u = 2\pi^2 u =: f \ \text{ in } \ \Omega, \quad u = 0 \ \text{ on } \ \partial\Omega. \tag{0.4.9}$$

For this setting the error analysis of the difference approximation yields the a priori estimate

$$\max_{\overline{\Omega}_h} |u - u_h| \approx \tfrac{1}{24} d_\Omega^2 M_4(u) h^2 \approx 8h^2, \tag{0.4.10}$$

where $M_4(u) = \max_{\bar{\Omega}} |\nabla^4 u| \approx \pi^4$ (see the lecture notes Rannacher [3]). In order to guarantee a relative error below $\mathrm{TOL} = 10^{-3}$, we have to choose $h \approx 10^{-2}$ corresponding to $n \approx 10^4$ in two and $n \approx 10^6$ in three dimension. The concrete structure of the matrix A depends on the numbering of mesh points used:

(i) Row-wise numbering: The lexicographical ordering of mesh points, $(x_i, y_j) \le (x_p, y_q)$ for $j \le q$ or $j = q$, $i \le p$, leads to a band matrix with band width $2m+1 \approx h^{-1}$. The sparsity within the band would be largely reduced by Gaussian elimination (so-called "fill-in")

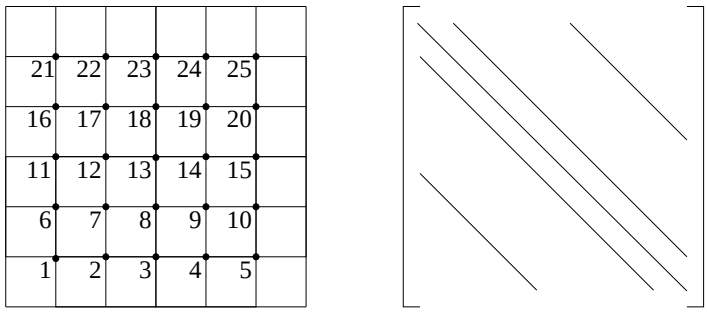


Figure 2: Lexicographical ordering of mesh points

(ii) Diagonal numbering: The successive numbering diagonally to the cartesian coordinate directions leads to a band matrix with less band volume. This results in less fill-in within Gaussian elimination.

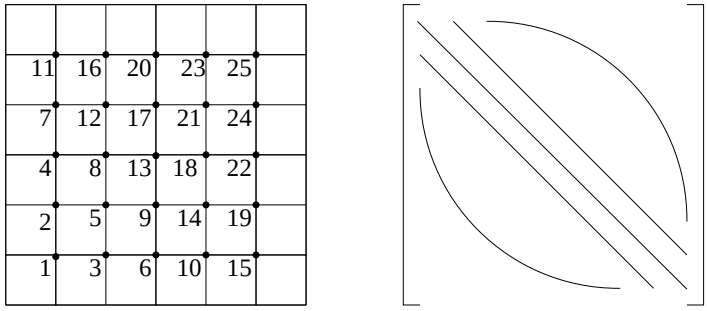


Figure 3: Diagonal mesh-point numbering

(iii) Checker-board numbering: The staggered row-wise and column-wise numbering leads to a 2×2-block matrix with diagonal main blocks and band width $2m + 1 \approx h^{-1}$.

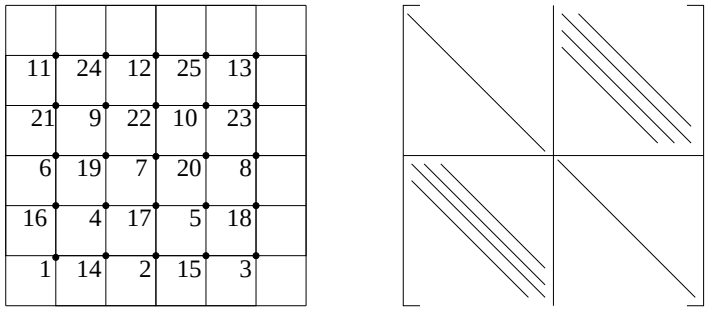


Figure 4: Checkerboard mesh-point numbering

For large linear systems of dimension $n \gg 10^5$ direct methods such as Gaussian elimination are difficult to realize since they are generally very storage and work demanding. For a matrix of dimension $n = 10^6$ and band width $m = 10^2$ Gaussian elimination requires already about 10^8 storage places. This is particularly undesirable if also the band is sparse as in the above example with at most 5 non-zero elements per row. In this case those *iterative* methods are more attractive, in which essentially only matrix-vector multiplications occur with matrices of similar sparsity pattern as that of A.

As illustrative examples, we consider simple fixed-point iterations for solving a linear system $Ax = b$ with a regular $n \times n$-coefficient matrix. The linear system is rewritten in the form

$$a_{jj}x_j + \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk}x_k = b_j, \quad j = 1, \ldots, n.$$

If $a_{jj} \neq 0$, this is equivalent to

$$x_j = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} x_k \Big\}, \quad j = 1, \ldots, n.$$

Then, the so-called "Jacobi method" generates iterates $x^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$, by successively solving

$$x_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} x_k^{t-1} \Big\}, \quad j = 1, \ldots, n. \tag{0.4.11}$$

When computing x_j^t the preceding components $x_r^t, r < j$, are already known. Hence, in order to accelerate the convergence of the method, one may use this new information in the computation of x_j^t. This idea leads to the "Gauß-Seidel[5] method":

$$x_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{k<j} a_{jk} x_k^t - \sum_{k>j} a_{jk} x_k^{t-1} \Big\}, \quad j = 1, \ldots, n. \tag{0.4.12}$$

The Gauß-Seidel method has the same arithmetic complexity as the Jacobi method but under certain conditions (satisfied in the above model situation) it converges twice as fast. However, though very simple and maximal storage economical both methods, Jacobi as well as Gauß-Seidel, are by far too slow in practical applications. Much more efficient iterative methods are the Krylov-space methods. The best known examples are the classical "conjugate gradient method" ("CG method") of Hestenes and Stiefel for solving linear systems with positive definite matrices and the "Arnoldi method" for solving corresponding eigenvalue problems. Iterative methods with minimal complexity can be constructed using multi-scale concepts (e. g., geometric or algebraic "multigrid methods"). The latter type of methods will be discussed below.

0.4.3 Hydrodynamic stability analysis

Another origin of large-scale eigenvalue problems is hydrodynamic stability analysis. Let $\{\hat{v}, \hat{p}\}$ be a solution (the "base flow") of the stationary Navier-Stokes equation

$$\begin{aligned} &-\nu \Delta \hat{v} + \hat{v} \cdot \nabla \hat{v} + \nabla \hat{p} = 0, \quad \nabla \cdot \hat{v} = 0, \quad \text{in } \Omega, \\ &\hat{v}|_{\Gamma_{\text{rigid}}} = 0, \;\; \hat{v}|_{\Gamma_{\text{in}}} = v^{\text{in}}, \;\; \nu \partial_n \hat{v} - \hat{p} n|_{\Gamma_{\text{out}}} = P, \;\; \hat{v}|_{\Gamma_Q} = q, \end{aligned} \tag{0.4.13}$$

where $\hat{v}$ is the velocity vector field of the flow, $\hat{p}$ its hydrostatic pressure, ν the kinematic viscosity (for normalized density $\rho \equiv 1$), and q the control pressure. The flow is driven by a prescribed flow velocity v^{in} at the Dirichlet (inflow) boundary (at the left end), a prescribed mean pressure P at the Neumann (outflow) boundary (at the right end) and the control pressure at the control boundary Γ_Q. The (artificial) "free outflow" (also called "do nothing") boundary condition in (0.4.13) has proven successful especially in modeling pipe flow since it is satisfied by Poiseuille flow (see Heywood et al. [39]).

[5] Philipp Ludwig von Seidel (1821-1896): German mathematician; prof. in Munich; contributions to analysis (method of least error-squares) and celestial mechanics and astronomy.

Figure 5: Configuration of the flow control problem.

Figure 5 shows the configuration of a channel flow around an obstacle controlled by pressure prescription at Γ_Q, and Figure 6 the computational mesh and streamline plots of two flows for different Reynolds numbers and control values, one stable and one unstable.

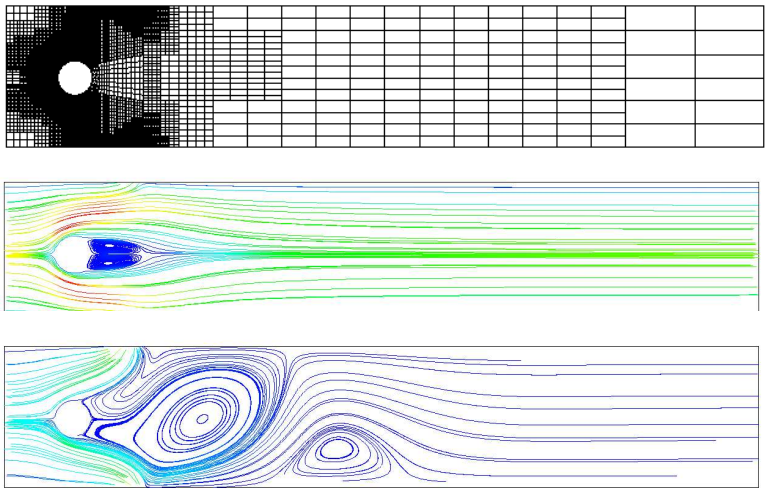


Figure 6: Computational mesh (top), uncontrolled stable (middle) and controlled unstable (bottom) stationary channel flow around an obstacle.

For deciding whether these base flows are stable or unstable, within the usual linearized stability analysis, one investigates the following eigenvalue problem corresponding to the Navier-Stokes operator linearized about the considered base flow:

$$\begin{aligned} &-\nu\Delta v + \hat{v}\cdot\nabla v + v\cdot\nabla\hat{v} + \nabla q = \lambda v, \quad \nabla\cdot v = 0, \quad \text{in } \Omega, \\ &v|_{\Gamma_{\text{rigid}}\cup\Gamma_{\text{in}}} = 0, \quad \nu\partial_n v - qn|_{\Gamma_{\text{out}}} = 0. \end{aligned} \tag{0.4.14}$$

From the location of the eigenvalues in the complex plane, one can draw the following conclusion: If an eigenvalue $\lambda \in \mathbb{C}$ of (0.4.14) has $\text{Re}\,\lambda < 0$, the base flow is unstable, otherwise it is said to be "linearly stable". This means that the solution of the linearized nonstationary perturbation problem

$$\begin{aligned} &\partial_t w - \nu\Delta w + \hat{v}\cdot\nabla w + w\cdot\nabla\hat{v} + \nabla q = 0, \quad \nabla\cdot w = 0, \quad \text{in } \Omega, \\ &w|_{\Gamma_{\text{rigid}}\cup\Gamma_{\text{in}}} = 0, \quad \nu\partial_n w - qn|_{\Gamma_{\text{out}}} = 0 \end{aligned} \tag{0.4.15}$$

corresponding to an initial perturbation $w_{|t=0} = w_0$ satisfies a bound

$$\sup_{t\geq 0} \|w(t)\| \leq A\|w_0\|, \tag{0.4.16}$$

with some constant $A \geq 1$. After discretization the eigenvalue problem (0.4.14) in function space is translated into an essentially nonsymmetric algebraic eigenvalue problem, which is usually of high dimension $n \approx 10^5 - 10^6$. Therefore its solution can only be achieved by iterative methods.

However, "linear stability" does not guarantee full "nonlinear stability" due to effects caused by the "non-normality" of the operator governing problem (0.4.14), which may cause the constant A to become large. This is related to the possible "deficiency" (discrepancy of geometric and algebraic multiplicity) or a large "pseudospectrum" (range of large resolvent norm) of the critical eigenvalue. This effect is commonly accepted as explanation of the discrepancy in the stability properties of simple base flows such as Couette flow and Poiseuille flow predicted by linear eigenvalue-based stability analysis and experimental observation (see, e.g., Trefethen & Embree [19] and the literature cited therein).

1 Matrix Analysis: Linear Systems and Eigenvalue Problems

In this chapter, we introduce the basic notation and facts about the normed real or complex vector spaces $\mathbb{K}^n$ of n-dimensional vectors and $\mathbb{K}^{n\times n}$ of corresponding $n \times n$-matrices. The emphasis is on square matrices as special representations of linear mappings in $\mathbb{K}^n$ and their spectral properties.

1.1 The normed Euclidean space $\mathbb{K}^n$

1.1.1 Vector norms and scalar products

We recall some basic topological properties of the finite dimensional "normed" (vector) space $\mathbb{K}^n$, where depending on the concrete situation $\mathbb{K} = \mathbb{R}$ (real space) or $\mathbb{K} = \mathbb{C}$ (complex space). In the following each point $x \in \mathbb{K}^n$ is expressed by its canonical coordinate representation $x = (x_1, \ldots, x_n)$ in terms of a (fixed) cartesian basis $\{e^1, \ldots, e^n\}$ of $\mathbb{K}^n$,

$$x = \sum_{i=1}^{n} x_i e^i.$$

Definition 1.1: *A mapping $\|\cdot\| : \mathbb{K}^n \to \mathbb{R}$ is a "(vector) norm" if it has the following properties:*

(N1) Definiteness: $\|x\| \geq 0, \quad \|x\| = 0 \;\Rightarrow\; x = 0, \quad x \in \mathbb{K}^n.$

(N2) Homogeneity: $\|\alpha x\| = |\alpha|\,\|x\|, \quad \alpha \in \mathbb{K},\ x \in \mathbb{K}^n.$

(N3) Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|, \quad x, y \in \mathbb{K}^n.$

The notion of a "norm" can be defined on any vector space V over $\mathbb{K}$, finite or infinite dimensional. The resulting pair $\{V, \|\cdot\|\}$ is called "normed space".

Remark 1.1: The property $\|x\| \geq 0$ is a consequence of the other conditions. With (N2), we obtain $0 = \|0\|$ and then with (N3) and (N2) $0 = \|x - x\| \leq \|x\| + \|-x\| = 2\|x\|$. With the help of (N3) we obtain the useful inequality

$$\big|\|x\| - \|y\|\big| \leq \|x - y\|, \quad x, y \in \mathbb{K}^n. \tag{1.1.1}$$

Example 1.1: The standard example of a vector norm is the "euclidian norm"

$$\|x\|_2 := \Big(\sum_{i=1}^{n} |x_i|^2\Big)^{1/2}.$$

The first two norm properties, (N1) and (N2), are obvious, while the triangle inequality is a special case of the "Minkowski inequality" provided below in Lemma 1.4. Other examples of useful norms are the "maximum norm" (or "l_∞ norm") and the "l_1 norm"

$$\|x\|_\infty := \max_{i=1,\ldots,n} |x_i|, \qquad \|x\|_1 := \sum_{i=1}^{n} |x_i|.$$

The norm properties of $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are immediate consequences of the corresponding properties of the modulus function. Between l_1 norm and maximum norm there are the so-called "l_p norms" for $1 < p < \infty$:

$$\|x\|_p := \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p}.$$

Again the first two norm properties, (N1) and (N2), are obvious and the triangle inequality is the Minkowski inequality provided in Lemma 1.4, below.

With the aid of a norm $\|\cdot\|$ the "distance" $d(x, x') := \|x - x'\|$ of two vectors in $\mathbb{K}^n$ is defined. This allows the definition of the usual topological terms "open", "closed", "compact", "diameter", and "neighborhood" for point sets in $\mathbb{K}^n$ in analogy to the corresponding situation in $\mathbb{K}$. We use the maximum norm $\|\cdot\|_\infty$ in the following discussion, but we will see later that this is independent of the chosen norm. For any $a \in \mathbb{K}^n$ and $r > 0$, we use the ball

$$K_r(a) := \{x \in \mathbb{K}^n : \|x - a\|_\infty < r\}$$

as standard neighborhood of a with radius r. This neighborhood is "open" since for each point $x \in K_r(a)$ there exists a neighborhood $K_\delta(x) \subset K_r(a)$; accordingly the complement $K_r(a)^c$ is "closed". The "closure" of $K_r(a)$ is defined by $\overline{K}_r(a) := K_r(a) \cup \partial K_r(a)$ with the "boundary" $\partial K_r(a)$ of $K_r(a)$.

Definition 1.2: *A sequence of vectors $(x^k)_{k\in\mathbb{N}}$ in $\mathbb{K}^n$ is called*

- *"bounded" if all its elements are contained in a ball $K_R(0)$, i. e., $\|x^k\|_\infty \le R$, $k \in \mathbb{N}$,*
- *"Cauchy sequence" if for each $\varepsilon \in \mathbb{R}_+$ there is an $N_\varepsilon \in \mathbb{N}$ such that $\|x^k - x^l\|_\infty < \varepsilon$, $k, l \ge N_\varepsilon$,*
- *"convergent" towards an $x \in \mathbb{K}^n$ if $\|x^k - x\|_\infty \to 0 \quad (k \to \infty)$.*

For a convergent sequence $(x^k)_{k\in\mathbb{N}}$, we also write $\lim_{k\to\infty} x^k = x$ or $x^k \to x$ $(k \to \infty)$. Geometrically this means that any standard neighborhood $K_\varepsilon(x)$ of x contains almost all (i. e., all but finitely many) elements x^k. This notion of "convergence" is obviously equivalent to the componentwise convergence:

$$\|x^k - x\|_\infty \to 0 \ \ (k \to \infty) \quad \Leftrightarrow \quad x_i^k \to x_i \ \ (k \to \infty),\ i = 1, \ldots, n.$$

This allows the reduction of the convergence of sequences of vectors in $\mathbb{K}^n$ to that of sequences of numbers in $\mathbb{K}$. As basic results, we obtain n-dimensional versions of the Cauchy criterion for convergence and the theorem of Bolzano-Weierstraß.

Theorem 1.1 (Theorems of Cauchy and Bolzano-Weierstraß):

(i) Each Cauchy sequence in $\mathbb{K}^n$ is convergent, i. e., the normed space $(\mathbb{K}^n, \|\cdot\|_\infty)$ is complete (a so-called "Banach space").

(ii) Each bounded sequence in $\mathbb{K}^n$ contains a convergent subsequence.

Proof. (i) For any Cauchy sequence $(x^k)_{k\in\mathbb{N}}$, in view of $|x_i| \le \|x\|_\infty$, $i = 1, \ldots, n$, for $x \in \mathbb{K}^n$, also the component sequences $(x_i^k)_{k\in\mathbb{N}}$, $i = 1, \ldots, n$, are Cauchy sequences in $\mathbb{K}$ and therefore

converge to limits $x_i \in \mathbb{K}$. Then, the vector $x := (x_1, \ldots, x_n) \in \mathbb{K}^n$ is limit of the vector sequence $(x^k)_{k\in\mathbb{N}}$ with respect to the maximum norm.

(ii) For any bounded vector sequence $(x^k)_{k\in\mathbb{N}}$ the component sequences $(x_i^k)_{k\in\mathbb{N}}$, $i = 1, \ldots, n$, are likewise bounded. By successively applying the theorem of Bolzano-Weierstraß in $\mathbb{K}$, in the first step, we obtain a convergent subsequence $(x_1^{k_{1j}})_{j\in\mathbb{N}}$ of $(x_1^k)_{k\in\mathbb{N}}$ with $x_1^{k_{1j}} \to x_1$ $(j \to \infty)$, in the next step a convergent subsequence $(x_2^{k_{2j}})_{j\in\mathbb{N}}$ of $(x_2^{k_{1j}})_{j\in\mathbb{N}}$ with $x_2^{k_{2j}} \to x_2$ $(j \to \infty)$, and so on. After n selection steps, we eventually obtain a subsequence $(x^{k_{nj}})_{j\in\mathbb{N}}$ of $(x^k)_{k\in\mathbb{N}}$, for which all component sequences $(x_i^{k_{nj}})_{j\in\mathbb{N}}$, $i = 1, \ldots, n$, converge. Then, with the limit values $x_i \in \mathbb{K}$, we set $x := (x_1, \ldots, x_n) \in \mathbb{K}^n$ and have the convergence $x^{k_{nj}} \to x$ $(j \to \infty)$. Q.E.D.

The following important result states that on the (finite dimensional) vector space $\mathbb{K}^n$ the notion of convergence, induced by any norm $\|\cdot\|$, is equivalent to the convergence with respect to the maximum norm, i. e., to the componentwise convergence.

Theorem 1.2 (Equivalence of norms): *All norms on the finite dimensional vector space $\mathbb{K}^n$ are equivalent to the maximum norm, i. e., for each norm $\|\cdot\|$ there are positive constants m, M such that*

$$m\,\|x\|_\infty \le \|x\| \le M\,\|x\|_\infty\,, \quad x \in \mathbb{K}^n\,. \tag{1.1.2}$$

Proof. Let $\|\cdot\|$ be an arbitrary vector norm. For any vector $x = \sum_{i=11}^n x_i e^i \in \mathbb{K}^n$ there holds

$$\|x\| \le \sum_{k=1}^n |x_k|\,\|e^k\| \le M\,\|x\|_\infty\,, \quad M := \sum_{k=1}^n \|e^k\|\,.$$

We set

$$S_1 := \{x \in \mathbb{K}^n :\ \|x\|_\infty = 1\}, \qquad m := \inf\{\|x\|,\ x \in S_1\} \ge 0.$$

We want to show that $m > 0$ since then, in view of $\|x\|_\infty^{-1} x \in S_1$, it follows that also $m \le \|x\|_\infty^{-1}\|x\|$ for $x \ne 0$, and consequently,

$$0 < m\|x\|_\infty \le \|x\|, \quad x \in \mathbb{K}^n.$$

Suppose that $m = 0$. Then, there is a sequence $(x^k)_{k\in\mathbb{N}}$ in S_1 such that $\|x^k\| \to 0$ $(k \to \infty)$. Since this sequence is bounded in the maximum norm, by the theorem of Bolzano-Weierstrass it possesses a subsequence, likewise denoted by x^k, which converges in the maximum norm to some $x \in \mathbb{K}^n$. Since

$$\big|1 - \|x\|_\infty\big| = \big|\|x^k\|_\infty - \|x\|_\infty\big| \le \|x^k - x\|_\infty \to 0 \quad (k \to \infty),$$

we have $x \in S_1$. On the other hand, for all $k \in \mathbb{N}$, there holds

$$\|x\| \le \|x - x^k\| + \|x^k\| \le M\|x - x^k\|_\infty + \|x^k\|.$$

This implies for $k \to \infty$ that $\|x\| = 0$ and therefore $x = 0$, which contradicts $x \in S_1$. Q.E.D.

Remark 1.2: (i) For the two foregoing theorems, the theorem of Bolzano-Weierstrass and the theorem of norm equivalence, the *finite* dimensionality of $\mathbb{K}^n$ is decisive. Both theorems do not hold in *infinite* dimensional normed spaces such as the space l_2 of (infinite) l_2-convergent

sequences or the space $C[a,b]$ of continuous functions on some interval $[a,b]$.

(ii) A subset $M \subset \mathbb{K}^n$ is called "compact" (or more precisely "sequentially compact"), if each sequence of vectors in M possesses a convergent subsequence with limit in M. Then, the theorem of Bolzano-Weierstrass implies that the compact subsets in $\mathbb{K}^n$ are exactly the bounded and closed subsets in $\mathbb{K}^n$.

(iii) A point $x \in \mathbb{K}^n$ is called "accumulation point" of a set $M \subset \mathbb{K}^n$ if each neighborhood of x contains at least one point from $M \setminus \{x\}$. The set of accumulation points of M is denoted by $\mathcal{H}(M)$ (closed "hull" of M). A point $x \in M \setminus \mathcal{H}(M)$ is called "isolated" .

Remark 1.3: In many applications there occur pairs $\{x,y\}$ (or more generally tuples) of points $x,y \in \mathbb{K}^n$. These form the so-called "product space" $V = \mathbb{K}^n \times \mathbb{K}^n$, which may be equipped with the generic norm

$$\|\{x,y\}\| := \big(\|x\|^2 + \|y\|^2\big)^{1/2}.$$

Since this space may be identified with the $2n$-dimensional euclidian space $\mathbb{K}^{2n}$ all results on subsets of $\mathbb{K}^n$ carry over to subsets of $\mathbb{K}^n \times \mathbb{K}^n$. This can be extended to more general product spaces of the form $V = \mathbb{K}^{n_1} \times \cdots \times \mathbb{K}^{n_m}$.

The basic concept in the geometry of $\mathbb{K}^n$ is that of "orthogonality" of vectors or subspaces. For its definition, we use a "scalar product".

Definition 1.3: *A mapping $(\cdot,\cdot) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$ is called "scalar product" if it has the following properties:*

(S1) Conjugate Symmetry: $(x,y) = \overline{(y,x)}\,, \quad x,y \in \mathbb{K}^n.$

(S2) Linearity: $(\alpha x + \beta y, z) = \alpha(x,z) + \beta(y,z)\,, \quad x,y,z \in \mathbb{K}^n,\ \alpha,\beta \in \mathbb{K}.$

(S3) Definiteness: $(x,x) \in \mathbb{R},\ (x,x) > 0, \ \ x \in \mathbb{K}^n \setminus \{0\}.$

In the following, we will mostly use the "euclidian" scalar product

$$(x,y)_2 = \sum_{j=1}^{n} x_j \overline{y_j}\,, \quad (x,x)_2 = \|x\|_2^2.$$

Remark 1.4: (i) If the requirement of *strict* definiteness, (S3), is relaxed, $(x,x) \in \mathbb{R},\ (x,x) \geq 0$, the sesquilinear form becomes a so-called "semi-scalar product".

(ii) From property (S2) (linearity in the first argument) and (S1) (conjugate symmetry), we obtain the conjugate linearity in the second argument. Hence, a scalar product is a special kind of "sesquilinear form" (if $\mathbb{K} = \mathbb{C}$) or "bilinear form" (if $\mathbb{K} = \mathbb{R}$)

(iii) In proving property (S1) (linearity) one may split the argument into two steps: "additivity", $(x^1 + x^2, y) = (x^1,y) + (x^2,y)$, and "homogeneity", $(\alpha x, y) = \alpha(x,y)$, $\alpha \in \mathbb{K}$.

Lemma 1.1: *For a scalar product $(\cdot,\cdot)$ on $\mathbb{K}^n$ there holds the "Cauchy-Schwarz inequality"*

$$|(x,y)|^2 \leq (x,x)(y,y), \quad x,y \in \mathbb{K}^n. \tag{1.1.3}$$

Proof. The assertion is obviously true for $y = 0$. Hence, we can now assume that $y \neq 0$. For arbitrary $\alpha \in \mathbb{K}$ there holds

$$0 \leq (x + \alpha y, x + \alpha y) = (x, x) + \alpha(y, x) + \overline{\alpha}(x, y) + \alpha\overline{\alpha}(y, y).$$

With $\alpha := -(x, y)(y, y)^{-1}$ this implies

$$\begin{aligned} 0 &\leq (x, x) - (x, y)(y, y)^{-1}(y, x) - \overline{(x, y)}(y, y)^{-1}(x, y) + (x, y)\overline{(x, y)}(y, y)^{-1} \\ &= (x, x) - |(x, y)|^2(y, y)^{-1} \end{aligned}$$

and, consequently, $0 \leq (x, x)(y, y) - |(x, y)|^2$. This completes the proof. Q.E.D.

The Cauchy-Schwarz inequality in $\mathbb{K}^n$ is a special case of the general "Hölder[1] inequality".

Corollary 1.1: *Any scalar product $(\cdot, \cdot)$ on $\mathbb{K}^n$ generates a norm $\|\cdot\|$ on $\mathbb{K}^n$ by*

$$\|x\| := (x, x)^{1/2}, \quad x \in \mathbb{K}^n.$$

The "euclidian" scalar product $(\cdot, \cdot)_2$ corresponds to the "euclidian" norm $\|x\|_2 := (x, x)_2^{1/2}$.

Proof. The norm properties (N1) and (N2) are obvious. It remains to show (N3). Using the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \|x + y\|^2 &= (x + y, x + y) = (x, x) + (x, y) + (y, x) + (y, y) \\ &\leq \|x\|^2 + 2|(x, y)| + \|y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = (\|x\| + ||y||)^2, \end{aligned}$$

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

Next, we provide a useful inequality, which is a special case of so-called "Young[2] inequalities".

Lemma 1.2 (Young inequality): *For $p, q \in \mathbb{R}$ with $1 < p, q < \infty$ and $1/p + 1/q = 1$, there holds the inequality*

$$|xy| \leq \frac{|x|^p}{p} + \frac{|y|^q}{q}, \quad x, y \in \mathbb{K}. \tag{1.1.4}$$

Proof. The logarithm $\ln(x)$ is on $\mathbb{R}_+$, in view of $\ln''(x) = -1/x^2 < 0$, a concave function. Hence, for $x, y \in \mathbb{K}$ there holds:

$$\ln\left(\tfrac{1}{p}|x|^p + \tfrac{1}{q}|y|^q\right) \geq \tfrac{1}{p}\ln(|x|^p) + \tfrac{1}{q}\ln(|y|^q) = \ln(|x|) + \ln(|y|).$$

Because of the monotonicity of the exponential function e^x it further follows that for $x, y \in \mathbb{K}$:

$$\tfrac{1}{p}|x|^p + \tfrac{1}{q}|y|^q \geq \exp\big(\ln(|x|) + \ln(|y|)\big) = \exp\big(\ln(|x|)\big)\exp\big(\ln(|y|)\big) = |x||y| = |xy|,$$

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

[1] Ludwig Otto Hölder (1859-1937): German mathematician; prof. in Tübingen; contributions first to the theory of Fourier series and later to group theory; found 1884 the inequality named after him.

[2] William Henry Young (1863-1942): English mathematician; worked at several universities world-wide, e. g., in Calcutta, Liverpool and Wales; contributions to differential and integral calculus, topological set theory and geometry.

Lemma 1.3 (Hölder inequality): *For the euclidian scalar product there holds, for arbitrary* $p, q \in \mathbb{R}$ *with* $1 < p, q < \infty$ *and* $1/p + 1/q = 1$, *the so-called "Hölder inequality"*

$$|(x, y)_2| \leq \|x\|_p \, \|y\|_q, \quad x, y \in \mathbb{K}^n. \tag{1.1.5}$$

This inequality also holds for the limit case $p = 1$, $q = \infty$.

Proof. For $x = 0$ or $y = 0$ the asserted estimate is obviously true. Hence, we can assume that $\|x\|_p \neq 0$ and $\|y\|_q \neq 0$. First, there holds

$$\frac{|(x, y)_2|}{\|x\|_p \|y\|_q} = \frac{1}{\|x\|_p \|y\|_q} \Big| \sum_{i=1}^{n} x_i \bar{y}_i \Big| \leq \sum_{i=1}^{n} \frac{|x_i||y_i|}{\|x\|_p \|y\|_q}.$$

Using the Young inequality it follows that

$$\frac{|(x, y)_2|}{\|x\|_p \|y\|_q} \leq \sum_{i=1}^{n} \Big\{ \frac{|x_i|^p}{p\|x\|_p^p} + \frac{|y_i|^q}{q\|y\|_q^q} \Big\} = \frac{1}{p\|x\|_p^p} \sum_{i=1}^{n} |x_i|^p + \frac{1}{q\|y\|_q^q} \sum_{i=1}^{n} |y_i|^q = \frac{1}{p} + \frac{1}{q} = 1.$$

This implies the asserted inequality. Q.E.D.

As consequence of the Hölder inequality, we obtain the so-called "Minkowski[3] inequality", which is the triangle inequality for the l_p norm.

Lemma 1.4 (Minkowski inequality): *For arbitrary* $p \in \mathbb{R}$ *with* $1 \leq p < \infty$ *as well as for* $p = \infty$ *there holds the "Minkowski inequality"*

$$\|x + y\|_p \leq \|x\|_p + \|y\|_p, \quad x, y \in \mathbb{K}^n. \tag{1.1.6}$$

Proof. For $p = 1$ and $p = \infty$ the inequality follows from the triangle inequality on $\mathbb{R}$:

$$\|x + y\|_1 = \sum_{i=1}^{n} |x_i + y_i| \leq \sum_{i=1}^{n} |x_i| + \sum_{i=1}^{n} |y_i| = \|x\|_1 + \|y\|_1,$$
$$\|x + y\|_\infty = \max_{1 \leq i \leq n} |x_i + y_i| \leq \max_{1 \leq i \leq n} |x_i| + \max_{1 \leq i \leq n} |y_i| = \|x\|_\infty + \|y\|_\infty.$$

Let now $1 < p < \infty$ and q be defined by $1/p + 1/q = 1$, i. e., $q = p/(p-1)$. We set

$$\xi_i := |x_i + y_i|^{p-1}, \;\; i = 1, \ldots, n, \quad \xi := (\xi_i)_{i=1}^n.$$

This implies that

$$\|x + y\|_p^p = \sum_{i=1}^{n} |x_i + y_i||x_i + y_i|^{p-1} \leq \sum_{i=1}^{n} |x_i|\xi_i + \sum_{i=1}^{n} |y_i|\xi_i$$

and further by the Hölder inequality

$$\|x + y\|_p^p \leq \|x\|_p \|\xi\|_q + \|y\|_p \|\xi\|_q = (\|x\|_p + \|y\|_p)\|\xi\|_q.$$

[3]Hermann Minkowski (1864-1909): Russian-German mathematician; prof. in Göttingen; several contributions to pure mathematics; introduced the non-euclidian 4-dimensional space-time continuum ("Minkowski space") for describing the theory of relativity of Einstein.

Observing $q = p/(p-1)$, we conclude

$$\|\xi\|_q^q = \sum_{i=1}^n |\xi_i|^q = \sum_{i=1}^n |x_i + y_i|^p = \|x+y\|_p^p,$$

and consequently,

$$\|x+y\|_p^p \le (\|x\|_p + \|y\|_p)\|x+y\|_p^{p/q} = (\|x\|_p + \|y\|_p)\|x+y\|_p^{p-1}.$$

This implies the asserted inequality. Q.E.D.

Using the euclidian scalar product, we can introduce a canonical notion of "orthogonality", i. e., two vectors $x, y \in \mathbb{K}^n$ are called "orthogonal" (abbreviated by $x \perp y$) if

$$(x,y)_2 = 0.$$

More general, two subspaces $N, M \subset \mathbb{K}^n$ are called "orthogonal" (abbreviated by $N \perp M$) if

$$(x,y)_2 = 0, \;\; x \in N,\, y \in M.$$

Accordingly to each subspace $M \in \mathbb{K}^n$, we can assign its "orthogonal complement" $M^\perp := \{x \in \mathbb{K}^n,\, \mathrm{span}(x) \perp M\}$, which is uniquely determined. Then, $\mathbb{K}^n = M \oplus M^\perp$, the "direct sum". Let $M \subset \mathbb{K}^n$ be a (nontrivial) subspace. Then, for any vector $x \in \mathbb{K}^n$ the "orthogonal projection" $P_M x \in M$ is determined by the relation

$$\|x - P_M x\|_2 = \min_{y \in M} \|x - y\|. \tag{1.1.7}$$

This "best approximation" property is equivalent to the relation

$$(x - P_M x, y)_2 = 0 \quad \forall y \in M, \tag{1.1.8}$$

which can be used to actually compute $P_M x$.

For arbitrary vectors there holds the "parallelogram identity" (exercise)

$$\|x+y\|_2^2 + \|x-y\|_2^2 = 2\|x\|_2^2 + 2\|y\|_2^2, \quad x, y \in \mathbb{K}^n, \tag{1.1.9}$$

and for orthogonal vectors the "Theorem of Pythagoras" (exercise):

$$\|x+y\|_2^2 = \|x\|_2^2 + \|y\|_2^2, \quad x, y \in \mathbb{K}^n,\; x \perp y. \tag{1.1.10}$$

A set of vectors $\{a^1, \ldots, a^m\}$, $a^i \neq 0$, of $\mathbb{K}^n$, which are mutually orthogonal, $(a^k, a^l) = 0$ for $k \neq l$ is necessarily linearly independent. Because for $\sum_{k=1}^m c_k a^k = 0$, successively taking the scalar product with a^l, $l = 1, \ldots, m$, yields

$$0 = \sum_{k=1}^m c_k (a^k, a^l)_2 = c_l (a^l, a^l)_2 \quad \Rightarrow \quad c_l = 0.$$

Definition 1.4: *A set of vectors $\{a^1, \ldots, a^m\}$, $a^i \neq 0$ of $\mathbb{K}^n$, which are mutually orthogonal, $(a^k, a^l)_2 = 0$, $k \neq l$, is called "orthogonal system" (in short "ONS") and in the case $m = n$ "orthogonal basis" (in short "ONB"). If $(a^k, a^k) = 1$, $k = 1, \ldots, m$, one speaks of an "orthonormal*

system" and an "orthonormal basis", respectively. The cartesian basis $\{e^1, \ldots, e^n\}$ *is obviously an orthonormal basis of* $\mathbb{R}^n$ *with respect to the euclidian scaler product. However, there are many other (actually infinitely many) of such orthonormal bases in* $\mathbb{R}^n$ *.*

Lemma 1.5: *Let* $\{a^i,\, i = 1, \ldots, n\}$ *be an orthonormal basis of* $\mathbb{K}^n$ *(with respect to the canonical euclidian scalar product). Then, each vector* $x \in \mathbb{K}^n$ *possesses a representation of the form (in analogy to the "Fourier expansion" with respect to the trigonometric functions)*

$$x = \sum_{i=1}^{n} (x, a^i)_2 \, a^i, \quad x \in \mathbb{K}^n, \tag{1.1.11}$$

and there holds the "Parseval[4] *identity"*

$$\|x\|_2^2 = \sum_{i=1}^{n} |(x, a^i)_2|^2, \quad x \in \mathbb{K}^n. \tag{1.1.12}$$

Proof. From the representation $x = \sum_{j=1}^{n} \alpha_j a^j$ taking the product with a^i it follows that

$$(x, a^i)_2 = \sum_{j=1}^{n} \alpha_j (a^j, a^i)_2 = \alpha_i, \quad i = 1, \ldots, n,$$

and consequently the representation (1.1.11). Further there holds:

$$\|x\|_2^2 = (x, x)_2 = \sum_{i,j=1}^{n} (x, a^i)_2 \overline{(x, a^j)_2} \, (a^i, a^j)_2 = \sum_{i=1}^{n} |(x, a^i)_2|^2,$$

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

By the following Gram-Schmidt algorithm, we can orthonormalize an arbitrary basis of $\mathbb{K}^n$, i. e., construct an *orthonormal* basis.

Theorem 1.3 (Gram-Schmidt algorithm): *Let* $\{a^1, \ldots, a^n\}$ *be any basis of* $\mathbb{K}^n$ *. Then, the following so-called "Gram*[5]*-Schmidt*[6] *orthonormalization algorithm",*

$$\begin{aligned} b^1 &:= \|a^1\|_2^{-1} a^1, \\ \tilde{b}^k &:= a^k - \sum_{j=1}^{k-1} (a^k, b^j)_2 b^j, \quad b^k := \|\tilde{b}^k\|_2^{-1} \tilde{b}^k, \quad k = 2, \ldots, n, \end{aligned} \tag{1.1.13}$$

yields an orthonormal basis $\{b^1, \ldots, b^n\}$ *of* $\mathbb{K}^n$*.*

[4] *Marc-Antoine Parseval des Chênes (1755-1836): French mathematician; worked on partial differential equations in physics (only five mathematical publications); known by the identity named after him, which he stated without proof and connection to Fourier series.*

[5] *Jørgen Pedersen Gram (1850-1916): Danish mathematician, employee and later owner of an insurance company, contributions to algebra (invariants theory), probability theory, numerics and forestry; the orthonormalization algorithm named after him had already been used before by Cauchy 1836.*

[6] *Erhard Schmidt (1876-1959): German mathematician, prof. in Berlin, there founder of the Institute for Applied Mathematics 1920, after the war Director of the Mathematical Institute of the Academy of Sciences of DDR; contributions to the theory of integral equations and Hilbert spaces and later to general topology.*

Proof. First, we show that the construction process of the b^k does not stop with $k < n$. By construction the vectors b^k are linear combinations of the $a^1, \ldots, a^k$. If for some $k \leq n$

$$a^k - \sum_{j=1}^{k-1} (a^k, b^j)_2 b^j = 0,$$

the vectors $\{a^1, \ldots, a^k\}$ would be linearly dependent contradicting the a priori assumption that $\{a^1, \ldots, a^n\}$ is a basis. Now, we show by induction that the Gram-Schmidt process yields an orthonormal basis. Obviously $\|b^1\|_2 = 1$. Let now $\{b^1, \ldots, b^k\}$, for $k \leq n$, be an already constructed orthonormal system. Then, for $l = 1, \ldots, k$, there holds

$$(b^{k+1}, b^l)_2 = (a^{k+1}, b^l)_2 - \sum_{j=1}^{k} (a^{k+1}, b^j)_2 \underbrace{(b^j, b^l)_2}_{= \delta_{jl}} = 0$$

and $\|b^{k+1}\|_2 = 1$, i. e., $\{b^1, \ldots, b^{k+1}\}$ is also an orthonormal system. Q.E.D.

The Gram-Schmidt algorithm in its "classical" form (1.1.13) is numerically unstable due to accumulation of round-off errors. Below, in Section 4.3.1, we will consider a stable version, the so-called "modified Gram-Schmidt algorithm", which for *exact* arithmetic yields the same result.

1.1.2 Linear mappings and matrices

We now consider linear mappings from the n-dimensional vector space $\mathbb{K}^n$ into the m-dimensional vector space $\mathbb{K}^m$, where not necessarily $m = n$. However, the special case $m = n$ plays the most important role. A mapping $\varphi = (\varphi_1, \ldots, \varphi_m) : \mathbb{K}^n \to \mathbb{K}^m$ is called "linear", if for $x, y \in \mathbb{K}^n$ and $\alpha, \beta \in \mathbb{K}$ there holds

$$\varphi(\alpha x + \beta y) = \alpha \varphi(x) + \beta \varphi(y). \tag{1.1.14}$$

The action of a linear mapping φ on a vector space can be described in several ways. It obviously suffices to prescribe the action of φ on the elements of a basis of the space, e. g., a cartesian basis $\{e^i, i = 1, \ldots, n\}$,

$$x = \sum_{i=1}^{n} x_i e^i \quad \to \quad \varphi(x) = \varphi\Big(\sum_{i=1}^{n} x_i e^i \Big) = \sum_{i=1}^{n} x_i \varphi(e^i).$$

Thereby, to each vector (or point) $x \in \mathbb{K}^n$ a "coordinate vector" $\hat{x} = (x_i)_{i=1}^n$ is uniquely associated. If also the images $\varphi(x)$ are expressed with respect to a cartesian basis of $\mathbb{K}^m$,

$$\varphi(x) = \sum_{j=1}^{m} \varphi_j(x) e^j = \sum_{j=1}^{m} \Big(\sum_{i=1}^{n} \underbrace{\varphi_j(e^i)}_{=: a_{ji}} x_i \Big) e^j,$$

with the coordinate vector $\hat{\varphi}(x) = (\varphi_j(x))_{j=1}^m$, we can write the action of the mapping φ on a vector $x \in \mathbb{K}^n$ in "matrix form" using the usual rules of matrix-vector multiplication as follows:

$$\varphi_j(x) = (A\hat{x})_j := \sum_{i=1}^{n} a_{ji} x_i, \quad j = 1, \ldots, m,$$

with the $n \times m$-array of numbers $A = (a_{ij})_{i,j=1}^{n,m} \in \mathbb{K}^{m\times n}$, a "matrix",

$$\begin{pmatrix} \varphi_1(e^1) & \cdots & \varphi_1(e^n) \\ \vdots & \ddots & \vdots \\ \varphi_m(e^1) & \cdots & \varphi_m(e^n) \end{pmatrix} =: \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} = A \in \mathbb{K}^{m\times n}.$$

By this matrix $A \in \mathbb{K}^{m\times n}$ the linear mapping φ is uniquely described with respect to the chosen bases of $\mathbb{K}^n$ and $\mathbb{K}^m$. In the following discussion, for simplicity, we identify the point $x \in \mathbb{K}^n$ with its special cartesian coordinate representation $\hat{x}$. Here, we follow the convention that in the notation $\mathbb{K}^{m\times n}$ for matrices the first parameter m stands for the dimension of the image space $\mathbb{K}^m$, i. e., the number of rows in the matrix, while the second one n corresponds to the dimension of the origin space $\mathbb{K}^n$, i. e., the number of columns. Accordingly, for a matrix entry a_{ij} the first index refers to the row number and the second one to the column number of its position in the matrix. We emphasize that this is only one of the possible concrete representations of the linear mapping $\varphi : \mathbb{K}^n \to \mathbb{K}^m$. In this sense each quadratic matrix $A \in \mathbb{K}^{n\times n}$ represents a linear mapping in $\mathbb{K}^n$. The identity map $\varphi(x) = x$ is represented by the "identity matrix" $I = (\delta_{ij})_{i,j=1}^n$ where $\delta_{ij} := 1$ for $i = j$ and $\delta_{ij} = 0$ else (the usual "Kronecker symbol").

Clearly, two matrices $A, A' \in \mathbb{K}^{m\times n}$ are identical, i. e., $a_{ij} = a'_{ij}$ if and only if $Ax = A'x$, $x \in \mathbb{K}^n$. To a general matrix $A \in \mathbb{K}^{m\times n}$, we associate the "adjoint transpose" $\bar{A}^T = (a^T_{i,j})^{n\times m}_{i,j=1}$ by setting $a^T_{ij} := \bar{a}_{ji}$. A quadratic matrix $A \in \mathbb{K}^{n\times n}$ is called "regular", if the corresponding linear mapping is injective and surjective, i. e., bijective, with "inverse" denoted by $A^{-1} \in \mathbb{K}^{n\times n}$. Further, to each matrix $A \in \mathbb{K}^{n\times n}$, we associate the following quantities, which are uniquely determined by the corresponding linear mapping φ:

- "determinant" of A: $\det(A)$.
- "adjugate" of A: $adj(A) := C^T$, $c_{ij} := (-1)^{i+j} A_{ij}$ (A_{ij} the cofactors of A).
- "trace" of A: $\mathrm{trace}(A) := \sum_{i=1}^n a_{ii}$.

The following property of the determinant will be useful below: $\det(\bar{A}^T) = \overline{\det(A)}$.

Lemma 1.6: *For a square matrix $A = (a_{ij})_{i,j=1}^n \in \mathbb{K}^{n\times n}$ the following statements are equivalent:*

(i) A is regular with inverse A^{-1}.

(ii) The equation $Ax = b$ has for any $b \in \mathbb{K}^n$ a unique solution (bijectivity).

(iii) The equation $Ax = 0$ has only the zero solution $x = 0$ (injectivity).

(iv) The equation $Ax = b$ has for any $b \in \mathbb{K}^n$ a solution (surjectivity).

(vi) $\det(A) \neq 0$.

(viii) The adjoint transpose $\bar{A}^T$ is regular with inverse $(\bar{A}^T)^{-1} = (\overline{A^{-1}})^T$.

Proof. For the proof, we refer to the standard linear algebra literature. Q.E.D.

Lemma 1.7: *For a general matrix $A \in \mathbb{K}^{m\times n}$, we introduce its "range"* $\mathrm{range}(A) := \{y \in \mathbb{K}^m | \, y = Ax$ *for some* $x \in \mathbb{K}^n\}$ *and its "kernel" (or "null space")* $\mathrm{kern}(A) := \{x \in \mathbb{K}^n \,|\, Ax = 0\}$. *There holds*

$$\mathrm{range}(A) = \mathrm{kern}(\bar{A}^T)^T, \qquad \mathrm{range}(\bar{A}^T) = \mathrm{kern}(A)^T, \tag{1.1.15}$$

i. e., the equation $Ax = b$ has a solution if and only if $(b, y)_2 = 0$ for all $y \in \mathrm{kern}(\bar{A}^T)$.

Proof. For the proof, we refer to the standard linear algebra literature. Q.E.D.

In many practical applications the governing matrices have special properties, which require the use of likewise special numerical methods. Some of the most important properties are those of "symmetry" or "normality" and "definiteness".

Definition 1.5: *(i) A quadratic matrix $A \in \mathbb{K}^{n\times n}$ is called "hermitian" if it satisfies*

$$A = \bar{A}^T \quad \Leftrightarrow \quad a_{ij} = \bar{a}_{ji},\ i, j = 1, \ldots, n, \tag{1.1.16}$$

or equivalently,

$$(Ax, y)_2 = (x, Ay)_2, \quad x, y \in \mathbb{K}^n. \tag{1.1.17}$$

(ii) It is called "normal" if $\bar{A}^T A = A\bar{A}^T$.

(iii) It is called "positive semi-definite" if

$$(Ax, x)_2 \in \mathbb{R}, \quad (Ax, x)_2 \geq 0, \ \ x \in \mathbb{K}^n. \tag{1.1.18}$$

and "positive definite" if

$$(Ax, x)_2 \in \mathbb{R}, \quad (Ax, x)_2 > 0, \ \ x \in \mathbb{K}^n \setminus \{0\}. \tag{1.1.19}$$

(iv) A real hermitian matrix $A \in \mathbb{R}^{n\times n}$ is called "symmetric".

Lemma 1.8: *For a hermitian positive definite matrix $A \in \mathbb{K}^{n\times n}$ the main diagonal elements are real and positive, $a_{ii} > 0$, and the element with largest modulus lies on the main diagonal.*

Proof. (i) From $a_{ii} = \bar{a}_{ii}$ it follows that $a_{ii} \in \mathbb{R}$. The positiveness follows via testing by the unit vector e^i yielding $a_{ii} = (Ae^i, e^i)_2 > 0$.
(ii) Let $a_{ij} \neq 0$ be an element of A with maximal modulus and suppose that $i \neq j$. Testing now by $x = e^i - \mathrm{sign}(a_{ij})e^j \neq 0$, we obtain the following contradiction to the positiveness of A:

$$\begin{aligned} 0 < (Ax, x)_2 &= (Ae^i, e^i)_2 - 2\,\mathrm{sign}(a_{ij})(Ae^i, e^j)_2 + \mathrm{sign}(a_{ij})^2(Ae^j, e^j)_2 \\ &= a_{ii} - 2\,\mathrm{sign}(a_{ij})a_{ij} + a_{jj} = a_{ii} - 2|a_{ij}| + a_{jj} \leq 0. \end{aligned}$$

This completes the proof. Q.E.D.

Remark 1.5 (Exercises): (i) If a matrix $A \in \mathbb{K}^{n\times n}$ is positive definite (or more generally just satisfies $(Ax, x)_2 \in \mathbb{R}$ for $x \in \mathbb{C}^n$), then it is necessarily hermitian. This does not need to be true for *real* matrices $A \in \mathbb{R}^{n\times n}$.

(ii) The general form of a scalar product $(\cdot,\cdot)$ on $\mathbb{K}^n$ is given by $(x,y) = (Ax,y)_2$ with a (hermitian) positive definite matrix $A \in \mathbb{K}^{n\times n}$.

Definition 1.6 (Orthonormal matrix): *A matrix $Q \in \mathbb{K}^{m\times n}$ is called "orthogonal" or "orthonormal" if its column vectors form an orthogonal or orthonormal system in $\mathbb{K}^n$, respectively. In the case $n = m$ such a matrix is called "unitary".*

Lemma 1.9: *A unitary matrix $Q \in \mathbb{K}^{n\times n}$ is regular and its inverse is $Q^{-1} = \bar{Q}^T$. Further, there holds:*

$$(Qx,Qy)_2 = (x,y)_2, \quad x,y \in \mathbb{K}^n, \tag{1.1.20}$$

$$\|Qx\|_2 = \|x\|_2, \quad x \in \mathbb{K}^n. \tag{1.1.21}$$

Proof. First, we show that $\overline{Q}^T$ is the inverse of Q. Let $q_i \in \mathbb{K}^n$ denote the column vectors of Q satisfying by definition $(q_i,q_j)_2 = q_i^T\overline{q}_j = \delta_{ij}$. This implies:

$$\overline{Q}^T Q = \begin{pmatrix} \overline{q}_1^T q_1 & \dots & \overline{q}_1^T q_n \\ \vdots & \ddots & \vdots \\ \overline{q}_n^T q_1 & \dots & \overline{q}_n^T q_n \end{pmatrix} = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{pmatrix} = I.$$

From this it follows that

$$(Qx,Qy)_2 = (x,\bar{Q}^T Qx)_2 = (x,y)_2, \quad x,y \in \mathbb{K}^n,$$

and further

$$\|Qx\|_2 = (Qx,Qx)_2^{1/2} = \|x\|_2, \quad x \in \mathbb{K}^n,$$

which completes the proof. Q.E.D.

Example 1.2: The real unitary matrix

$$Q_\theta^{(ij)} = \begin{array}{c} \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & i & & j \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \cos(\theta) & 0 & -\sin(\theta) & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \sin(\theta) & 0 & \cos(\theta) & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \\ i \\ \\ j \end{array}$$

describes a rotation in the (x_i,x_j)-plane about the origin $x = 0$ with angle $\theta \in [0,2\pi)$.

Remark 1.6: (i) In view of the relations (1.1.20) and (1.1.21) euclidian scalar product and euclidian norm of vectors are invariant under unitary transformations. This explains why it is the euclidian norm, which is used for measuring length or distance of vectors in $\mathbb{R}^n$.

(ii) The Schwarz inequality (1.1.3) allows the definition of an "angle" between two vectors in

$\mathbb{R}^n$. For any number $\alpha \in [-1,1]$ there is exactly one $\theta \in [0,\pi]$ such that $\alpha = \cos(\theta)$. By

$$\cos(\theta) = \frac{(x,y)_2}{\|x\|_2\|y\|_2}, \quad x,y \in \mathbb{K}^n \setminus \{0\},$$

a $\theta \in [0,\pi]$ is uniquely determined. This is then the "angle" between the two vectors x and y. The relation (1.1.20) states that the euclidian scalar product of two vectors in $\mathbb{K}^n$ is invariant under rotations. By some rotation Q in $\mathbb{R}^n$, we can achieve that $Qx, Qy \in \operatorname{span}\{e^{(1)}, e^{(2)}\}$ and $Qx = \|x\|_2 e^{(1)}$. Then, there holds

$$(x,y)_2 = (Qx,Qy)_2 = \|x\|_2(e^{(1)},Qy)_2 = \|x\|_2(Qy)_1 = \|x\|_2\|Qy\|_2\cos(\theta) = \|x\|_2\|y\|_2\cos(\theta),$$

i. e., θ is actually the "angle" between the two vectors in the sense of elementary geometry.

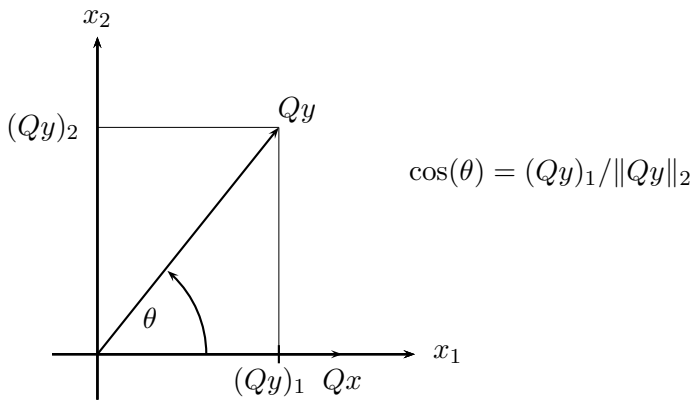


Figure 1.1: *Angle between two vectors* $x = \|x\|e^1$ *and* y *in* $\mathbb{R}^2$.

1.1.3 Non-quadratic linear systems

Let $A \in \mathbb{R}^{m\times n}$ be a not necessarily quadratic coefficient matrix and $b \in \mathbb{R}^m$ a right-hand side vector. We concentrate in the case $m \neq n$ and consider the non-quadratic linear system

$$Ax = b, \tag{1.1.22}$$

for $x \in \mathbb{R}^n$. Here, $\operatorname{rank}(A) < \operatorname{rank}[A,b]$ is allowed, i. e., the system does not need to possess a solution in the normal sense. In this case an appropriately extended notion of "solution" is to be used. In the following, we consider the so-called "method of least error-squares", which goes back to Gauss. In this approach a vector $\bar{x} \in \mathbb{R}^n$ is seeked with minimal defect norm $\|d\|_2 = \|b - A\bar{x}\|_2$. Clearly, this extended notion of "solution" coincides with the traditional one if $\operatorname{rank}(A) = \operatorname{rank}[A,b]$.

Theorem 1.4 ("Least error-squares" solution): *There exists always a "solution"* $\bar{x} \in \mathbb{R}^n$ *of (2.3.25) in the sense of least error-squares ("least error-squares" solution)*

$$\|A\bar{x} - b\|_2 = \min_{x\in\mathbb{R}^n} \|Ax - b\|_2. \tag{1.1.23}$$

This is equivalent to $\bar{x}$ *being solution of the so-called "normal equation":*

$$A^T A\bar{x} = A^T b. \tag{1.1.24}$$

If $m \geq n$ *and* $\operatorname{rank}(A) = n$ *the "least error-squares" solution* $\bar{x}$ *is uniquely determined. Otherwise each other solution has the form* $\bar{x} + y$ *with* $y \in \operatorname{kern}(A)$*. In this case, there always exists such a solution with minimal euclidian norm, i. e., a "minimal" solution with least error-squares,*

$$\|x^{\min}\|_2 = \min\{\|\bar{x} + y\|_2,\ y \in \operatorname{kern}(A)\}. \tag{1.1.25}$$

Proof. (i) Let $\bar{x}$ be a solution of the normal equation. Then, for arbitrary $x \in \mathbb{R}^n$ there holds

$$\begin{aligned} \|b - Ax\|_2^2 &= \|b - A\bar{x} + A(\bar{x} - x)\|_2^2 \\ &= \|b - A\bar{x}\|_2^2 + 2\underbrace{(b - A\bar{x}}_{\in \operatorname{kern}(A^T)}, \underbrace{A[\bar{x} - x])}_{\in \operatorname{range}(A)} + \|A(\bar{x} - x)\|_2^2 \geq \|b - A\bar{x}\|_2^2, \end{aligned}$$

i. e., $\bar{x}$ has least error-squares. In turn, for such a least error-squares solution $\bar{x}$ there holds

$$\begin{aligned} 0 = \frac{\partial}{\partial x_i}\|Ax - b\|_{2|x=\bar{x}}^2 &= \frac{\partial}{\partial x_i}\Big(\sum_{j=1}^n \Big|\sum_{k=1}^n a_{jk}x_k - b_j\Big|^2\Big)_{|x=\bar{x}} \\ &= 2\sum_{j=1}^n a_{ji}\Big(\sum_{k=1}^n a_{jk}\bar{x}_k - b_j\Big) = 2(A^T A\bar{x} - A^T b)_i, \end{aligned}$$

i. e., $\bar{x}$ solves the normal equation.

(ii) We now consider the solvability of the normal equation. The orthogonal complement of $\operatorname{range}(A)$ in $\mathbb{R}^m$ is $\operatorname{kern}(A^T)$. Hence the element b has a unique decomposition

$$b = s + r\,, \quad s \in \operatorname{range}(A)\,, \quad r \in \operatorname{kern}(A^T).$$

Then, for any $\bar{x} \in \mathbb{R}^n$ satisfying $A\bar{x} = s$ there holds

$$A^T A\bar{x} = A^T s \;=\; A^T s + A^T r = A^T b,$$

i. e., $\bar{x}$ solves the normal equation. In case that $\operatorname{range}(A) = n$ there holds $\operatorname{kern}(A) = \{0\}$ and $\operatorname{range}(A) = \mathbb{R}^n$. Observing $A^T Ax = 0$ and $\operatorname{kern}(A^T) \perp \operatorname{range}(A)$, we conclude $Ax = 0$ and $x = 0$. The matrix $A^T A \in \mathbb{R}^{n\times n}$ is regular and consequently $\bar{x}$ uniquely determined. In case that $\operatorname{range}(A) < n$, for any other solution x_1 of the normal equation, we have

$$b = Ax_1 + (b - Ax_1) \;\in\; \operatorname{range}(A) + \operatorname{kern}(A^T) = \operatorname{range}(A) + \operatorname{range}(A)^T.$$

In view of the uniqueness of the orthogonal decomposition, we necessarily obtain $Ax_1 = A\bar{x}$ and $\bar{x} - x_1 \in \operatorname{kern}(A)$.

(iii) We finally consider the case $\operatorname{rank}(A) < n$. Among the solutions $\bar{x} + \operatorname{kern}(A)$ of the normal equation, we can find one with minimal euclidian norm,

$$\|x^{\min}\|_2 = \min\{\|\bar{x} + y\|_2,\ y \in \operatorname{kern}(A)\}.$$

This follows from the non-negativity of the function $F(y) := \|\bar{x} + y\|_2$ and its uniform strict convexity, which also implies uniqueness of the minimal solution. Q.E.D.

For the computation of the "solution with smallest error-squares" of a non-quadratic system $Ax = b$, we have to solve the normal equation $A^T Ax = A^T b$. Efficient methods for this task will be discussed in the next chapter.

Lemma 1.10: *For any matrix $A \in \mathbb{K}^{m\times n}$ the matrices $\bar{A}^T A \in \mathbb{K}^{n\times n}$ and $A\bar{A}^T \in \mathbb{K}^{m\times m}$ are hermitian (symmetric) and positive semi-definite. In the case $m \geq n$ and if $\operatorname{rank}(A) = n$ the matrix $\bar{A}^T A$ it is even positive definite.*

Proof. Following the rules of matrix arithmetic there holds

$$(\bar{A}^T A)^T = A^T \bar{A} = \overline{\bar{A}^T A}, \qquad \bar{x}^T (\bar{A}^T A)x = \overline{(Ax)}^T Ax = \|Ax\|_2^2 \geq 0,$$

i. e., $\bar{A}^T A$ is hermitian and positive semi-definite. The argument for $A\bar{A}^T$ is analogous. In case that $m \geq n$ and $\operatorname{rank}(A) = n$ the matrix viewed as mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ is injective, i. e., $\|Ax\|_2 = 0$ implies $x = 0$. Hence, the matrix $\bar{A}^T A$ is positive definite. Q.E.D.

1.1.4 Eigenvalues and eigenvectors

In the following, we consider square matrices $A = (a_{ij})_{i,j=1}^n \in \mathbb{K}^{n\times n}$.

Definition 1.7: *(i) A number $\lambda \in \mathbb{C}$ is called "eigenvalue" of A, if there is a corresponding "eigenvector" $w \in \mathbb{C}^n$, $w \neq 0$, such that the following "eigenvalue equation" holds:*

$$Aw = \lambda w. \tag{1.1.26}$$

(ii) The vector space of all eigenvectors of an eigenvalue λ is called "eigenspace" and denoted by E_λ. Its dimension is the "geometric multiplicity" of λ. The set of all eigenvalue of a matrix $A \in \mathbb{K}^{n\times n}$ is called its "spectrum" and denoted by $\sigma(A) \subset \mathbb{C}$. The matrix function $R_A(z) := zI - A$ is called the "resolvent" of A and $\operatorname{Res}(A) := \{z \in \mathbb{C} \,|\, zI - A$ is regular$\}$ the corresponding "resolvent set".

(iii) The eigenvalues are just the zeros of the "characteristic polynomial" $\chi_A \in P_n$ of A,

$$\chi_A(z) := \det(zI - A) = z^n + b_1 z^{n-1} + \ldots + b_n.$$

Hence, by the fundamental theorem of algebra there are exactly n eigenvalues counted accordingly to their multiplicity as zeros of χ_A, their so-called "algebraic multiplicities". The algebraic multiplicity is always greater or equal than the geometric multiplicity. If it is strictly greater, then the eigenvalue is called "deficient" and the difference the "defect" of the eigenvalue.

(iv) The eigenvalues of a matrix can be determined independently of each other. One speaks of the "partial eigenvalue problem" if only a small number of the eigenvalues (e.g., the largest or the smallest) and the corresponding eigenvectors are to be determined. In the "full eigenvalue problem" one seeks all eigenvalues with their corresponding eigenvectors. For a given eigenvalue $\lambda \in \mathbb{C}$ (e. g., obtained as a zero of the characteristic polynomial) a corresponding eigenvector

can be determined as any solution of the (singular) problem

$$(A - \lambda I)w = 0. \tag{1.1.27}$$

Conversely, for a given eigenvector $w \in \mathbb{K}^n$ *(e. g., obtained by the "power method" described below), one obtains the corresponding eigenvalue by evaluating any of the quotients (choosing* $w_i \neq 0$ *)*

$$\lambda = \frac{(Aw)_i}{w_i},\ i = 1, \ldots, n, \qquad \lambda = \frac{(Aw, w)_2}{\|w\|_2^2}.$$

The latter quotient is called the "Rayleigh[7] *quotient".*

The characteristic polynomial of a matrix $A \in \mathbb{K}^{n\times n}$ has the following representation with its mutually distinct zeros λ_i:

$$\chi_A(z) = \prod_{i=1}^{m} (z - \lambda_i)^{\sigma_i}, \quad \sum_{i=1}^{m} \sigma_i = n,$$

where σ_i is the algebraic multiplicity of eigenvalue λ_i. Its geometric multiplicity is $\rho_i := \dim(\mathrm{kern}(A - \lambda_i I))$. We recall that generally $\rho_i \leq \sigma_i$, i. e., the defect satisfies $\alpha_i := \sigma_i - \rho_i \geq 0$. The latter corresponds to the largest integer $\alpha = \alpha(\lambda)$ such that

$$\mathrm{kern}(A - \lambda I)^{\alpha+1}) \neq \mathrm{kern}((A - \lambda I)^{\alpha}). \tag{1.1.28}$$

Since

$$\det(\bar{A}^T - \bar{z}I) = \det(\overline{A^T - zI}) = \det(\overline{A - zI)^T}) = \overline{\det(A - zI)}$$

the eigenvalues of the matrices A and $\bar{A}^T$ are related by

$$\lambda(\bar{A}^T) = \overline{\lambda(A)}. \tag{1.1.29}$$

Hence, associated to a normalized "primal" (right) eigenvector $w \in \mathbb{K}^n$, $\|w\|_2 = 1$, corresponding to an eigenvalue λ of A there is a "dual" (left) eigenvector $w^* \in \mathbb{K}^n \backslash \{0\}$ corresponding to the eigenvalue $\bar{\lambda}$ of $\bar{A}^T$ satisfying the "adjoint" eigenvalue equation

$$\bar{A}^T w^* = \bar{\lambda}\, w^* \quad \Leftrightarrow \quad \bar{w}^{*T} A = \lambda \bar{w}^{*T}. \tag{1.1.30}$$

The dual eigenvector w^* may also be normalized by $\|w^*\|_2 = 1$ or, what is more suggested by numerical purposes, by $(w, w^*)_2 = 1$. In the "degenerate" case $(w, w^*)_2 = 0$, and only then, the problem

$$Aw^1 - \lambda w^1 = w \tag{1.1.31}$$

has a solution $w^1 \in \mathbb{K}^n$. This follows from the relations $w^* \in \mathrm{kern}(\bar{A}^T - \bar{\lambda}I)$, $w \perp \mathrm{kern}(\bar{A}^T - \bar{\lambda}I)$, and $\mathrm{range}(A - \lambda I) = \mathrm{kern}(\bar{A}^T - \bar{\lambda}I)^T$, the latter following from the result of Lemma 1.7. The vector w^1 is called "generalized eigenvector (of level one)" of A (or "Hauptvektor erster Stufe" in German) corresponding to the eigenvalue λ. Within this notion, eigenvectors are

[7] *John William Strutt (Lord Rayleigh) (1842-1919): English mathematician and physicist; worked at the beginning as (aristocratic) private scholar, 1879-1884 professor for experimental physics in Cambridge; fundamental contributions to theoretical physics: scattering theory, acoustics, electro-magnetics, gas dynamics.*

"generalized eigenvectors" of level zero. By definition, there holds

$$(A - \lambda I)^2 w^1 = (A - \lambda I) w = 0,$$

i. e., $w^1 \in \mathrm{kern}((A - \lambda I)^2)$ and, consequently, in view of the above definition, the eigenvalue λ has "defect" $\alpha(\lambda) \geq 1$. If this construction can be continued, i. e., if $(w^1, w^*)_2 = 0$, such that also the problem $Aw^2 - \lambda w^2 = w^1$ has a solution $w^2 \in \mathbb{K}^n$, which is then a "generalized eigenvector" of level two, by construction satisfying $(A - \lambda I)^3 w^2 = 0$. In this way, we may obtain "generalize eigenvectors" $w^m \in \mathbb{K}^n$ of level m for which $(A - \lambda I)^{m+1} w^m = 0$ and $(w^m, w^*)_2 \neq 0$. Then, the eigenvalue λ has defect $\alpha(\lambda) = m$.

Example 1.3: The following special matrices $C_m(\lambda)$ occur as building blocks, so-called "Jordan blocks", in the "Jordan[8] normal form" of a matrix $A \in \mathbb{K}^{n \times n}$ (see below):

$$C_m(\lambda) = \begin{bmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ 0 & & & & \lambda \end{bmatrix} \in \mathbb{K}^{m \times m}, \quad \text{eigenvalue} \quad \lambda \in \mathbb{C}$$

$$\chi_{C_m(\lambda)}(z) = (z - \lambda)^m \quad \Rightarrow \quad \sigma = m, \quad \mathrm{rank}(C_m(\lambda) - \lambda I) = m - 1 \quad \Rightarrow \quad \rho = 1.$$

1.1.5 Similarity transformations

Two matrices $A, B \in \mathbb{K}^{n \times n}$ are called "similar (to each other)", if there is a regular matrix $T \in \mathbb{K}^{n \times n}$ such that

$$B = T^{-1} A T.$$

The transition $A \to B$ is called "similarity transformation". Suppose that the matrix $A \in \mathbb{K}^{n \times n}$ is the representation of a linear mapping $\varphi : \mathbb{K}^n \to \mathbb{K}^n$ with respect to a basis $\{a^1, \ldots, a^n\}$ of $\mathbb{K}^n$. Then, using the regular matrix $T \in \mathbb{K}^{n \times n}$, we obtain a second basis $\{Ta^1, \ldots, Ta^n\}$ of $\mathbb{K}^n$ and B is the representation of the mapping φ with respect to this new basis. Hence, similar matrices are representations of the same linear mapping and any two representations of the same linear mapping are similar. In view of this fact, we expect that two similar matrices, representing the same linear mapping, have several of their characteristic quantities as matrices in common.

Lemma 1.11: *For any two similar matrices $A, B \in \mathbb{K}^{n \times n}$ there holds:*

a) $\det(A) = \det(B)$.

b) $\sigma(A) = \sigma(B)$.

c) $\mathrm{trace}(A) = \mathrm{trace}(B)$.

[8] Marie Ennemond Camille Jordan (1838-1922): French mathematician; prof. in Paris; contributions to algebra, group theory, calculus and topology.

Proof. i) The product theorem for determinants implies that $\det(AB) = \det(A)\det(B)$ and further $\det(T^{-1}) = \det(T)^{-1}$. This implies that

$$\det(B) = \det(T^{-1}AT) = \det(T^{-1})\det(A)\det(T) = \det(T)^{-1}\det(A)\det(T) = \det(A).$$

ii) Further, for any $z \in \mathbb{C}$ there holds

$$\begin{aligned}\det(zI - B) &= \det(zT^{-1}T - T^{-1}AT) = \det(T^{-1}(zI - A)T)\\ &= \det(T^{-1})\det(zI - A)\det(T) = \det(zI - A),\end{aligned}$$

which implies that A and B have the same eigenvalues.
iii) The trace of A is just the coefficient of the monom z^{n-1} in the characteristic polynomial $\chi_A(z)$. Hence by (i) the trace of A equals that of B. Q.E.D.

Any matrix $A \in \mathbb{K}^{n\times n}$ is similar to its "canonical form" (Jordan normal form) which has the eigenvalues λ_i of A on its main diagonal counted accordingly to their algebraic multiplicity. Hence, in view of Lemma 1.11 there holds

$$\det(A) = \prod_{i=1}^{n} \lambda_i, \qquad \text{trace}(A) = \sum_{i=1}^{n} \lambda_i. \tag{1.1.32}$$

Definition 1.8 (Normal forms): *(i) Any matrix $A \in \mathbb{K}^{n\times n}$ is similar to its "canonical normal form" J_A ("Jordan normal form") which is a block diagonal matrix with main diagonal blocks, the "Jordan blocks", of the form as shown in Example 1.3. Here, the "algebraic" multiplicity of an eigenvalue corresponds to the number of occurrences of this eigenvalue on the main diagonal of J_A, while its "geometric" multiplicity corresponds to the number of Jordan blocks containing λ.*

(ii) A matrix $A \in \mathbb{K}^{n\times n}$, which is similar to a diagonal matrix, then having its eigenvalues on the main diagonal, is called "diagonalizable",

$$WAW^{-1} = \Lambda = \text{diag}(\lambda_i) \quad (\lambda_i \text{ eigenvalues of } A).$$

This relation implies that the transformation matrix $W = [w^1, \ldots, w^n]$ has the eigenvectors w^i corresponding to the eigenvalues λ_i as column vectors. This means that orthogonalizability of a matrix is equivalent to the existence of a basis of eigenvectors.

(iii) A matrix $A \in \mathbb{K}^{n\times n}$ is called "unitarily diagonalizable" if it is diagonalizable with a unitary transformation matrix. This is equivalent to the existence of an orthonormal basis of eigenvectors.

Positive definite hermitian matrices $A \in \mathbb{K}^{n\times n}$ have very special spectral properties. These are collected in the following lemma and theorem, the latter one being the basic result of matrix analysis ("spectral theorem").

Lemma 1.12: *(i) A hermitian matrix has only real eigenvalues and eigenvectors to different eigenvalues are mutually orthogonal.*

(ii) A hermitian matrix is positive definite if and only if all its (real) eigenvalues are positive.

(iii) Two normal matrices $A, B \in \mathbb{K}^{n\times n}$ commute, $AB = BA$, if and only if they possess a common basis of eigenvectors.

Proof. For the proofs, we refer to the standard linear algebra literature. Q.E.D.

Theorem 1.5 (Spectral theorem): *For square hermitian matrices, $A = \bar{A}^T$, or more general for "normal" matrices, $\bar{A}^T A = A\bar{A}^T$, algebraic and geometric multiplicities of eigenvalues are equal, i. e., these matrices are diagonalizable. Further, they are even unitarily diagonalizable, i. e., there exists an orthonormal basis of eigenvectors.*

Proof. For the proof, we refer to the standard linear algebra literature. Q.E.D.

1.1.6 Matrix analysis

We now consider the vector space of all $m \times n$-matrices $A \in \mathbb{K}^{m\times n}$. This vector space may be identified with the vector space of mn-vectors, $\mathbb{K}^{n\times n} \cong \mathbb{K}^{mn}$. Hence, all statements for vector norms carry over to norms for matrices. In particular, all norms for $m \times n$-matricen are equivalent and the convergence of sequences of matrices is again the componentwise convergence

$$A^k \to A \,(k \to \infty) \quad \Longleftrightarrow \quad a_{ij}^k \to a_{ij} \,(k \to \infty)\,, \quad i = 1, \ldots, m\,, j = 1, \ldots, n\,.$$

Now, we restrict the further discussion to square matrices $A \in \mathbb{K}^{n\times n}$. For an arbitrary vector norm $\|\cdot\|$ on $\mathbb{K}^n$ a norm for matrices $A \in \mathbb{K}^{n\times n}$ is generated by

$$\|A\| := \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{K}^n, \|x\|=1} \|Ax\|.$$

The definiteness and homogeneity are obvious and the triangle inequality follows from that holding for the given vector norm. This matrix norm is called the "natural matrix norm" corresponding to the vector norm $\|\cdot\|$. In the following for both norms, the matrix norm and the generating vector norm, the same notation is used. For a natural matrix norm there always holds $\|I\| = 1$. Such a "natural" matrix norm is automatically "compatible" with the generating vector norm, i. e., it satisfies

$$\|Ax\| \le \|A\|\,\|x\|\,, \quad x \in \mathbb{K}^n,\ A \in \mathbb{K}^{n\times n}\,. \tag{1.1.33}$$

Further it is "submultiplicative":

$$\|AB\| \le \|A\|\,\|B\|\,, \quad A, B \in \mathbb{K}^{n\times n}. \tag{1.1.34}$$

Not all matrix norms are "natural" in the above sense. For instance, the square-sum norm (also called "Frobenius[9]-norm")

$$\|A\|_F := \Big(\sum_{j,k=1}^{n} |a_{jk}|^2\Big)^{1/2}$$

is compatible with the euclidian norm and submultiplicative but cannot be a *natural* matrix norm since $\|I\|_F = \sqrt{n}$ (for $n \ge 2$). The natural matrix norm generated from the euclidian vector norm is called "spectral norm". This name is suggested by the following result.

[9]Ferdinand Georg Frobenius (1849-1917): German mathematician; prof. in Zurich and Berlin; contributions to the theory of differential equations, to determinants and matrices as well as to group theory.

Lemma 1.13 (Spectral norm): *For an arbitrary square matrix $A \in \mathbb{K}^{n\times n}$ the product matrix $\overline{A}^T A \in \mathbb{K}^{n\times n}$ is always hermitian and positive semi-definitsemi-definite. For the spectral norm of A there holds:*

$$\|A\|_2 = \max\{|\lambda|^{1/2},\, \lambda \in \sigma(\bar{A}^T A)\}. \tag{1.1.35}$$

If A is hermitian (or symmetric), then:

$$\|A\|_2 = \max\{|\lambda|,\, \lambda \in \sigma(A)\}. \tag{1.1.36}$$

Proof. (i) Let the matrix $A \in \mathbb{K}^{n\times n}$ be hermitian. For any eigenvalue λ of A and corresponding eigenvector x there holds

$$|\lambda| = \frac{\|\lambda x\|_2}{\|x\|_2} = \frac{\|Ax\|_2}{\|x\|_2} \le \|A\|_2.$$

Conversely, let $\{a^i, i = 1, \cdots, n\} \subset \mathbb{C}^n$ be an ONB of eigenvectors of A and $x = \sum_i x_i a^i \in \mathbb{C}^n$ be arbitrary. Then,

$$\|Ax\|_2 = \big\|A\big(\sum\nolimits_i x_i a^i\big)\big\|_2 = \big\|\sum\nolimits_i \lambda_i x_i a^i\big\|_2 \le \max_i |\lambda_i|\, \big\|\sum\nolimits_i x_i a^i\big\|_2 = \max_i |\lambda_i|\, \|x\|_2,$$

and consequently,

$$\frac{\|Ax\|_2}{\|x\|_2} \le \max_i |\lambda_i|.$$

(ii) For a general matrix $A \in \mathbb{K}^{n\times n}$ there holds

$$\|A\|_2^2 = \max_{x\in C^n\setminus 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \max_{x\in C^n\setminus 0} \frac{(\bar{A}^T Ax, x)_2}{\|x\|_2^2} \le \max_{x\in C^n\setminus 0} \frac{\|\bar{A}^T Ax\|_2}{\|x\|_2} = \|\bar{A}^T A\|_2.$$

and $\|\bar{A}^T A\|_2 \le \|\bar{A}^T\|_2 \|A\|_2 = \|A\|_2^2$ (observe that $\|A\|_2 = \|\bar{A}^T\|_2$ due to $\|Ax\|^2 = \|\bar{A}^T \bar{x}\|^2$). This completes the proof. Q.E.D.

Lemma 1.14 (Natural matrix norms): *The natural matrix norms generated by the l_∞ norm $\|\cdot\|_\infty$ and the l_1 Norm $\|\cdot\|_1$ are the so-called "maximal-row-sum norm"*

$$\|A\|_\infty := \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|, \tag{1.1.37}$$

and the "maximal-column-sum norm"

$$\|A\|_1 := \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|\,. \tag{1.1.38}$$

Proof. We give the proof only for the l_∞ norm. For the l_1 norm the argument is analogous.
(i) The maximal row sum $\|\cdot\|_\infty$ is a matrix norm. The norm properties (N1) - (N3) follow from

the corresponding properties of the modulus. For the matrix product AB there holds

$$\|AB\|_\infty = \max_{1\le i\le n}\Big|\sum_{j=1}^n\Big(\sum_{k=1}^n a_{ik}b_{kj}\Big)\Big| \le \max_{1\le i\le n}\sum_{k=1}^n\Big(|a_{ik}|\sum_{j=1}^n|b_{kj}|\Big)$$
$$\le \max_{1\le i\le n}\sum_{k=1}^n|a_{ik}|\ \max_{1\le k\le n}\sum_{j=1}^n|b_{kj}| = \|A\|_\infty\|B\|_\infty.$$

(ii) Further, in view of

$$\|Ax\|_\infty = \max_{1\le j\le n}|\sum_{k=1}^n a_{jk}x_k| \le \max_{1\le j\le n}\sum_{k=1}^n|a_{jk}|\ \max_{1\le k\le n}|x_k| = \|A\|_\infty\|x\|_\infty$$

the maximal row-sum is compatible with the maximum norm $\|\cdot\|_\infty$ and there holds

$$\sup_{\|x\|_\infty=1}\|Ax\|_\infty \le \|A\|_\infty.$$

(iii) In the case $\|A\|_\infty = 0$ also $A = 0$, i. e.,

$$\|A\|_\infty = \sup_{\|x\|_\infty=1}\|Ax\|_\infty.$$

Therefore, let $\|A\|_\infty > 0$ and $m \in \{1,\dots,n\}$ an index such that

$$\|A\|_\infty = \max_{1\le j\le n}\sum_{k=1}^n|a_{jk}| = \sum_{k=1}^n|a_{mk}|.$$

For $k = 1,\dots,n$, we set

$$z_k := \begin{cases} |a_{mk}|/a_{mk} & \text{für } a_{mk}\neq 0,\\ 0, & \text{sonst,}\end{cases}$$

i. e., $z = (z_k)_{k=1}^n \in \mathbb{K}^n$, $\|z\|_\infty = 1$. For $v := Az$ it follows that

$$v_m = \sum_{k=1}^n a_{mk}z_k = \sum_{k=1}^n|a_{mk}| = \|A\|_\infty.$$

Consequently,

$$\|A\|_\infty = v_m \le \|v\|_\infty = \|Az\|_\infty \le \sup_{\|y\|_\infty=1}\|Ay\|_\infty,$$

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

Let $\|\cdot\|$ be an arbitrary vector norm and $\|\cdot\|$ a corresponding compatible matrix norm. Then, with a normalized eigenvector $\|w\| = 1$ corresponding to the eigenvalue λ there holds

$$|\lambda| = |\lambda|\,\|w\| = \|\lambda w\| = \|Aw\| \le \|A\|\,\|w\| = \|A\|, \tag{1.1.39}$$

i. e., all eigenvalues of A are contained in a circle in $\mathbb{C}$ with center at the origin and radius $\|A\|$. Especially with $\|A\|_\infty$, we obtain the eigenvalue bound

$$\max_{\lambda\in\sigma(A)} |\lambda| \le \|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|. \tag{1.1.40}$$

Since the eigenvalues of $\bar{A}^T$ and A are related by $\lambda(\bar{A}^T) = \bar{\lambda}(A)$, using the bound (1.1.40) simultaneously for $\bar{A}^T$ and A yields the refined bound

$$\begin{aligned} \max_{\lambda\in\sigma(A)} |\lambda| &\le \min\{\|A\|_\infty, \|\bar{A}^T\|_\infty\} \\ &= \min\Big\{ \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|, \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}| \Big\}. \end{aligned} \tag{1.1.41}$$

The following lemma contains a useful result on the regularity of small perturbations of the unit matrix.

Lemma 1.15 (Perturbation of unity): *Let $\|\cdot\|$ be any natural matrix norm on $\mathbb{K}^{n\times n}$ and $B \in \mathbb{K}^{n\times n}$ a matrix with $\|B\| < 1$. Then, the perturbed matrix $I + B$ is regular and its inverse is given as the "Neumann[10] series"*

$$(I + B)^{-1} = \sum_{k=0}^{\infty} B^k. \tag{1.1.42}$$

Further, there holds

$$\|(I + B)^{-1}\| \le \frac{1}{1 - \|B\|}. \tag{1.1.43}$$

Proof. (i) First, we show the regularity of $I + B$ and the bound (1.1.43). For all $x \in \mathbb{K}^n$ there holds

$$\|(I + B)x\| \ge \|x\| - \|Bx\| \ge \|x\| - \|B\|\|x\| = (1 - \|B\|)\|x\|.$$

In view of $1 - \|B\| > 0$ this implies that $I + B$ is injective and consequently regular. Then, the following estimate implies (1.1.43):

$$\begin{aligned} 1 = \|I\| \;&=\; \|(I + B)(I + B)^{-1}\| = \|(I + B)^{-1} + B(I + B)^{-1}\| \\ &\ge \|(I + B)^{-1}\| - \|B\|\,\|(I + B)^{-1}\| = \|(I + B)^{-1}\|(1 - \|B\|) > 0. \end{aligned}$$

(ii) Next, we define

$$S := \lim_{k\to\infty} S_k, \quad S_k = \sum_{s=0}^{k} B^s.$$

[10] *Carl Gottfried Neumann (1832-1925): German mathematician; since 1858 "Privatdozent" and since 1863 apl. prof. in Halle. After holding professorships in Basel and Tübingen he moved 1868 to Leipzig where he worked for more than 40 years. He contributed to the theory of (partial) differential and integral equations, especially to the Dirichlet problem. The "Neumann boundary condition" and the "Neumann series" are named after him. In mathematical physics he worked on analytical mechanics and potential theory. together with A. Clebsch he founded the journal "Mathematische Annalen".*

S is well defined due to the fact that $\{S_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence with respect to the matrix norm $\|\cdot\|$ (and, by the norm equivalence in finite dimensional normed spaces, with respect to any matrix norm). By employing the triangle inequality, using the matrix norm property and the limit formula for the geometric series, we see that

$$\|S\| = \lim_{k\to\infty} \|S_k\| = \lim_{k\to\infty} \Big\| \sum_{s=0}^{k} B^s \Big\| \leq \lim_{n\to\infty} \sum_{s=0}^{k} \|B\|^s = \lim_{k\to\infty} \frac{1-\|B\|^{k+1}}{1-\|B\|} = \frac{1}{1-\|B\|}.$$

Furthermore, $S_k\,(I-B) = I - B^{k+1}$ and due to the fact that multiplication with $I-B$ is continuous,

$$I = \lim_{k\to\infty} \big(S_k(I-B)\big) = \big(\lim_{k\to\infty} S_k\big)(I-B) = S(I-B).$$

Hence, $S = (I-B)^{-1}$ and the proof is complete. Q.E.D.

Corollary 1.2: *Let $A \in \mathbb{K}^{n\times n}$ be a regular matrix and $\tilde{A}$ another matrix such that*

$$\|\tilde{A} - A\| < \frac{1}{\|A^{-1}\|}. \tag{1.1.44}$$

Then, also $\tilde{A}$ is regular. This means that the "resolvent set" $\mathrm{Res}(A)$ of a matrix $A \in \mathbb{K}^{n\times n}$ is open in $\mathbb{K}^{n\times n}$ and the only "singular" points are just the eigenvalues of A, i. e., there holds $\mathbb{C} = \mathrm{Res}(A) \cup \sigma(A)$.

Proof. Notice that $\tilde{A} = A + \tilde{A} - A = A(I + A^{-1}(\tilde{A} - A))$. In view of

$$\|A^{-1}(\tilde{A} - A)\| \leq \|A^{-1}\|\,\|\tilde{A} - A\| < 1$$

by Lemma 1.15 the matrix $I + A^{-1}(\tilde{A} - A)$ is regular. Then, also the product $A(I + A^{-1}(\tilde{A} - A))$ is regular, which implies the regularity of $\tilde{A}$. Q.E.D.

1.2 Spectra and pseudo-spectra of matrices

1.2.1 Stability of dynamical systems

We consider a finite dimensional dynamical system of the form

$$u'(t) = F(t, u(t)), \quad t \geq 0, \quad u(0) = u^0, \tag{1.2.45}$$

where $u : [0,\infty) \to \mathbb{R}^n$ is a continuously differentiable vector function and the system function $F(\cdot,\cdot)$ is assumed (for simplicity) to be defined on all of $\mathbb{R}\times\mathbb{R}^n$ and twice continuously differentiable. The system (1.2.45) may originate from the discretization of an infinite dimensional dynamical system such as the nonstationary Navier-Stokes equations mentioned in the introductory Chapter 0. Suppose that u is a particular solution of (1.2.45). We want to investigate its stability against small perturbations $u(t_0) \to u(t_0) + w^0 =: v(t_0)$ at any time $t_0 \geq 0$. For this, we use the strongest concept of stability, which is suggested by the corresponding properties of solutions of the Navier-Stokes equations.

Definition 1.9: *The solution* $u \in C^1[0,\infty;\mathbb{R}^n)$ *of (1.2.45) is called "exponentially stable" if there are constants* $\delta, K, \kappa \in \mathbb{R}_+$ *such that for any perturbation* $w^0 \in \mathbb{R}^n$, $\|w^0\|_2 \le \delta$, *at any time* $t_0 \ge 0$, *there exists a secondary solution* $v \in C^1(t_0,\infty;\mathbb{R}^n)$ *of the perturbed system*

$$v'(t) = F(t,v(t)), \quad t \ge 0, \quad v(t_0) = u(t_0) + w^0, \tag{1.2.46}$$

and there holds

$$\|v(t) - u(t)\|_2 \le K e^{-\kappa(t-t_0)} \|w^0\|_2, \quad t \ge t_0. \tag{1.2.47}$$

For simplicity, we restrict the following discussion to the special situation of an autonomous system, i. e., $F(t,\cdot) \equiv F(\cdot)$ and a stationary particular solution $u(t) \equiv u \in \mathbb{R}^n$, i. e., to the solution of the nonlinear system

$$F(u) = 0. \tag{1.2.48}$$

The investigation of the stability of u leads us to consider the so-called "perturbation equation" for the perturbation $w(t) := v(t) - u$,

$$w'(t) = F(v(t)) - F(u) = F'(u)w(t) + \mathcal{O}(\|w(t)\|_2^2), \;\; t \ge 0, \quad w(0) = w^0, \tag{1.2.49}$$

where the higher-order term depends on bounds on u and u' as well as on the smoothness properties of $F(\cdot)$.

Theorem 1.6: *Suppose that the Jacobian* $A := F'(u)$ *is diagonalizable and that all its eigenvalues have real parts less than zero. Then, the solution* u *of (1.2.48) is exponentially stable in the sense of Definition 1.9 with the constants* $\kappa = -\mathrm{Re}\lambda_{\min}$ *and* $K = \mathrm{cond}_2(W)$, *where* $\lambda_{\min}$ *is the eigenvalue of* A *with smallest real part and* $W = [w^1,\dots,w^n]$ *the column matrix formed by the (normalized) eigenbasis of* A. *If* A *is normal then* $K = \mathrm{cond}_2(W) = 1$.

Proof. (i) Consider the linearized system (linearized perturbation equation)

$$w'(t) = Aw(t), \;\; t \ge t_0, \quad w(0) = w^0. \tag{1.2.50}$$

Since the Jacobian A is diagonalizable there exists an ONB $\{w^1,\dots,w^n\}$ of eigenvectors of A:

$$Aw^i = \lambda_i w^i, \quad i = 1,\dots,n.$$

With the matrices $W := [w^1,\dots,w^n]$ and $\Lambda := \mathrm{diag}(\lambda_i)$ there holds

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

Using this notation the perturbation equation can be rewritten in the form

$$w'(t) = Aw(t) \quad \Leftrightarrow \quad w'(t) = W\Lambda W^{-1}w(t) \quad \Leftrightarrow \quad (W^{-1}w)'(t) = \Lambda W^{-1}w(t),$$

or for the transformed variable $v := W^{-1}w$ componentwise:

$$v_i'(t) = \lambda_i v_i(t), \;\; t \ge 0, \quad v_i(0) = (W^{-1}w)_i(0).$$

The solution behavior is (observe that $e^{i\mathrm{Im}\lambda_i t} = 1$)

$$|v_i(t)| \le e^{\mathrm{Re}\lambda_i t}|(W^{-1}w)_i(0)|, \quad t \ge 0.$$

This implies:

$$\|v(t)\|_2^2 \le \sum_{i=1}^{n} |v_i(t)|^2 \le \sum_{i=1}^{n} e^{2\mathrm{Re}\lambda_i t}|(W^{-1}w)_i(0)|^2 \le e^{2\mathrm{Re}\lambda_{\min} t}\|(W^{-1}w)(0)\|_2^2,$$

and consequently,

$$\begin{aligned}\|w(t)\|_2 &\le \|Wv(t)\|_2 \le \|W\|_2\|v(t)\|_2 \le \|W\|_2\, e^{\mathrm{Re}\lambda_{\min} t}\,\|(W^{-1}w)(0)\|_2 \\ &\le \|W\|_2\, e^{\mathrm{Re}\lambda_{\min} t}\,\|W^{-1}\|_2\|w(0)\|_2 \\ &= \mathrm{cond}_2(W)\, e^{\mathrm{Re}\lambda_{\min} t}\|w(0)\|_2.\end{aligned} \tag{1.2.51}$$

The condition number of W can become arbitrarily large depending on the "non-orthogonality" of the eigenbasis of the Jacobian A.

(ii) The assertion now follows by combining (1.2.51) and (1.2.49) within a continuation argument. The proof is complete. Q.E.D.

Following the argument in the proof of Theorem 1.6, we see that the occurrence of just one eigenvalue with $\mathrm{Re}\,\lambda > 0$ inevitably causes dynamic instability of the solution u, i. e., arbitrarily small perturbations may grow in time without bound. Denoting by $S : \mathbb{R}^n \to C^1[0,\infty;\mathbb{R}^n)$ the "solution operator" of the *linearized* perturbation equation (1.2.50), i. e., $w(t) = S(t)w^0$, this can be formulated as

$$\max_{\lambda\in\sigma(A)} \mathrm{Re}\,\lambda > 0 \quad \Rightarrow \quad \sup_{t\ge 0}\|S(t)\|_2 = \infty. \tag{1.2.52}$$

The result of Theorem 1.6 can be extended to the case of a non-diagonalizable Jacobian $A = F'(u)$. In this case, one obtains a stability behavior of the form

$$\|S(t)\|_2 \approx K(1+t^\alpha)e^{\mathrm{Re}\lambda_{\min} t}, \quad t \ge 0, \tag{1.2.53}$$

where $\alpha \ge 1$ is the defect of the most critical eigenvalue $\lambda_{\min}$, i. e., that eigenvalue with largest real part $\mathrm{Re}\lambda_{\min} < 0$. This implies that

$$\sup_{t>0}\|S(t)\| \approx \Big(\frac{\alpha}{e}\Big)^\alpha \frac{1}{|\,\mathrm{Re}\,\lambda_{\min}|^\alpha}, \tag{1.2.54}$$

i. e., for $-1 \ll \mathrm{Re}\,\lambda_{\min} < 0$ initially small perturbations may grow beyond a value at which nonlinear instability is triggered. Summarizing, we are interested in the case that all eigenvalues of $A = F'(u)$ have negative real part, suggesting stability in the sense of Theorem 1.6, and especially want to compute the most "critical" eigenvalue, i. e., that $\lambda \in \sigma(A)$ with maximal $\mathrm{Re}\,\lambda < 0$ to detect whether the corresponding solution operator $S(t)$ may behave in a critical way.

The following result, which is sometimes addressed as the "easy part of the Kreiss matrix theorem" indicates in which direction this analysis has to go.

Lemma 1.16: *Let $A := F'(u)$ and $z \in \mathbb{C}\setminus\sigma(A)$ with $\mathrm{Re}\,z > 0$. Then, for the solution*

operator $S(t)$ *of the linearized perturbation equation (1.2.50), there holds*

$$\sup_{t\geq 0} \|S(t)\|_2 \geq |\operatorname{Re} z|\, \|(zI-A)^{-1}\|_2. \tag{1.2.55}$$

Proof. We continue using the notation from the proof of Theorem 1.6. If $\|S(t)\|_2$ is unbounded over $[0,\infty)$, the asserted estimate holds trivially. Hence, let us assume that

$$\sup_{t\geq 0} \|w(t)\|_2 = \sup_{t\geq 0} \|S(t)w^0\|_0 \leq \sup_{t\geq 0} \|S(t)\|_2 \|w^0\|_2 < \infty.$$

For $z \notin \sigma(A)$ the resolvent $R_A(z) = zI - A$ is regular. Let $w^0 \in \mathbb{K}^n$ be an arbitrary but nontrivial initial perturbation and $w(t) = S(t)w^0$. We rewrite equation (1.2.50) in the form

$$\partial_t w - zw + (zI - A)w = 0,$$

and multiply by e^{-tz}, to obtain

$$\partial_t (e^{-tz} w) + e^{-tz}(zI - A)w = 0.$$

Next, integrating this over $0 \leq t < T$ and observing $\operatorname{Re} z > 0$ and $\lim_{t\to\infty} e^{-tz} w = 0$ yields

$$-(zI-A)^{-1} w^0 = \Big(\int_0^\infty e^{-tz} S(t)\, dt \Big)\, w^0.$$

From this, we conclude

$$\|(zI-A)^{-1}\|_2 \leq \Big(\int_0^\infty e^{-t|\operatorname{Re} z|}\, dt \Big) \sup_{t>0} \|S(t)\|_2 \leq |\operatorname{Re} z|^{-1} \sup_{t>0} \|S(t)\|_2,$$

which implies the asserted estimate. Q.E.D.

The above estimate (1.2.55) for the solution operator $S(t)$ can be interpreted as follows: Even if all eigenvalues of the matrix A have negative real parts, which in view of Theorem 1.6 would indicate stability of solutions to (1.2.50), there may be points z in the right complex half plane for which $\|(zI-A)^{-1}\|_2 \gg |\operatorname{Re} z|^{-1}$ and consequently,

$$\sup_{t\geq 0} \|S(t)\|_2 \gg 1. \tag{1.2.56}$$

Hence, even small perturbations of the particular solution u may be largely amplified eventually triggering nonlinear instability.

1.2.2 Pseudospectrum of a matrix

The estimate (1.2.55) makes us search for points $z \in \mathbb{C} \setminus \sigma(A)$ with $\operatorname{Re} z > 0$ and

$$\|(zI-A)^{-1}\|_2 \gg |\operatorname{Re} z|^{-1}.$$

This suggests the concept of the "pseudospectrum" of the matrix A, which goes back to Landau [6] and has been extensively described and applied in the stability analysis of dynamical systems, e. g., in Trefethen [17] and Trefethen & Embree [19].

Definition 1.10 (Pseudospectrum): *For $\varepsilon \in \mathbb{R}_+$ the "ε-pseudo-spectrum" $\sigma_\varepsilon(A) \subset \mathbb{C}$ of a matrix $A \in \mathbb{K}^{n\times n}$ is defined by*

$$\sigma_\varepsilon(A) := \big\{ z \in \mathbb{C} \setminus \sigma(A) \big|\ \|(A - zI)^{-1}\|_2 \geq \varepsilon^{-1} \big\} \cup \sigma(A). \tag{1.2.57}$$

Remark 1.7: The concept of a pseudospectrum is interesting only for non-normal operators, since for a normal operator $\sigma_\varepsilon(A)$ is just the union of ε-circles around its eigenvalues. This follows from the estimate (see Dunford & Schwartz [5] or Kato [9])

$$\|(A - zI)^{-1}\|_2 \geq \operatorname{dist}(z, \sigma(A))^{-1}, \quad z \notin \sigma(A), \tag{1.2.58}$$

where equality holds if A is normal.

Remark 1.8: The concept of the "pseudospectrum" can be introduced in much more general situations, such as that of closed linear operators in abstract Hilbert or Banach spaces (see Trefethen & Embree [19]). Typically hydrodynamic stability analysis concerns differential operators defined on bounded domains. This situation fits into the Hilbert-space framework of "closed unbounded operators with compact inverse".

Using the notion of the pseudospectrum the estimate (1.2.55) can be expressed in the following form

$$\sup_{t\geq 0} \|S(t)\|_2 \geq \sup \Big\{ \frac{|\operatorname{Re} z|}{\varepsilon} \ \Big|\ \varepsilon > 0, z \in \sigma_\varepsilon(A),\ \operatorname{Re} z > 0 \Big\}, \tag{1.2.59}$$

or

$$\max_{\lambda\in\sigma_\varepsilon(A)} \operatorname{Re}\lambda > K\varepsilon \quad \Rightarrow \quad \sup_{t\geq 0} \|S(t)\|_2 > K. \tag{1.2.60}$$

Below, we will present methods for computing estimates to the pseudospectrum of a matrix. This will be based on related methods for solving the partial eigenvalue problem. To this end, we provide some results on several basic properties of the pseudospectrum.

Lemma 1.17: *(i) For a matrix $A \in \mathbb{K}^{n\times n}$ the following definitions of the ε-pseudospectrum are equivalent:*

$$\begin{aligned}
&(a) \quad \sigma_\varepsilon(A) := \big\{ z \in \mathbb{C} \setminus \sigma(A) \big|\ \|(A - zI)^{-1}\|_2 \geq \varepsilon^{-1} \big\} \cup \sigma(A). \\
&(b) \quad \sigma_\varepsilon(A) := \big\{ z \in \mathbb{C} \big|\ z \in \sigma(A + E) \ \text{for some}\ E \in \mathbb{K}^{n\times n} \ \text{with}\ \|E\|_2 \leq \varepsilon \big\}. \\
&(c) \quad \sigma_\varepsilon(A) := \big\{ z \in \mathbb{C} \big|\ \|(A - zI)v\|_2 \leq \varepsilon \ \text{for some}\ v \in \mathbb{K}^n \ \text{with}\ \|v\|_2 = 1 \big\}.
\end{aligned}$$

(ii) Let $0 \notin \sigma(A)$. Then, the ε-pseudospectra of A and that of its inverse A^{-1} are related by

$$\sigma_\varepsilon(A) \subset \big\{ z \in \mathbb{C} \setminus \{0\} \ \big|\ z^{-1} \in \sigma_{\delta(z)}(A^{-1}) \big\} \cup \{0\}, \tag{1.2.61}$$

where $\delta(z) := \varepsilon\|A^{-1}\|_2/|z|$ and, for $0 < \varepsilon < 1$, by

$$\sigma_\varepsilon(A^{-1}) \cap B_1(0)^c \subset \big\{ z \in \mathbb{C} \setminus \{0\} \ \big|\ z^{-1} \in \sigma_\delta(A) \big\}, \tag{1.2.62}$$

where $B_1(0) := \{z \in \mathbb{C},\ |z| \leq 1\}$ and $\delta := \varepsilon/(1 - \varepsilon)$.

Proof. The proof of part (i) can be found in Trefethen & Embree [19]. For completeness, we recall a sketch of the argument. The proof of part (ii) is taken from Gerecht et al. [32].

(ia) In all three definitions, we have $\sigma(A) \subset \sigma_\varepsilon(A)$. Let $z \in \sigma_\varepsilon(A)$ in the sense of definition (a). There exists a $w \in \mathbb{K}^n$ with $\|w\|_2 = 1$, such that $\|(zI - A)^{-1}w\|_2 \geq \varepsilon^{-1}$. Hence, there is a $v \in \mathbb{K}^n$ with $\|v\|_2 = 1$, and $s \in (0, \varepsilon)$, such that $(zI - A)^{-1}w = s^{-1}v$ or $(zI - A)v = sw$. Let $Q(v, w) \in \mathbb{K}^{n\times n}$ denote the unitary matrix, which rotates the unit vector v into the unit vector w, such that $sw = sQ(v, w)v$. Then, $z \in \sigma(A + E)$ where $E := sQ(v, w)$ with $\|E\|_2 \leq \varepsilon$, i.e., $z \in \sigma_\varepsilon(A)$ in the sense of definition (b). Let now be $z \in \sigma_\varepsilon(A)$ in the sense of definition (b), i.e., there exists $E \in \mathbb{K}^{n\times n}$ with $\|E\|_2 \leq \varepsilon$ such that $(A + E)w = zw$, with some $w \in \mathbb{K}^n$, $w \neq 0$. Hence, $(A - zI)w = -Ew$, and therefore,

$$\begin{aligned}\|(A - zI)^{-1}\|_2 &= \sup_{v\in\mathbb{K}^n\setminus\{0\}} \frac{\|(A - zI)^{-1}v\|_2}{\|v\|_2} = \sup_{v\in\mathbb{K}^n\setminus\{0\}} \frac{\|v\|_2}{\|(A - zI)v\|_2} \\ &= \Big(\inf_{v\in\mathbb{K}^n\setminus\{0\}} \frac{\|(A - zI)v\|_2}{\|v\|_2}\Big)^{-1} \geq \Big(\frac{\|(A - zI)w\|_2}{\|w\|_2}\Big)^{-1} \\ &= \Big(\frac{\|Ew\|_2}{\|w\|_2}\Big)^{-1} \geq \|E\|_2^{-1} \geq \varepsilon^{-1}.\end{aligned}$$

Hence, $z \in \sigma_\varepsilon(A)$ in the sense of (a). This proves the equivalence of definitions (a) and (b).

(ib) Next, let again $z \in \sigma_\varepsilon(A) \setminus \sigma(A)$ in the sense of definition (a). Then,

$$\varepsilon \geq \|(A - zI)^{-1}\|_2^{-1} = \Big(\sup_{w\in\mathbb{K}^n\setminus\{0\}} \frac{\|(A - zI)^{-1}w\|_2}{\|w\|_2}\Big)^{-1} = \inf_{v\in\mathbb{K}^n\setminus\{0\}} \frac{\|(A - zI)v\|_2}{\|v\|_2}.$$

Hence, there exists a $v \in \mathbb{K}^n$ with $\|v\|_2 = 1$, such that $\|(A - zI)v\| \leq \varepsilon$, i.e., $z \in \sigma_\varepsilon(A)$ in the sense of definition (c). By the same argument, now used in the reversed direction, we see that $z \in \sigma_\varepsilon(A)$ in the sense of definition (c) implies that also $z \in \sigma_\varepsilon(A)$ in the sense of definition (a). Thus, definition (a) is also equivalent to condition (c).

(iia) We use the definition (c) from part (i) for the ε-pseudospectrum. Let $z \in \sigma_\varepsilon(A)$ and accordingly $v \in \mathbb{K}^n$, $\|v\|_2 = 1$, satisfying $\|(A - zI)v\|_2 \leq \varepsilon$. Then,

$$\|(A^{-1} - z^{-1}I)v\|_2 = \|z^{-1}A^{-1}(zI - A)v)\|_2 \leq |z|_2^{-1}\|A^{-1}\|_2\,\varepsilon.$$

This proves the asserted relation (1.2.61).

(iib) To prove the relation (1.2.62), we again use the definition (c) from part (i) for the ε-pseudospectrum. Accordingly, for $z \in \sigma_\varepsilon(A^{-1})$ with $|z| \geq 1$ there exists a unit vector $v \in \mathbb{K}^n$, $\|v\|_2 = 1$, such that

$$\varepsilon \geq \|(zI - A^{-1})v\|_2 = |z|\|(A - z^{-1}I)A^{-1}v\|_2.$$

Then, setting $w := \|A^{-1}v\|_2^{-1}A^{-1}v$ with $\|w\|_2 = 1$, we obtain

$$\|(A - z^{-1}I)w\|_2 \leq |z|^{-1}\|A^{-1}v\|_2^{-1}\varepsilon.$$

Hence, observing that

$$\|A^{-1}v\|_2 = \|(A^{-1} - zI)v + zv\|_2 \geq \|zv\|_2 - \|(A^{-1} - zI)v\|_2 \geq |z| - \varepsilon,$$

we conclude that

$$\|(A - z^{-1}I)w\|_2 \leq \frac{\varepsilon}{|z|(|z| - \varepsilon)} \leq \frac{\varepsilon}{1 - \varepsilon}.$$

This completes the proof. Q.E.D.

The next proposition relates the size of the resolvent norm $\|(zI - A)^{-1}\|_2$ to easily computable quantities in terms of the eigenvalues and eigenfunctions of the matrix $A = F'(u)$.

Theorem 1.7: *Let $\lambda \in \mathbb{C}$ be a non-deficient eigenvalue of the matrix $A := F'(u)$ with corresponding primal and dual eigenvectors $v, v^* \in \mathbb{K}^n$ normalized by $\|v\|_2 = (v, v^*)_2 = 1$. Then, there exists a continuous function $\omega : \mathbb{R}_+ \to \mathbb{C}$ with $\lim_{\varepsilon \searrow 0_+} \omega(\varepsilon) = 1$, such that for $\lambda_\varepsilon := \lambda - \varepsilon\omega(\varepsilon)\|v^*\|_2$, there holds*

$$\|(A - \lambda_\varepsilon I)^{-1}\|_2 \geq \varepsilon^{-1}, \tag{1.2.63}$$

i.e., the point λ_ε lies in the ε-pseudospectrum of the matrix A.

Proof. The argument of the proof is recalled from Gerecht et. al. [32] where it is developed within a function space setting and has therefore to be simplified here for the finite dimensional situation.

Let $B \in \mathbb{K}^{n \times n}$ be a matrix with $\|B\|_2 \leq 1$. We consider the perturbed eigenvalue problem

$$(A + \varepsilon B)v_\varepsilon = \lambda_\varepsilon v_\varepsilon. \tag{1.2.64}$$

Since this is a regular perturbation and λ non-deficient, there exist corresponding eigenvalues $\lambda_\varepsilon \in \mathbb{C}$ and eigenvectors $v_\varepsilon \in \mathbb{K}^n$, $\|v_\varepsilon\|_2 = 1$, such that

$$|\lambda_\varepsilon - \lambda| = \mathcal{O}(\varepsilon), \quad \|v_\varepsilon - v\|_2 = \mathcal{O}(\varepsilon).$$

Furthermore, from the relation

$$(Av - \lambda_\varepsilon I)v_\varepsilon = -\varepsilon B v_\varepsilon, \quad \varphi \in \mathbf{J}_1,$$

we conclude that

$$\|(A - \lambda_\varepsilon I)v_\varepsilon\|_2 \leq \varepsilon\|B\|_2\|v_\varepsilon\|_2 \leq \varepsilon\|v_\varepsilon\|_2,$$

and from this, if λ_ε is not an eigenvalue of A,

$$\begin{aligned}\|(A - \lambda_\varepsilon I)^{-1}\|_2^{-1} &= \Big(\sup_{y \in \mathbb{K}^n} \frac{\|(A - \lambda_\varepsilon I)^{-1}y\|_2}{\|y\|_2}\Big)^{-1} = \Big(\sup_{x \in \mathbb{K}^n} \frac{\|x\|_2}{\|(A - \lambda_\varepsilon I)x\|_2}\Big)^{-1} \\ &= \inf_{x \in \mathbb{K}^n} \frac{\|(A - \lambda_\varepsilon I)x\|_2}{\|x\|_2} \leq \frac{\|(A - \lambda_\varepsilon I)v_\varepsilon\|_2}{\|v_\varepsilon\|_2} \leq \varepsilon.\end{aligned}$$

This implies the asserted estimate

$$\|(A - \lambda_\varepsilon I)^{-1}\|_2 \geq \varepsilon^{-1}. \tag{1.2.65}$$

(ii) Next, we analyze the dependence of the eigenvalue λ_ε on ε in more detail. Subtracting the equation for v from that for v_ε, we obtain

$$A(v_\varepsilon - v) + \varepsilon B v_\varepsilon = (\lambda_\varepsilon - \lambda)v_\varepsilon + \lambda(v_\varepsilon - v).$$

Multiplying this by v^* yields

$$(A(v_\varepsilon - v), v^*)_2 + \varepsilon(Bv_\varepsilon, v^*)_2 = (\lambda_\varepsilon - \lambda)v_\varepsilon, v^*)_2 + \lambda(v_\varepsilon - v, v^*)_2$$

and, using the equation satisfied by v^*,

$$\varepsilon(Bv_\varepsilon, v^*)_2 = (\lambda_\varepsilon - \lambda)(v_\varepsilon, v^*)_2.$$

This yields $\lambda_\varepsilon = \lambda + \varepsilon\omega(\varepsilon)(Bv, v^*)_2$, where, observing $v_\varepsilon \to v$ and $(v, v^*) = 1$,

$$\omega(\varepsilon) := \frac{(Bv_\varepsilon, v^*)_2}{(v_\varepsilon, v^*)_2(Bv, v^*)_2} \to 1 \quad (\varepsilon \to 0).$$

(iii) It remains to construct an appropriate perturbation matrix B. For convenience, we consider the renormalized dual eigenvectors $\tilde{v}^* := v^*\|v^*\|_2^{-1}$, satisfying $\|\tilde{v}^*\|_2 = 1$. With the vector $w := (v - \tilde{v}^*)\|v - \tilde{v}^*\|_2^{-1}$, we set for $\psi \in \mathbb{K}^n$:

$$S\psi := \psi - 2\,\mathrm{Re}\,(\psi, w)_2 w, \qquad B := -S.$$

The unitary matrix S acts like a Householder transformation mapping v into $\tilde{v}^*$ (s. the discussion in Section 2.3.1). In fact, observing $\|v\|_2 = \|\tilde{v}^*\|_2 = 1$, there holds

$$\begin{aligned} Sv = v - \frac{2\,\mathrm{Re}\,(v, v - \tilde{v}^*)_2}{\|v - \tilde{v}^*\|_2^2}(v - \tilde{v}^*) &= \frac{\{2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2\}v - 2\,\mathrm{Re}\,(v, v - \tilde{v}^*)_2(v - \tilde{v}^*)}{2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2} \\ = \frac{2v - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2 v - 2v + 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2 v + (2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2)\tilde{v}^*}{2 - 2\,\mathrm{Re}\,(v, \tilde{v}^*)_2} &= \tilde{v}^*. \end{aligned}$$

This implies that

$$(Bv, v^*)_2 = -(Sv, v^*)_2 = -(\tilde{v}^*, v^*)_2 = -\|v^*\|_2.$$

Further, observing $\|w\|_2 = 1$ and

$$\|Sv\|_2^2 = \|v\|_2^2 - 2\,\mathrm{Re}\,(v, w)_2(v, w)_2 - 2\,\mathrm{Re}\,(v, w)_2(w, v)_2 + 4\,\mathrm{Re}\,(v, w)_2^2\|w\|_2^2 = \|v\|_2^2,$$

we have $\|B\|_2 = \|S\|_2 = 1$. Hence, for this particular choice of the matrix B, we have

$$\lambda_\varepsilon = \lambda - \varepsilon\omega(\varepsilon)\|v^*\|_2, \qquad \lim_{\varepsilon \to 0} \omega(\varepsilon) = 1,$$

as asserted. Q.E.D.

Remark 1.9: (i) We note that the statement of Theorem 1.7 becomes trivial if the matrix A is *normal*. In this case primal and dual eigenvectors coincide and, in view of Remark 1.7, $\sigma_\varepsilon(A)$ is the union of ε-circles around its eigenvalues λ. Hence, observing $\|w^*\|_2 = \|w\|_= 1$ and setting $\omega(\varepsilon) \equiv 1$, we trivially have $\lambda_\varepsilon := \lambda - \varepsilon \in \sigma_\varepsilon(A)$ as asserted.

(ii) If A is non-normal it may have a nontrivial pseudospectrum. Then, a large norm of the dual eigenfunction $\|w^*\|_2$ corresponding to a critical eigenvalue λ_{crit} with $-1 \ll \mathrm{Re}\,\lambda_{\mathrm{crit}} < 0$, indicates that the ε-pseudospectrum $\sigma_\varepsilon(A)$, even for small ε, reaches into the right complex half plane.

(iii) If the eigenvalue $\lambda \in \sigma(A)$ considered in Theorem 1.7 is deficient, the normalization $(w, w^*)_2 = 1$ is not possible. In this case, as discussed above, there is still another mechanism for triggering nonlinear instability.

1.3 Perturbation theory and conditioning

First, we analyze the "conditioning" of quadratic linear systems. There are two main sources of errors in solving an equation $Ax = b$:

a) errors in the "theoretical" solution caused by errors in the data, i. e,, the elements of A and b,

b) errors in the "numerical" solution caused by round-off errors in the course of the solution process.

1.3.1 Conditioning of linear algebraic systems

We give an error analysis for linear systems

$$Ax = b \tag{1.3.66}$$

with regular coefficient matrix $A \in \mathbb{K}^{n\times n}$. The matrix A and the vector b are faulty by small errors δA and δb, so that actually the perturbed system

$$\tilde{A}\tilde{x} = \tilde{b}, \tag{1.3.67}$$

is solved with $\tilde{A} = A + \delta A$, $\tilde{b} = b + \delta b$ and $\tilde{x} = x + \delta x$. We want to estimate the error δx in dependence of δA and δb. For this, we use an arbitrary vector norm $\|\cdot\|$ and the associated natural matrix norm.

Theorem 1.8 (Perturbation theorem): *Let the matrix $A \in \mathbb{K}^{n\times n}$ be regular and the perturbation satisfy $\|\delta A\| < \|A^{-1}\|^{-1}$. Then, the perturbed matrix $\tilde{A} = A + \delta A$ is also regular and for the resulting relative error in the solution there holds*

$$\frac{\|\delta x\|}{\|x\|} \le \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A)\|\delta A\|/\|A\|} \left\{ \frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right\}, \tag{1.3.68}$$

with the so-called "condition number" $\operatorname{cond}(A) := \|A\|\,\|A^{-1}\|$ *of the matrix A.*

Proof. The assumptions imply

$$\|A^{-1}\delta A\| \le \|A^{-1}\|\,\|\delta A\| < 1,$$

such that also $A + \delta A = A[I + A^{-1}\delta A]$ is regular by Lemma 1.15. From

$$(A + \delta A)\tilde{x} = b + \delta b\,, \quad (A + \delta A)x = b + \delta A x$$

it follows that then for $\delta x = \tilde{x} - x$

$$(A + \delta A)\delta x = \delta b - \delta A x,$$

and consequently using the estimate of Lemma 1.15,

$$\begin{aligned}
\|\delta x\| &\le \|(A+\delta A)^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&= \|\big(A(I+A^{-1}\delta A)\big)^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&= \|(I+A^{-1}\delta A)^{-1}A^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&\le \|(I+A^{-1}\delta A)^{-1}\|\,\|A^{-1}\| \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&\le \frac{\|A^{-1}\|}{1-\|A^{-1}\delta A\|} \left\{\|\delta b\| + \|\delta A\|\,\|x\|\right\} \\
&\le \frac{\|A^{-1}\|\,\|A\|\,\|x\|}{1-\|A^{-1}\|\,\|\delta A\|\,\|A\|\,\|A\|^{-1}} \left\{\frac{\|\delta b\|}{\|A\|\,\|x\|} + \frac{\|\delta A\|}{\|A\|}\right\}.
\end{aligned}$$

Since $\|b\| = \|Ax\| \le \|A\|\,\|x\|$ it eventually follows that

$$\|\delta x\| \le \frac{\mathrm{cond}(A)}{1-\mathrm{cond}(A)\|\delta A\|\|A\|^{-1}} \left\{\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right\} \|x\|,$$

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

The condition number $\mathrm{cond}(A)$ depends on the chosen vector norm in the estimate (1.3.68). Most often the max-norm $\|\cdot\|_\infty$ or the euclidian norm $\|\cdot\|_2$ are used. In the first case there holds

$$\mathrm{cond}_\infty(A) := \|A\|_\infty \|A^{-1}\|_\infty$$

with the maximal row sum $\|\cdot\|_\infty$. Especially for hermitian matrices Lemma 1.13 yields

$$\mathrm{cond}_2(A) := \|A\|_2 \|A^{-1}\|_2 = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

with the eigenvalues $\lambda_{\max}$ and $\lambda_{\min}$ of A with largest and smallest modulus, respectively. Accordingly, the quantity $\mathrm{cond}_2(A)$ is called the "spectral condition (number)" of A. In the case $\mathrm{cond}(A)\|\delta A\|\,\|A\|^{-1} \ll 1$, the stability estimate (1.3.68) takes the form

$$\frac{\|\delta x\|}{\|x\|} \;\approx\; \mathrm{cond}(A) \left\{\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|}\right\},$$

i. e., $\mathrm{cond}(A)$ is the amplification factor by which relative errors in the data A and b affect the relative error in the solution x.

Corollary 1.3: *Let the condition of A be of size $\mathrm{cond}(A) \sim 10^s$. Are the elements of A and b faulty with a relative error if size*

$$\frac{\|\delta A\|}{\|A\|} \approx 10^{-k}, \quad \frac{\|\delta b\|}{\|b\|} \approx 10^{-k} \quad (k > s),$$

then the relative error in the solution can be at most of size

$$\frac{\|\delta x\|}{\|x\|} \approx 10^{s-k}.$$

In the case $\|\cdot\| = \|\cdot\|_\infty$, one may lose s decimals in accuracy.

Example 1.4: Consider the following coefficient matrix A and its inverse A^{-1}:

$$A = \begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix}, \quad A^{-1} = 10^8 \begin{bmatrix} 0.1441 & -0.8648 \\ -0.2161 & 1.2969 \end{bmatrix}$$

$$\|A\|_\infty = 2.1617, \quad \|A^{-1}\|_\infty = 1.513 \cdot 10^8 \Rightarrow \operatorname{cond}(A) \approx 3.3 \cdot 10^8.$$

In solving the linear system $Ax = b$ one may lose 8 decimals in accuracy by which the elements a_{jk} and b_j are given. Hence, this matrix is very ill-conditioned.

Finally, we demonstrate that the stability estimate (1.3.68) is essentially sharp. Let A be a positive definite $n \times n$-matrix with smallest and largest eigenvalues λ_1 and λ_n and corresponding normalized eigenvectors w_1 and w_n, respectively. We choose

$$\delta A \equiv 0, \quad b \equiv w_n, \quad \delta b \equiv \varepsilon w_1 \ (\varepsilon \neq 0).$$

Then, the equations $Ax = b$ and $A\tilde{x} = b + \delta b$ have the solutions

$$x = \lambda_n^{-1} w_n, \quad \tilde{x} = \lambda_n^{-1} w_n + \varepsilon \lambda_1^{-1} w_1.$$

Consequently, for $\delta x = \tilde{x} - x$ there holds

$$\frac{\|\delta x\|_2}{\|x\|_2} = \varepsilon \frac{\lambda_n}{\lambda_1} \frac{\|w_1\|_2}{\|w_n\|_2} = \operatorname{cond}_2(A) \frac{\|\delta b\|_2}{\|b\|_2},$$

i. e., in this very special case the estimate (1.3.68) is sharp.

1.3.2 Conditioning of eigenvalue problems

The most natural way of computing eigenvalues of a matrix $A \in \mathbb{K}^{n \times n}$ appears to go via its definition as zeros of the characteristic polynomial $\chi_A(\cdot)$ of A and to compute corresponding eigenvectors by solving the singular system $(A - \lambda I)w = 0$. This approach is not advisable in general since the determination of zeros of a polynomial may be highly ill-conditioned, at least if the polynomial is given in canonical form as sum of monomials. We will see that the determination of eigenvalues may be well- or ill-conditioned depending on the properties of A, i. e., its deviation from being "normal".

Example 1.5: A symmetric matrix $A \in \mathbb{R}^{20 \times 20}$ with eigenvalues $\lambda_j = j$, $j = 1, \ldots, 20$, has the characteristic polynomial

$$\chi_A(z) = \prod_{j=1}^{20} (z - j) = z^{20} \underbrace{-210}_{b_1} z^{19} + \ldots + \underbrace{20!}_{b_{20}}.$$

The coefficient b_1 is perturbed: $\tilde{b}_1 = -210 + 2^{-23} \sim -210,000000119\ldots$, which results in

$$\text{relative error} \quad \Big|\frac{\tilde{b}_1 - b_1}{b_1}\Big| \sim 10^{-10}.$$

Then, the perturbed polynomial $\tilde{\chi}_A(z)$ has two roots $\lambda_\pm \sim 16.7 \pm 2.8i$, far away from the trues.

The above example shows that via the characteristic polynomial eigenvalues may be computed reliably only for very special matrices, for which $\chi_A(z)$ can be computed without determining its monomial from. Examples of some practical importance are, e. g., "tridiagonal matrices" or more general "Hessenberg[11] matrices".

$$\begin{bmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & & b_{n-1} \\ & & c_n & a_n \end{bmatrix} \qquad \begin{bmatrix} a_{11} & \cdots & & a_{1n} \\ a_{21} & \ddots & & \vdots \\ & \ddots & & a_{n-1,n} \\ 0 & & a_{n,n-1} & a_{nn} \end{bmatrix}$$

tridiagonal matrix Hessenberg matrix

Next, we provide a useful estimate which will be the basis for estimating the conditioning of the eigenvalue problem.

Lemma 1.18: *Let $A, B \in \mathbb{K}^{n\times n}$ be arbitrary matrices and $\|\cdot\|$ a natural matrix norm. Then, for any eigenvalue λ of A, which is not eigenvalue of B there holds*

$$\|(\lambda I - B)^{-1}(A - B)\| \geq 1. \tag{1.3.69}$$

Proof. If w is an eigenvector corresponding to the eigenvalue λ of A it follows that

$$(A - B)\, w = (\lambda I - B)\, w,$$

and for λ not being an eigenvalue of B,

$$(\lambda I - B)^{-1}(A - B)\, w = w.$$

Consequently

$$1 \leq \sup_{x\in\mathbb{K}^n\setminus\{0\}} \frac{\|(\lambda I - B)^{-1}(A - B)\, x\|}{\|x\|} = \|(\lambda I - B)^{-1}(A - B)\|,$$

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

As consequence of Lemma 1.18, we obtain the following important inclusion theorem of Gerschgorin[12] (1931).

Theorem 1.9 (Theorem of Gerschgorin): *All eigenvalues of a matrix $A \in \mathbb{K}^{n\times n}$ are contained in the union of the corresponding "Gerschgorin circles"*

$$K_j := \Big\{ z \in \mathbb{C} : |z - a_{jj}| \leq \sum_{k=1,k\neq j}^{n} |a_{jk}| \Big\}, \quad j = 1, \ldots, n. \tag{1.3.70}$$

[11] Karl Hessenberg (1904-1959): German mathematician; dissertation "Die Berechnung der Eigenwerte und Eigenl"osungen linearer Gleichungssysteme", TU Darmstadt 1942.

[12] Semyon Aranovich Gershgorin (1901-1933): Russian mathematician; since 1930 prof. in Leningrad (St. Petersburg); worked in algebra, complex function theory differential equations and numerics.

If the sets $U \equiv \cup_{i=1}^{m} K_{j_i}$ *and* $V \equiv \overline{\cup_{j=1}^{n} K_j \setminus U}$ *are disjoint then* U *contains exactly* m *and* V *exactly* $n-m$ *eigenvalues of* A *(counted accordingly to their algebraic multiplicities).*

Proof. (i) We set $B \equiv D = \mathrm{diag}(a_{jj})$ in Lemma 1.18 and take the "maximal row sum" as natural matrix norm. Then, it follows that for $\lambda \neq a_{jj}$:

$$\|(\lambda I - D)^{-1}(A - D)\|_\infty = \max_{j=1,\ldots,n} \frac{1}{|\lambda - a_{jj}|} \sum_{k=1,k\neq j}^{n} |a_{jk}| \geq 1,$$

i. e., λ is contained in one of the Gerschgorin circles.

(ii) For proving the second assertion, we set $A_t \equiv D + t(A-D)$. Obviously exactly m eigenvalues of $A_0 = D$ are in U and $n-m$ eigenvalues in V. The same then also follows for $A_1 = A$ since the eigenvalues of A_t (ordered accordingly to their algebraic multiplicities) are continuous functions of t. Q.E.D.

The theorem of Gerschgorin yields much more accurate information on the position of eigenvalues λ of A than the rough estimate $|\lambda| \leq \|A\|_\infty$ derived above. The eigenvalues of the matrices A and $\bar{A}^T$ are related by $\lambda(\bar{A}^T) = \overline{\lambda(A)}$. By applying the Gerschgorin theorem simultaneously to A and $\bar{A}^T$, one may obtain a sharpening of the estimates for the eigenvalues.

Example 1.6: *Consider the* 3×3*-matrix:*

$$A = \begin{bmatrix} 1 & 0.1 & -0.2 \\ 0 & 2 & 0.4 \\ -0.2 & 0 & 3 \end{bmatrix} \quad \|A\|_\infty = 3.2\,, \quad \|A\|_1 = 3.6.$$

$$\begin{aligned} K_1 &= \{z \in \mathbb{C} : |z-1| \leq 0.3\} & K_1^T &= \{z \in \mathbb{C} : |z-1| \leq 0.2\} \\ K_2 &= \{z \in \mathbb{C} : |z-2| \leq 0.4\} & K_2^T &= \{z \in \mathbb{C} : |z-2| \leq 0.1\} \\ K_3 &= \{z \in \mathbb{C} : |z-3| \leq 0.2\} & K_3^T &= \{z \in \mathbb{C} : |z-3| \leq 0.6\} \end{aligned}$$

$$|\lambda_1 - 1| \leq 0.2\,, \quad |\lambda_2 - 2| \leq 0.1\,, \quad |\lambda_3 - 3| \leq 0.2$$

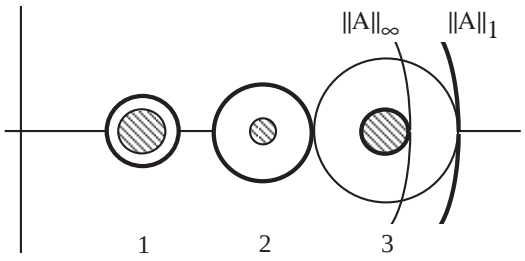


Figure 1.2: Gerschgorin circles of A *and* A^T

Next, from the estimate of Lemma 1.18, we derive the following basic stability result for the eigenvalue problem.

Theorem 1.10 (Stability theorem): *Let $A \in \mathbb{K}^{n\times n}$ be a diagonalizable matrix, i. e., one for which n linearly independent eigenvectors $\{w^1, \ldots, w^n\}$ exist, and let $B \in \mathbb{K}^{n\times n}$ be an arbitrary second matrix. Then, for each eigenvalue $\lambda(B)$ of B there is a corresponding eigenvalue $\lambda(A)$ of A such that with the matrix $W = [w^1, \ldots, w^n]$ there holds*

$$|\lambda(A) - \lambda(B)| \;\leq\; \mathrm{cond}_2(\mathrm{W})\, \|\mathrm{A} - \mathrm{B}\|_2\,. \tag{1.3.71}$$

Proof. The eigenvalue equation $Aw^i = \lambda_i(A)w^i$ can be rewritten in matrix form $AW = W\,\mathrm{diag}(\lambda_i(A))$ with the regular matrix $W = [w_1, \ldots, w_n]$. Consequently,

$$A = W\,\mathrm{diag}(\lambda_i(A))\,W^{-1},$$

i. e., A is "similar" to the diagonal matrix $\Lambda = \mathrm{diag}(\lambda_i(A))$. Since $\lambda = \lambda(B)$ is not an eigenvalue of A,

$$\begin{aligned}\|(\lambda I - A)^{-1}\|_2 &= \|W(\lambda I - \Lambda)^{-1}W^{-1}\|_2\\ &\leq \|W^{-1}\|_2\|W\|_2\|(\lambda I - \Lambda)^{-1}\|_2\\ &= \mathrm{cond}_2(W)\max_{i=1,\ldots,n}|\lambda - \lambda_i(A)|^{-1}.\end{aligned}$$

Then, Lemma 1.18 yields the estimate,

$$\begin{aligned}1 &\leq \|(\lambda I - A)^{-1}\,(B - A)\| \leq \|(\lambda I - A)^{-1}\|\|(B - A)\|\\ &\leq \mathrm{cond}_2(W)\max_{i=1,\ldots,n}|\lambda - \lambda_i(A)|^{-1}\|(B - A)\|,\end{aligned}$$

from which the assertion follows. Q.E.D.

For hermitian matrices $A \in \mathbb{K}^{n\times n}$ there exists an ONB in $\mathbb{K}^n$ of eigenvectors so that the matrix W in the estimate (1.3.71) can be assumed to be unitary, $W\bar{W}^T = I$. In this special case there holds

$$\mathrm{cond}_2(W) = \|\bar{W}^T\|_2\,\|W\|_2 = 1, \tag{1.3.72}$$

i. e., the eigenvalue problem of "hermitian" (or more general "normal") matrices is well conditioned. For general "non-normal" matrices the conditioning of the eigenvalue problem may be arbitrarily bad, $\mathrm{cond}_2(W) \gg 1$.

1.4 Exercises

Exercise 1.1 (About the geometry of $\mathbb{R}^n$): The unit sphere with respect to a vector norm $\|\cdot\|$ on $\mathbb{R}^n$ is defined by

$$S := \{x \in \mathbb{R}^n \,|\, \|x\| = 1\}.$$

Sketch the unit spheres in $\mathbb{R}^2$ corresponding to the l_1-norm, the euclidian norm and the l_∞-norm:

$$\|x\|_1 := |x_1| + |x_2|, \qquad \|x\|_2 := \left(|x_1|^2 + |x_2|^2\right)^{1/2}, \qquad \|x\|_\infty := \max\left\{|x_1|, |x_2|\right\}.$$

How do the unit spheres corresponding to the general l_p-norms look like?

Exercise 1.2 (Some useful facts about norms and scalar products): Verify the following claims for vectors $x, y \in \mathbb{R}^n$ and the euclidean norm $\|\cdot\|_2$ and scalar product $(\cdot,\cdot)_2$:

a) $2\|x\|_2^2 + 2\|y\|_2^2 = \|x+y\|_2^2 + \|x-y\|_2^2$ (parallelogram identity).

b) $|(x,y)_2| \le \|x\|_2\|y\|_2$ (Cauchy-Schwarz inequality).

c) For any symmetric, positive definite matrix $A \in \mathbb{R}^{n\times n}$ the bilinear form $(x,y)_A := (Ax,y)_2$ is a scalar product. i) Can any scalar product on $\mathbb{R}^{n\times n}$ be written in this form? ii) How has this to be formulated for complex Matrices $A \in \mathbb{C}^{n\times n}$?

Exercise 1.3: a) Let $(V, \|\cdot\|)$ be a *real* normed vector space the norm of which, $\|\cdot\|$, satisfies the "parallelogram identity":

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2, \quad x, y \in V.$$

Show that then

$$(x,y) := \tfrac{1}{4}\|x+y\|^2 - \tfrac{1}{4}\|x-y\|^2, \quad x, y \in V,$$

defines a scalar product on V, by which the given norm can be generated. i. e., is the corresponding "natural" vector norm.
(Hint: The proofs of the several scalar product properties are of very different difficulty. The assertion also holds in *complex* vector spaces, but with an appropriately adjusted definition of the scalar product $(\cdot,\cdot)$.)

b) Show that for $p \in [1,\infty) \cup \{\infty\} \setminus \{2\}$ the l_p-norms on $\mathbb{R}^n$ with $n > 1$,

$$\|x\|_p := \Big(\sum_{i=1}^{n} |x_i|^p\Big)^{1/p}, \quad x \in \mathbb{R}^n,$$

cannot be generated from a scalar product.

Exercise 1.4 (Some useful facts about matrix norms): Verify the following relations for matrices $A, B \in \mathbb{K}^{n\times n}$ and the euclidean norm $\|\cdot\|_2$:

a) $\|A\|_2 := \max\big\{\|Ax\|_2/\|x\|_2,\, x \in \mathbb{K}^n, x \neq 0\big\} = \max\big\{\|Ax\|_2,\, x \in \mathbb{R}^n, \|x\|_2 = 1\big\}$.

b) $\|Ax\|_2 \le \|A\|_2\|x\|_2$.

c) $\|AB\|_2 \le \|A\|_2\|B\|_2$ (Is this relation true for any matrix norm?).

d) For hermitian matrices $A \in \mathbb{C}^{n\times n}$ there holds $\|A\|_2 = \max\{|\lambda|,\, \lambda \text{ eigenvalue of } A\}$.

e) For general matrices $A \in \mathbb{C}^{n\times n}$ there holds $\|A\|_2 = \max\{|\lambda|^{1/2},\, \lambda \text{ eigenvalue of } \bar{A}^T A\}$.

Exercise 1.5 (Some useful facts about vector spaces and matrices):
a) Formulate the Gram-Schmidt algorithm for orthonormalizing a set of linearly independent vectors $\{x_1, \ldots, x_m\} \subset \mathbb{R}^n$:

b) How can one define the square root $A^{1/2}$ of a symmetric, positive definite matrix $A \in \mathbb{R}^{n\times n}$?

c) Show that a positive definite matrix $A \in \mathbb{C}^{n\times n}$ is automatically hermitian, i.e., $A = \bar{A}^T$. This is not necessarily true for real matrices $A \in \mathbb{R}^{n\times n}$, i.e., for real matrices the definition of positiveness usually goes together with the requirement of symmetry.

Exercise 1.6: Recall the definitions of the following quantities:

a) The "maximum-norm" $\|\cdot\|_\infty$ and the "l_1-norm" $\|\cdot\|_1$ on $\mathbb{K}^n$.

b) The "spectrum" $\sigma(A)$ of a matrix $A \in \mathbb{K}^{n\times n}$.

c) The "Gerschgorin circles" $K_i \subset \mathbb{C}$, $i = 1, \ldots, n$, of a matrix $A \in \mathbb{K}^{n\times n}$.

d) The "spectral radius" $\rho(A)$ of a matrix $A \in \mathbb{K}^{n\times n}$.

e) The "spectral condition number" $\kappa_2(A)$ of a matrix $A \in \mathbb{K}^{n\times n}$.

Exercise 1.7: Recall the proofs of the following facts about matrices:

a) The diagonal elements of a (hermitian) positive definite matrix $A \in \mathbb{K}^{n\times n}$ are real positive.

b) For the trace $\mathrm{tr}(A) := \sum_{i=1}^n a_{ii}$ of a hermitian matrix $A \in \mathbb{K}^{n\times n}$ with eigenvalues $\lambda_i \in \Sigma(A)$ there holds

$$\mathrm{tr}(A) = \sum_{i=1}^n \lambda_i .$$

c) A strictly diagonally dominant matrix $A \in \mathbb{K}^{n\times n}$ is regular. If it is also hermitian with (real) positive diagonal entries, then it is positive definite.

Exercise 1.8: Show that for any vector norm $\|\cdot\|$ on $\mathbb{K}^n$

$$\|A\| := \sup\left\{\frac{\|Ax\|}{\|x\|},\, x \in \mathbb{K}^n, x \neq 0\right\} = \sup\left\{\|Ax\|,\, x \in \mathbb{K}^n, \|x\| = 1\right\}$$

defines a compatible matrix norm. This is called the "natural" matrix norm generated by the given vector norm. Why can the square-sum norm (so-called "Frobenius norm")

$$\|A\|_{\mathrm{FR}} = \Big(\sum_{i,j=1}^n |a_{ij}|^2\Big)^{1/2}$$

not be such a "natural" matrix norm ?

Exercise 1.9: Let $B \in \mathbb{K}^{n\times n}$ be a matrix which for some matrix norm $\|\cdot\|$ (not necessarily compatible with a vector norm) satisfies $\|B\| < 1$. Recall the proof that the matrix $I - B$ is regular with inverse satisfying

$$\|(I-B)^{-1}\| \le \frac{1}{1-\|B\|} .$$

Give an explicit power series representation of the inverse $(I-B)^{-1}$.

Exercise 1.10: Prove that (i) each connected component of k Gerschgorin circles (that are disjoined to all other $n-k$ circles) of a matrix $A \in \mathbb{K}^{n\times n}$ contains exactly k eigenvalues of A (counted accordingly to their algebraic multiplicities). (ii) This implies that a matrix, for which all Gerschgorin circles are mutually disjoint, is diagonalizable.

Exercise 1.11: Let $A, B \in \mathbb{K}^{n\times n}$ be two hermitian matrices. Then, the following statements are equivalent:

(i) A and B commute, i. e., $AB = BA$.

(ii) A and B possess a common basis of eigenvectors.

(iii) AB is hermitian.

Does the above equivalence in an appropriate sense also hold for two general "normal" (or even more general "diagonalizable") matrices $A, B \in \mathbb{K}^{n\times n}$?

Exercise 1.12: A "sesquilinear form" on $\mathbb{K}^n$ is a mapping $\varphi(\cdot,\cdot) : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$, which is bilinear in the following sense:

$$\varphi(\alpha x + \beta y, z) = \bar{\alpha}\varphi(x,z) + \bar{\beta}\varphi(y,z), \quad \varphi(z, \alpha x + \beta y) = \alpha\varphi(z,x) + \beta\varphi(z,y), \quad \alpha,\beta \in \mathbb{K}.$$

(i) Show that for any regular matrix $A \in \mathbb{K}^{n\times n}$ the sesquilinear form $\varphi(x,y) := (Ax, Ay)_2$ is a scalar product on $\mathbb{K}^n$.

(ii) In an earlier exercise, we have seen that each scalar product (x,y) on $\mathbb{K}^n$ can be written in the form $(x,y) = (x, Ay)_2$ with a (hermitian) positive definite matrix $A \in \mathbb{K}^{n\times n}$. Why does this statement not contradict (i)?

Exercise 1.13: Let $A \in \mathbb{K}^{n\times n}$ be hermitian.

(i) Show that eigenvectors corresponding to different eigenvalues $\lambda_1(A)$ and $\lambda_2(A)$ are orthogonal. Is this also true for (non-hermitian) "normal" matrices, i. e., if $\bar{A}^T A = A\bar{A}^T$?

(ii) Show that there holds

$$\lambda_{\min}(A) = \min_{x\in\mathbb{K}^n\setminus\{0\}} \frac{(Ax,x)_2}{\|x\|_2^2} \le \max_{x\in\mathbb{K}^n\setminus\{0\}} \frac{(Ax,x)_2}{\|x\|_2^2} = \lambda_{\max}(A),$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimal and maximal (real) eigenvalues of A, respectively. (Hint: Use that a hermitian matrix possesses an ONS basis of eigenvectors.)

Exercise 1.14: Let $A \in \mathbb{K}^{n\times n}$ and $0 \notin \sigma(A)$. Show that the ε-pseudospectra of A and that of its inverse A^{-1} are related by

$$\sigma_\varepsilon(A) \subset \big\{ z \in \mathbb{C}\setminus\{0\} \,\big|\, z^{-1} \in \sigma_{\delta(z)}(A^{-1}) \big\} \cup \{0\},$$

where $\delta(z) := \varepsilon\|A^{-1}\|/|z|$ and, for $0 < \varepsilon < 1$, by

$$\sigma_\varepsilon(A^{-1}) \setminus B_1(0) \subset \big\{ z \in \mathbb{C}\setminus\{0\} \,\big|\, z^{-1} \in \sigma_\delta(A) \big\},$$

where $B_1(0) := \{z \in \mathbb{C},\ \|z\| \le 1\}$ and $\delta := \varepsilon/(1-\varepsilon)$.

Exercise 1.15: Consider the linear system

$$\begin{pmatrix} 0,5 & 0,5 \\ 0,5 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

How large are the relative errors $\|\delta x\|_1/\|x\|_1$ and $\|\delta x\|_\infty/\|x\|_\infty$ if the corresponding relative errors in the matrix elements are $\pm 1\%$ and that in the right hand side elements $\pm 3\%$? Sketch

the point sets in $\mathbb{R}^2$ in which the solution $x + \delta x$ of the perturbed system is located. (Hint: Compute the inverse of the coefficient matrix and determine its l_1- and l_∞-condition.)

Exercise 1.16: Let $A \in \mathbb{K}^{n\times n}$ be arbitrary. Show that

(i) The matrix $\bar{A}^T A$ is hermitian and positive semi-definite. For regular A it is even positive definite.

(ii) The spectral condition of A is given by

$$\|A\|_2 = \max\{|\lambda|^{1/2},\, \lambda \in \sigma(\bar{A}^T A)\}.$$

Exercise 1.17: a) Show that the set M of regular matrices,

$$M := \{A \in \mathbb{K}^{n\times n} \,|\, A \text{ regular}\,\} \subset \mathbb{K}^{n\times n},$$

is open in $\mathbb{K}^{n\times n}$.

b) For a matrix $A \in \mathbb{K}^{n\times n}$ the "resolvent" is defined by $R(z) := (A - zI)^{-1}$ and the corresponding "resolvent set" by

$$\mathrm{Res}(A) := \{z \in \mathbb{C} \,|\, A - zI \in \mathbb{K}^{n\times n} \text{ regulär}\} \subset \mathbb{C}.$$

As complement of the (discrete) spectrum $\sigma(A)$ the resolvent set is open. Show that the resolvent $R(\cdot)$ is on $\mathrm{Res}(A)$ continuous and on any compact subset of $\mathrm{Res}(A)$ uniformly Lipschitz-continuous.

Exercise 1.18: For any matrix $A \in \mathbb{K}^{n\times n}$ the exponential matrix $e^A \in \mathbb{K}^{n\times n}$ may formally be defined by the following Taylor series:

$$e^A := \sum_{k=0}^{\infty} \frac{A^k}{k!} := \lim_{n\to\infty} \sum_{k=0}^{n} \frac{A^k}{k!}.$$

a) Show that this series converges, i. e., that the sequence of partial sums has a limit in $\mathbb{K}^{n\times n}$. This limit is uniquely determined and is denoted by e^A. (Hint: Use the Cauchy convergence criterium.)

b) Show that for diagonalizable matrices $A, B \in \mathbb{K}^{n\times n}$, which have a common basis of eigenvectors there holds,

$$e^{A+B} = e^A e^B.$$

(Hint: Observe the property $\lambda_i(A+B) = \lambda_i(A) + \lambda_i(B)$ for eigenvalues and the known relation $e^{a+b} = e^a e^b$ for numbers $a, b \in \mathbb{K}$. In general, $e^{A+B} \neq e^A e^B$ for non-commuting matrices.)

2 Direct Solution Methods

2.1 Gaussian elimination, LR and Cholesky decomposition

In this chapter, we collect some basic results on so-called "direct" methods for solving linear systems and matrix eigenvalue problems. A "direct" method delivers the exact solution theoretically in finitely many arithmetic steps, at least under the assumption of "exact" real arithmetic. However, to get useful results a "direct" method has to be carried to its very end. In contrast to this, so-called "iterative" methods produce sequences of approximate solutions of increasing accuracy, which theoretically converge to the exact solution in infinitely many arithmetic steps. However, 'terative" methods may yield useful results already after a small number of iterations. Usually "direct" methods are very robust but, due to their usually high storage requirements and computational work, feasible only for problems of moderate size. Here, the meaning of "moderate size" depends very much on the currently available computer power, i. e., today reaches up to dimension $n \approx 10^5 - 10^6$. Iterative methods need less storage and as multi-level algorithms may even show optimal arithmetic complexity, i. e., a fixed improvement in accuracy is achieved in $\mathcal{O}(n)$ arithmetic operations. These methods can be used for really large-scale problems of dimension reaching up to $n \approx 10^6 - 10^9$ but at the prize of less robustness and higher algorithmic complexity. Such modern "iterative" methods are the main subject of this course and will be discussed in the next chapters.

2.1.1 Gaussian elimination and LR decomposition

In the following, we discuss "direct methods" for solving (real) quadratic linear systems

$$Ax = b\,. \tag{2.1.1}$$

It is particularly easy to solve staggered systems, e. g., those with an upper triangular matrix $A = (a_{jk})$ as coefficient matrix

$$\begin{array}{rcrcrcl} a_{11}x_1 & + & a_{12}x_2 & + & \ldots & + & a_{1n}x_n = b_1 \\ & & a_{22}x_2 & + & \ldots & + & a_{2n}x_n = b_2 \\ & & & & & & \vdots \\ & & & & & & a_{nn}x_n = b_n \end{array} \quad .$$

In case that $a_{jj} \neq 0$, $j = 1, \ldots, n$, we obtain the solution by so-called "backward substitution":

$$x_n = \frac{b_n}{a_{nn}}\,, \quad j = n-1, \ldots, 1: \quad x_j = \frac{1}{a_{jj}}\,(b_j - \sum_{k=j+1}^{n} a_{jk}\,x_k).$$

This requires $N_{\text{back subst}} = n^2/2 + O(n)$ *arithmetic operations.* The same holds true if the coefficient matrix is lower triangular and the system is solved by the corresponding "forward substitution" .

Definition 2.1: *For quantifying the arithmetic work required by an algorithm, i. e., its "(arithmetical) complexity", we use the notion "arithmetical operation" (in short "a. op."), which*

means the equivalent of "1 multiplication + 1 addition" or "1 division" (assuming that the latter operations take about the same time on a modern computer).

The *classical* direct method for solving linear systems is the elimination method of Gauß[1] which transforms the system $Ax = b$ in several "elimination steps" (assuming "exact" arithmetic) into an equivalent upper triangular system $Rx = c$, which is then solved by backward substitution. In practice, due to round-off errors, the resulting upper triangular system is not exactly equivalent to the original problem and this unavoidable error needs to be controlled by additional algorithmical steps ("final iteration", or "Nachiteration", in German). In the elimination process two elementary transformations are applied to the matrix A, which do not alter the solution of system (2.1.1): "*permutation of two rows of the matrix*" or "*addition of a scalar multiple of a row to another row of the matrix*". Also the "*permutation of columns*" of A is admissible if the unknowns x_i are accordingly renumbered.

In the practical realization of Gaussian elimination the elementary transformations are applied to the composed matrix $[A, b]$. In the following, we assume the matrix A to be regular. First, we set $A^{(0)} \equiv A$, $b^{(0)} \equiv b$ and determine $a_{r1}^{(0)} \neq 0$, $r \in \{1, \ldots, n\}$. (Such an element exists since otherwise A would be singular.). Permute the 1st and the rth row. Let the result be the matrix $[\tilde{A}^{(0)}, \tilde{b}^{(0)}]$. Then, for $j = 2, \ldots, n$, we multiply the 1st row by q_{j1} and subtract the result from the jth row,

$$q_{j1} \equiv \tilde{a}_{j1}^{(0)} / \tilde{a}_{11}^{(0)} \; (= a_{r1}^{(0)} / a_{rr}^{(0)}), \quad a_{ji}^{(1)} := \tilde{a}_{ji}^{(0)} - q_{j1} \tilde{a}_{1i}^{(0)}, \quad b_j^{(1)} := \tilde{b}_j^{(0)} - q_{j1} \tilde{b}_1^{(0)}.$$

The result is

$$[A^{(1)}, b^{(1)}] = \left[\begin{array}{cccc|c} \tilde{a}_{11}^{(0)} & \tilde{a}_{12}^{(0)} & \ldots & \tilde{a}_{1n}^{(0)} & \tilde{b}_1^{(0)} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & & & \\ 0 & a_{n2}^{(1)} & \ldots & a_{nn}^{(1)} & b_n^{(1)} \end{array}\right].$$

The transition $[A^{(0)}, b^{(0)}] \to [\tilde{A}^{(0)}, \tilde{b}^{(0)}] \to [A^{(1)}, b^{(1)}]$ can be expressed in terms of matrix multiplication as follows:

$$[\tilde{A}^{(0)}, \tilde{b}^{(0)}] = P_1 [A^{(0)}, b^{(0)}], \quad [A^{(1)}, b^{(1)}] = G_1 [\tilde{A}^{(0)}, \tilde{b}^{(0)}],$$

where P_1 is a "permutation matrix" und G_1 is a "Frobenius matrix" of the following form:

[1] Carl Friedrich Gauß (1777-1855): eminent German mathematician, astronomer and physicist; worked in Göttingen; fundamental contributions to arithmetic, algebra and geometry; founder of modern number theory, determined the planetary orbits by his "equalization calculus", further contributions to earth magnetism and construction of an electro-magnetic telegraph.

$$P_1 = \begin{array}{c} \begin{matrix} 1 & & & & r & & \end{matrix} \\ \begin{bmatrix} 0 & & \cdots & & 1 & & \\ & 1 & & & & & \\ \vdots & & \ddots & & \vdots & & \\ & & & 1 & & & \\ 1 & & \cdots & & 0 & & \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & & 1 \end{bmatrix} \end{array} \begin{matrix} 1 \\ \\ \\ \\ r \\ \\ \\ \end{matrix} \qquad G_1 = \begin{array}{c} \begin{matrix} & 1 & & & \end{matrix} \\ \begin{bmatrix} 1 & & & \\ -q_{21} & 1 & & \\ \vdots & & \ddots & \\ -q_{n1} & & & 1 \end{bmatrix} \end{array} 1$$

Both matrices, P_1 and G_1, are regular regular with determinants $\det(P_1) = \det(G_1) = 1$ and there holds

$$P_1^{-1} = P_1\,, \qquad G_1^{-1} = \begin{bmatrix} 1 & & & \\ q_{21} & 1 & & \\ \vdots & & \ddots & \\ q_{n1} & & & 1 \end{bmatrix}.$$

The systems $Ax = b$ and $A^{(1)}x = b^{(1)}$ have obviously the same solution,

$$Ax = b \quad \Longleftrightarrow \quad A^{(1)}x = G_1P_1Ax = G_1P_1b = b^{(1)}.$$

Definition 2.2: *The element $a_{r1} = \tilde{a}_{11}^{(0)}$ is called "pivot element" and the whole substep of its determination "pivot search". For reasons of numerical stability one usually makes the choice*

$$|a_{r1}| = \max_{1\le j\le n} |a_{j1}|\,. \tag{2.1.2}$$

The whole process incl. permutation of rows is called "column pivoting" . If the elements of the matrix A are of very different size "total pivoting" is advisable. This consists in the choice

$$|a_{rs}| = \max_{1\le j,k\le n} |a_{jk}|, \tag{2.1.3}$$

and subsequent permutation of the $1st$ row with the rth row and the $1st$ column with the sth column. According to the column permutation also the unknowns x_k have to be renumbered. However, "total pivoting" is costly so that simple "column pivoting" is usually preferred.

$$P_i = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccccc} & & & & \text{i} & & & & \text{r} & & \\ \end{array} \\ \begin{bmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & 0 & \cdots & & 1 & & & & \\ & & & & 1 & & & & & & \\ & & & \vdots & & \ddots & & \vdots & & & \\ & & & & & & 1 & & & & \\ & & & 1 & \cdots & & 0 & & & & \\ & & & & & & & 1 & & & \\ & & & & & & & & \ddots & & \\ & & & & & & & & & 1 & \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \text{i} \\ \\ \\ \\ \text{r} \\ \\ \\ \end{array}$$

$$G_i = \begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccc} & & \text{i} & & \\ \end{array} \\ \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -q_{i+1,i} & 1 & \\ & & \vdots & & \ddots \\ & & -q_{ni} & & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \text{i} \\ \\ \end{array}$$

The matrix $A^{(1)}$ generated in the first step is again regular. The same is true for the reduced submatrix to which the next elimination step is applied. By repeating this elimination process, one obtains in $n-1$ steps a sequence of matrices,

$$[A, b] \to [A^{(1)}, b^{(1)}] \to \ldots \to [A^{(n-1)}, b^{(n-1)}] =: [R, c]\,,$$

where

$$[A^{(i)}, b^{(i)}] = G_i P_i [A^{(i-1)}, b^{(i-1)}]\,, \quad [A^{(0)}, b^{(0)}] := [A, b]\,,$$

with (unitary) permutation matrices P_i and (regular) Frobenius matrices G_i of the above form. The end result

$$[R, c] = G_{n-1} P_{n-1} \ldots G_1 P_1 [A, b] \tag{2.1.4}$$

is an upper triangular system $Rx = c$, which has the same solution as the original system $Ax = b$. By the ith elimination step $[A^{(i-1)}, b^{(i-1)}] \to [A^{(i)}, b^{(i)}]$ the subdiagonal elements in the ith column are made zero. The resulting free places are used for storing the elements $q_{i+1,i}, \ldots, q_{n,i}$ of the matrices G_i^{-1} $(i = 1, \ldots, n-1)$. Since in this elimination step the preceding rows 1 to i are not changed, one works with matrices of the form

$$\left[\begin{array}{ccccccc|c} r_{11} & r_{12} & \cdots & r_{1i} & r_{1,i+1} & \cdots & r_{1n} & c_1 \\ \lambda_{21} & r_{22} & \cdots & r_{2i} & r_{2,i+1} & \cdots & r_{2n} & c_2 \\ \lambda_{31} & \lambda_{32} & & r_{3i} & r_{3,i+1} & \cdots & r_{3n} & c_3 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{i1} & \lambda_{i2} & & r_{ii} & r_{i,i+1} & \cdots & r_{in} & c_i \\ \lambda_{i+1,1} & \lambda_{i+2,2} & & \lambda_{i+1,i} & a_{i+1,i+1}^{(i)} & \cdots & a_{i+1,n}^{(i)} & b_{i+1}^{(i)} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{n,1} & \lambda_{n,2} & \cdots & \lambda_{n,i} & a_{n,i+1}^{(i)} & \cdots & a_{n,n}^{(i)} & b_n^{(i)} \end{array}\right].$$

Here, the subdiagonal elements $\lambda_{k+1,k}, \ldots, \lambda_{nk}$ in the kth column are permutations of the elements $q_{k+1,k}, \ldots, q_{nk}$ of G_k^{-1} since the permutations of rows (and only those) are applied to the whole composed matrix. As end result, we obtain the matrix

$$\left[\begin{array}{cccc|c} r_{11} & & \cdots & r_{1n} & c_1 \\ l_{21} & r_{22} & & r_{2n} & c_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ l_{n1} & \cdots & l_{n,n-1} & r_{nn} & c_n \end{array}\right].$$

Theorem 2.1 (LR decomposition): *The matrices*

$$L = \begin{bmatrix} 1 & & & 0 \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \cdots & l_{n,n-1} & 1 \end{bmatrix}, \quad R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{bmatrix}$$

are the factors in a so-called (multiplicative) "LR decomposition" of the matrix PA,

$$PA = LR, \qquad P := P_{n-1} \cdots P_1 . \tag{2.1.5}$$

If the LR decomposition is possible with $P = I$, then it is uniquely determined. Once an LR decomposition is computed the solution of the linear system $Ax = b$ can be accomplished by successively solving two triangular systems,

$$Ly = Pb, \qquad Rx = y, \tag{2.1.6}$$

by forward and backward substitution, respectively.

Proof. (i) We give the proof only for the case that pivoting is not necessary, i. e., $P_i = I$. Then, $R = G_{n-1} \cdots G_1 A$ and $G_1^{-1} \cdots G_{n-1}^{-1} R = A$. In view of $L = G_1^{-1} \cdots G_{n-1}^{-1}$ the first assertion follows.
(ii) To prove uniqueness let $A = L_1 R_1 = L_2 R_2$ be two LR decompositions. Then, $L_2^{-1} L_1 = R_2 R_1^{-1} = I$ since $L_2^{-1} L_1$ is lower triangular with ones on the main diagonal and $R_2 R_1^{-1}$ is upper triangular. Consequently, $L_1 = L_2$ and $R_1 = R_2$, what was to be shown. Q.E.D.

Lemma 2.1: *The solution of a linear $n \times n$ system $Ax = b$ by Gaussian elimination requires*

$$N_{Gauß}(n) = \tfrac{1}{3} n^3 + O(n^2) \tag{2.1.7}$$

arithmetical operations. This is just the work count of computing the corresponding decomposition $PA = LR$, while the solution of the two triangular systems (2.1.6) only requires $n^2 + O(n)$ arithmetical operations.

Proof. The kth elimination step

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \quad b_i^{(k)} = b_i^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_k^{(k-1)}, \quad i, j = k, \ldots, n,$$

requires $n-k$ divisions and $(n-k)+(n-k)^2$ combined multiplications and additions resulting altogether in

$$\sum_{k=1}^{n-1} k^2 + O(n^2) = \tfrac{1}{3}n^3 + O(n^2) \quad \text{a. Op.}$$

for the $n-1$ steps of forward elimination. By this all elements of the matrices L and R are computed. The work count of the forward and backward elimination in (2.1.6) follows by similar considerations. Q.E.D.

Example 2.1: The pivot elements are marked by $\boxed{\cdot}$.

$$\begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix} \quad \rightarrow \quad \begin{array}{c} \text{pivoting} \\ \hline \boxed{3} \;\; 1 \;\; 6 \;\vert\; 2 \\ 2 \;\; 1 \;\; 3 \;\vert\; 7 \\ 1 \;\; 1 \;\; 1 \;\vert\; 4 \end{array}$$

$$\begin{array}{c} \text{elimination} \\ \hline 3 \;\; 1 \;\; 6 \;\vert\; 2 \\ 2/3 \;\; 1/3 \;\; -1 \;\vert\; 17/3 \\ 1/3 \;\; \boxed{2/3} \;\; -1 \;\vert\; 10/3 \end{array} \quad \rightarrow \quad \begin{array}{c} \text{pivoting} \\ \hline 3 \;\; 1 \;\; 6 \;\vert\; 2 \\ 1/3 \;\; 2/3 \;\; -1 \;\vert\; 10/3 \\ 2/3 \;\; 1/3 \;\; -1 \;\vert\; 17/3 \end{array}$$

$$\begin{array}{c} \text{elimination} \\ \hline 3 \;\; 1 \;\; 6 \;\vert\; 2 \\ 1/3 \;\; 2/3 \;\; -1 \;\vert\; 10/3 \\ 2/3 \;\; 1/2 \;\; -1/2 \;\vert\; 4 \end{array} \quad \rightarrow \quad \begin{aligned} x_3 &= -8 \\ x_2 &= \tfrac{3}{2}(\tfrac{10}{3} - x_3) = -7 \\ x_1 &= \tfrac{1}{3}(2 - x_2 - 6x_3) = 19\,. \end{aligned}$$

LR decomposition:

$$P_1 = I\,, \quad P_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

$$PA = \begin{bmatrix} 3 & 1 & 6 \\ 1 & 1 & 1 \\ 2 & 1 & 3 \end{bmatrix} = LR = \begin{bmatrix} 1 & 0 & 0 \\ 1/3 & 1 & 0 \\ 2/3 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 6 \\ 0 & 2/3 & -1 \\ 0 & 0 & -1/2 \end{bmatrix}.$$

Example 2.2: For demonstrating the importance of the pivoting process, we consider the following linear 2×2-system:

$$\begin{bmatrix} 10^{-4} & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{2.1.8}$$

with the exact solution $x_1 = 1.00010001$, $x_2 = 0.99989999$. Using 3-decimal floating point arithmetic with correct rounding yields

a) without pivoting:

x_1	x_2	
$0.1 \cdot 10^{-3}$	$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$
0	$-0.1 \cdot 10^{5}$	$-0.1 \cdot 10^{5}$
$x_2 = 1\,,$	$x_1 = 0$	

b) with pivoting:

x_1	x_2	
$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$	$0.2 \cdot 10^{1}$
0	$0.1 \cdot 10^{1}$	$0.1 \cdot 10^{1}$
$x_2 = 1\,,$	$x_1 = 1$	

Example 2.3: The positive effect of *column* pivoting is achieved only if all row sums of the matrix A are of similar size. As an example, we consider the 2×2-system

$$\begin{bmatrix} 2 & 20000 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 20000 \\ 2 \end{bmatrix},$$

which results from (2.1.8) by scaling the first row by the factor 20.000. Since in the first column the element with largest modulus is on the main diagonal the Gauß algorithm with and without pivoting yields the same unacceptable result $(x_1, x_2)^T = (0, 1)^T$. To avoid this effect, we apply an "equilibration" step before the elimination, i. e., we multiply A by a diagonal matrix D,

$$Ax = b \quad \rightarrow \quad DAx = Db\,, \qquad d_i = \Big(\sum_{j=1}^{n} |a_{ij}| \Big)^{-1}, \tag{2.1.9}$$

such that all row sums of A are scaled to 1. An even better stabilization in the case of matrix elements of very different size is "total pivoting". Here, an equilibration step, row-wise and column-wise, is applied before the elimination.

Conditioning of Gaussian elimination

We briefly discuss the conditioning of the solution of a linear system by Gaussian elimination. For any (regular) matrix A there exists an LR decomposition like $PA = LR$. Then, there holds

$$R = L^{-1}PA, \quad R^{-1} = (PA)^{-1}L.$$

Due to column pivoting the elements of the triangular matrices L and L^{-1} are all less or equal one and there holds

$$\mathrm{cond}_\infty(L) = \|L\|_\infty \|L^{-1}\|_\infty \le n^2.$$

Consequently,

$$\begin{aligned} \mathrm{cond}_\infty(R) &= \|R\|_\infty \|R^{-1}\|_\infty = \|L^{-1}PA\|_\infty \|(PA)^{-1}L\|_\infty \\ &\le \|L^{-1}\|_\infty \|PA\|_\infty \|(PA)^{-1}\|_\infty \|L\|_\infty \le n^2\, \mathrm{cond}_\infty(PA). \end{aligned}$$

Then, the general perturbation theorem, Theorem1.8, yields the following estimate for the solution of the equation $LRx = Pb$ (considering only perturbations of the right-hand side b):

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \le \mathrm{cond}_\infty(L)\mathrm{cond}_\infty(R) \frac{\|\delta Pb\|_\infty}{\|Pb\|_\infty} \le n^4\, \mathrm{cond}_\infty(PA) \frac{\|\delta Pb\|_\infty}{\|Pb\|_\infty}.$$

Hence the conditioning of the original system $Ax = b$ is by the LR decomposition, in the worst case, amplified by n^4. However, this is an extremely pessimistic estimate, which can significantly be improved.

Theorem 2.2 (Round-off error influence): *The matrix $A \in \mathbb{R}^{n\times n}$ be regular, and the linear system $Ax = b$ be solved by Gaussian elimination with column pivoting. Then, the actually computed perturbed solution $x + \delta x$ under the influence of round-off error is exact solution of a perturbed system $(A + \delta A)(x + \delta x) = b$, where (*eps = *"machine accuracy")*

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \leq 1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\text{eps}. \tag{2.1.10}$$

In combination with the perturbation estimate of Theorem 1.8 Wilkinson's result yields the following bound on the effect of round-off errors in the Gaussian elimination:

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \leq \frac{\text{cond}(A)}{1 - \text{cond}(A)\|\delta A\|_\infty/\|A\|_\infty}\,\{1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\text{eps}\}\,. \tag{2.1.11}$$

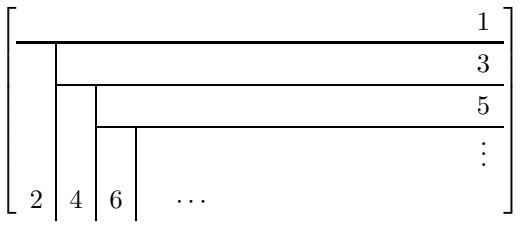
This estimate is, as practical experience shows, by far too pessimistic since it is oriented at the worst case scenario and does not take into account round-off error cancellations. Incorporating the latter effect would require a statistical analysis. Furthermore, the above estimate applies to arbitrary full matrices. For "sparse" matrices with many zero entries much more favorable estimates are to be expected. Altogether, we see that Gaussian elimination is, in principle, a well-conditioned algorithm, i. e., the influence of round-off errors is bounded in terms of the problem dimension n and the condition $\text{cond}(A)$, which described the conditioning of the numerical problem to be solved.

Direct LR and Cholesky decomposition

The Gaussian algorithm for the computation of the LR decomposition $A = LR$ (if it exists) can also be written in direct form, in which the elements l_{jk} of L and r_{jk} of R are computed recursively. The equation $A = LR$ yields n^2 equations for the n^2 unknown elements r_{jk}, $j \leq k$, l_{jk}, $j > k$ $(l_{jj} = 1)$:

$$a_{jk} = \sum_{i=1}^{\min(j,k)} l_{ji}\, r_{ik}. \tag{2.1.12}$$

Here, the ordering of the computation of l_{jk}, r_{jk} is not prescribed a priori. In the so-called "algorithm of Crout[2] " the matrix $A = LR$ is tessellated as follows:



[2]Prescott D. Crout (1907-1984): US-American mathematician and engineer; prof. at Massachusetts Institute of Technology (MIT); contributions to numerical linear algebra ("A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients", Trans. Amer. Inst. Elec. Eng. 60, 1235-1241, 1941) and to numerical electro dynamics.

The single steps of this algorithm are $(l_{ii} \equiv 1)$:

$$
\begin{aligned}
k = 1, \cdots, n: \qquad & a_{1k} = \sum_{i=1}^{1} l_{1i} r_{ik} \;\Rightarrow\; r_{1k} := a_{1k}, \\
j = 2, \cdots, n: \qquad & a_{j1} = \sum_{i=1}^{1} l_{ji} r_{i1} \;\Rightarrow\; l_{j1} := r_{11}^{-1} a_{j1}, \\
k = 2, \cdots, n: \qquad & a_{2k} = \sum_{i=1}^{2} l_{2i} r_{ik} \;\Rightarrow\; r_{2k} := a_{2k} - l_{21} r_{1k}, \\
& \vdots
\end{aligned}
$$

and generally for $j = 1, \cdots, n$:

$$
\begin{aligned}
r_{jk} &:= a_{jk} - \sum_{i=1}^{j-1} l_{ji} r_{ik}\,, \quad k = j, j+1, \cdots, n\,, \\
l_{kj} &:= r_{jj}^{-1} \left(a_{kj} - \sum_{i=1}^{j-1} l_{ki} r_{ij} \right), \quad k = j+1, j+2, \cdots, n\,.
\end{aligned}
\tag{2.1.13}
$$

The Gaussian elimination and the direct computation of the LR decomposition differ only in the ordering of the arithmetical operations and are algebraically equivalent.

2.1.2 Accuracy improvement by final iteration

The Gaussian elimination algorithm transforms a linear system $Ax = b$ into an upper triangular system $Rx = c$, from which the solution x can be obtained by simple backward substitution. Due to Theorem 2.1 this is equivalent to the determination of the decomposition $PA = LR$ and the subsequent solution of the two triangular systems

$$
Ly = Pb\,, \quad Rx = y\,. \tag{2.1.14}
$$

This variant of the Gaussian algorithm is preferable if the same linear system is successively to be solved for several right-hand sides b. Because of the unavoidable round-off error one usually obtains an only approximate LR decomposition

$$
\tilde{L}\tilde{R} \neq PA
$$

and using this in (2.1.14) an only approximate solution $x^{(0)}$ with (exact) "defect"

$$
\hat{d}^{(0)} := b - Ax^{(0)} \neq 0\,.
$$

Using the already computed approximate trianguler decomposition $\tilde{L}\tilde{R} \sim PA$, one solves (again approximately) the so-called "defect equation"

$$
Ak = \hat{d}^{(0)}, \qquad \tilde{L}\tilde{R}k^{(1)} = \hat{d}^{(0)}, \tag{2.1.15}
$$

and from this obtains a correction $k^{(1)}$ for $x^{(0)}$:

$$x^{(1)} := x^{(0)} + k^{(1)} . \tag{2.1.16}$$

Had the defect equation be solved exactly, i. e., $k^{(1)} \equiv k$, then

$$Ax^{(1)} = Ax^{(0)} + Ak = Ax^{(0)} - b + b + \hat{d}^{(0)} = b,$$

i. e., $x^{(1)} = x$ would be the exact solution of the system $Ax = b$. In general, $x^{(1)}$ is a better approximation to x than $x^{(0)}$ even if the defect equation is solved only approximately. This, however, requires the computation of the defect d with *higher* accuracy by using extended floating point arithmetic. This is supported by the following error analysis.

For simplicity, let us assume that $P = I$. We suppose the relative error in the LR decomposition of the matrix A to be bounded by a small number ε. Due to the general perturbation result of Theorem 1.8 there holds the estimate

$$\frac{\|x^{(0)} - x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)}{1 - \operatorname{cond}(A)\frac{\|A-\tilde{L}\tilde{R}\|}{\|A\|}} \underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim \varepsilon} .$$

Here, the loss of exact decimals corresponds to the condition $\operatorname{cond}(A)$. Additionally round-off errors are neglected. The exact defect $\hat{d}^{(0)}$ is replaced by the expression $d^{(0)} := b - \tilde{A}x^{(0)}$ where $\tilde{A}$ is a more accurate approximation to A,

$$\frac{\|A - \tilde{A}\|}{\|A\|} \leq \tilde{\varepsilon} \ll \varepsilon .$$

By construction there holds

$$\begin{aligned} x^{(1)} = x^{(0)} + k^{(1)} &= x^{(0)} + (\tilde{L}\tilde{R})^{-1}[b - \tilde{A}x^{(0)}] \\ &= x^{(0)} + (\tilde{L}\tilde{R})^{-1}[Ax - Ax^{(0)} + (A - \tilde{A})\, x^{(0)}], \end{aligned}$$

and, consequently,

$$\begin{aligned} x^{(1)} - x &= x^{(0)} - x - (\tilde{L}\tilde{R})^{-1}A(x^{(0)} - x) + (\tilde{L}\tilde{R})^{-1}(A - \tilde{A})\, x^{(0)} \\ &= (\tilde{L}\tilde{R})^{-1}[\tilde{L}\tilde{R} - A](x^{(0)} - x) + (\tilde{L}\tilde{R})^{-1}(A - \tilde{A})\, x^{(0)}. \end{aligned}$$

Since

$$\tilde{L}\tilde{R} = A - A + \tilde{L}\tilde{R} = A\big(I - A^{-1}(A - \tilde{L}\tilde{R})\big),$$

we can use Lemma 1.15 to conclude

$$\begin{aligned} \|(\tilde{L}\tilde{R})^{-1}\| &\leq \|A^{-1}\|\, \|\,[I - A^{-1}(A - \tilde{L}\tilde{R})]^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}(A - \tilde{L}\tilde{R})\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\, \|A - \tilde{L}\tilde{R}\|} = \frac{\|A^{-1}\|}{1 - \operatorname{cond}(A)\, \frac{\|A-\tilde{L}\tilde{R}\|}{\|A\|}}. \end{aligned}$$

This eventually implies

$$\frac{\|x^{(1)}-x\|}{\|x\|} \sim \operatorname{cond}(A)\Big[\underbrace{\frac{\|A-\tilde{L}\tilde{R}\|}{\|A\|}}_{\sim\varepsilon}\underbrace{\frac{\|x^{(0)}-x\|}{\|x\|}}_{\sim\operatorname{cond}(A)\varepsilon}+\underbrace{\frac{\|A-\tilde{A}\|}{\|A\|}}_{\sim\tilde{\varepsilon}}\frac{\|x^{(0)}\|}{\|x\|}\Big].$$

This correction procedure can be iterated to a so-called "final iteration" ("Nachiteration" in German). It may be continued until the obtained solution has an error (usually achieved after 2-3 steps) of the order of the defect computation, i. e., $\|x^{(3)}-x\|/\|x\| \sim \tilde{\varepsilon}$.

Example 2.4: The linear system

$$\begin{bmatrix} 1.05 & 1,02 \\ 1.04 & 1,02 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}=\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

has the exact solution $x=(-100,\,103.921\ldots)^T$. Gaussian elimination, with 3-decimal arithmetic and correct rounding, yields the approximate triangular matrices

$$\tilde{L}=\begin{bmatrix} 1 & 0 \\ 0.990 & 1 \end{bmatrix},\qquad \tilde{R}=\begin{bmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{bmatrix},$$

$$\tilde{L}\tilde{R}-A=\begin{bmatrix} 0 & 0 \\ 5\cdot 10^{-4} & 2\cdot 10^{-4} \end{bmatrix}\qquad \text{(correct within machine accuracy)}.$$

The resulting "solution" $x^{(0)}=(-97,1.101)^T$ has the defect

$$d^{(0)}=b-Ax^{(0)}=\begin{cases}(0,0)^T & \text{3-decimal computation,}\\ (0,065,0,035)^T & \text{6-decimal computation.}\end{cases}$$

The approximate correction equation

$$\begin{bmatrix} 1 & 0 \\ 0.990 & 1 \end{bmatrix}\begin{bmatrix} 1.05 & 1.02 \\ 0 & 0.01 \end{bmatrix}\begin{bmatrix} k_1^{(1)} \\ k_2^{(1)} \end{bmatrix}=\begin{bmatrix} 0.065 \\ 0.035 \end{bmatrix}$$

has the solution $k^{(1)}=(-2.9,102.899)^T$ (obtained by 3 decimal computation). Hence, one correction step yields the approximate solution

$$x^{(1)}=x^{(0)}+k^{(1)}=(-99.9,104)^T,$$

which is significantly more accurate than the first approximation $x^{(0)}$.

2.1.3 Inverse computation and the Gauß-Jordan algorithm

In principle, the inverse A^{-1} of a regular matrix A can be computed as follows:

(i) Computation of the LR decomposition of PA.

(ii) Solution of the staggered systems

$$Ly^{(i)} = Pe^{(i)},\ Rx^{(i)} = y^{(i)}, \quad i = 1, \ldots, n,$$

with the cartesian basis vectors e^i of $\mathbb{R}^n$.

(iii) Then, $A^{-1} = [x^{(1)}, \ldots, x^{(n)}]$.

More practical is the simultaneous elimination (without explicit determination of the matrices L and R), which directly leads to the inverse (without row perturbation):

$$\begin{array}{c|ccc} & 1 & & 0 \\ A & & \ddots & \\ & 0 & & 1 \end{array} \quad \rightarrow \quad \begin{array}{cccc|ccc} & & \text{forward elimination} & & & & \\ r_{11} & \cdots & & r_{1n} & 1 & & 0 \\ & \ddots & & \vdots & & \ddots & \\ & & & r_{nn} & * & & 1 \end{array}$$

$$\begin{array}{cccc|c} \text{backward elimination} & & & & \\ r_{11} & & 0 & & \\ & \ddots & & & * \\ 0 & & r_{nn} & & \end{array} \quad \rightarrow \quad \begin{array}{ccc|c} \text{scaling} & & & \\ 1 & & 0 & \\ & \ddots & & A^{-1} \\ 0 & & 1 & \end{array}$$

Example 2.5: The pivot elements are marked by $\boxed{\cdot}$.

$$A = \begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \quad : \quad \begin{array}{ccc|ccc} \text{forward elimination} & & & & & \\ \boxed{3} & 1 & 6 & 1 & 0 & 0 \\ 2 & 1 & 3 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{ccc|ccc} \text{row permutation} & & & & & \\ 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \\ 0 & \boxed{2/3} & -1 & -1/3 & 0 & 1 \end{array} \quad \rightarrow \quad \begin{array}{ccc|ccc} \text{forward elimination} & & & & & \\ 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 1/3 & -1 & -2/3 & 1 & 0 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{ccc|ccc} \text{backward elimination} & & & & & \\ 3 & 1 & 6 & 1 & 0 & 0 \\ 0 & 2/3 & -1 & -1/3 & 0 & 1 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array} \quad \rightarrow \quad \begin{array}{ccc|ccc} \text{backward elimination} & & & & & \\ 3 & 1 & 0 & -5 & 12 & -6 \\ 0 & 2/3 & 0 & 2/3 & -2 & 2 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{ccc|ccc} \text{scaling} & & & & & \\ 3 & 0 & 0 & -6 & 15 & -9 \\ 0 & 2/3 & 0 & 2/3 & -2 & 2 \\ 0 & 0 & -1/2 & -1/2 & 1 & -1/2 \end{array} \quad \rightarrow \quad \begin{array}{ccc|ccc} & & & & & \\ 1 & 0 & 0 & -2 & 5 & -3 \\ 0 & 1 & 0 & 1 & -3 & 3 \\ 0 & 0 & 1 & 1 & -2 & 1 \end{array}$$

$$\Rightarrow \quad A^{-1} = \begin{bmatrix} -2 & 5 & -3 \\ 1 & -3 & 3 \\ 1 & -2 & 1 \end{bmatrix}.$$

An alternative method for computing the inverse of a matrix is the so-called "exchange algorithm" (sometimes called "Gauß-Jordan[3]algorithm"). Let be given a not necessarily quadratic linear system

$$Ax = y, \qquad \text{where} \quad A \in \mathbb{R}^{m\times n}, \;\; x \in \mathbb{R}^n, \;\; y \in \mathbb{R}^m. \tag{2.1.17}$$

A solution is computed by successive substitution of components of x by those of y. If a matrix element $a_{pq} \neq 0$, then the pth equation can be solved for x_q:

$$x_q = -\frac{a_{p1}}{a_{pq}}x_1 - \ldots - \frac{a_{p,q-1}}{a_{pq}}x_{q-1} + \frac{1}{a_{pq}}y_p - \frac{a_{p,q+1}}{a_{pq}}x_{q+1} - \ldots - \frac{a_{pn}}{a_{pq}}x_n\,.$$

Substitution of x_q into the other equations

$$a_{j1}x_1 + \ldots + a_{j,q-1}x_{q-1} + a_{jq}\boxed{x_q} + a_{j,q+1}x_{q+1} + \ldots + a_{jn}x_n = y_j\,,$$

yields for $j = 1,\ldots,m$, $j \neq p$:

$$\left[a_{j1} - \frac{a_{jq}a_{p1}}{a_{pq}}\right] x_1 + \ldots + \left[a_{j,q-1} - \frac{a_{jq}a_{p,q-1}}{a_{pq}}\right] x_{q-1} + \frac{a_{jq}}{a_{pq}}y_p \quad +$$

$$+ \left[a_{j,q+1} - \frac{a_{jq}a_{p,q+1}}{a_{pq}}\right] x_{q+1} + \ldots + \left[a_{jn} - \frac{a_{jq}a_{pn}}{a_{pq}}\right] x_n \;=\; y_j\,.$$

The result is a new system, which is equivalent to the original one,

$$\tilde{A}\begin{bmatrix} x_1 \\ \vdots \\ y_p \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ x_q \\ \vdots \\ y_m \end{bmatrix}, \tag{2.1.18}$$

where the elements of the matrix $\tilde{A}$ are determined as follows:

pivot element	:	$\tilde{a}_{pq} = 1/a_{pq}\,,$
pivot row	:	$\tilde{a}_{pk} = a_{pk}/a_{pq}\,, \quad k = 1,\ldots,n\,, \quad k \neq q\,,$
pivot column	:	$\tilde{a}_{jq} = a_{jq}/a_{pq}\,, \quad j = 1,\ldots,m\,, \quad j \neq p\,,$
others	:	$\tilde{a}_{jk} = a_{jk} - a_{jq}a_{pk}/a_{pq}\,, \quad j = 1,\ldots,m,\; j \neq p,\; k = 1,\ldots,n\,,\; k \neq q.$

If we succeed with replacing all components of x by those of y the result is the solution of the system $y = A^{-1}x$. In the case $m = n$, we obtain the inverse A^{-1}, but in general with permutated rows and columns. In determining the pivot element it is advisable, for stability reasons, to chose an element a_{pq} of maximal modulus.

Theorem 2.3 (Gauß-Jordan algorithm): *In the Gauß-Jordan algorithm* $r = \operatorname{rank}(A)$ *exchange steps can be done.*

[3]Marie Ennemond Camille Jordan (1838-1922): French mathematician; prof. in Paris; contributions to algebra, group theory, calculus and topology.

Proof. Suppose the algorithm stops after r exchange steps. Let at this point $x_1, \ldots, x_r$ be exchanged against $y_1, \ldots, y_r$ so that the resulting system has the form

$$\begin{array}{l} r \\ m-r \end{array} \left\{ \begin{array}{c|c} * & * \\ \hline * & 0 \end{array} \right. \begin{bmatrix} y_1 \\ \vdots \\ y_r \\ x_{r+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_r \\ y_{r+1} \\ \vdots \\ y_m \end{bmatrix}.$$

$$\underbrace{}_{r} \quad \underbrace{}_{n-r}$$

If one chooses now $y_1 = \cdots = y_r = 0$, $x_{r+1} = \lambda_1, \cdots$, $x_n = \lambda_{n-r}$ so are all $x_1, \cdots, x_r$ uniquely determined and it follows that $y_{r+1} = \cdots = y_m = 0$. For arbitrary values $\lambda_1, \cdots, \lambda_{n-r}$ there also holds

$$A \begin{bmatrix} x_1(\lambda_1, \cdots, \lambda_{n-r}) \\ \vdots \\ x_r(\lambda_1, \cdots, \lambda_{n-r}) \\ \lambda_1 \\ \vdots \\ \lambda_{n-r} \end{bmatrix} = \begin{bmatrix} 0 \\ \\ \vdots \\ \\ 0 \end{bmatrix}.$$

Hence, $\dim(\mathrm{kern}(A)) \geq n - r$. On the other hand, because $y_1, \cdots, y_r$ can be freely chosen, we have $\dim(\mathrm{range}(A)) \geq r$. Further, observing $\dim(\mathrm{range}(A)) + \dim(\mathrm{kern}(A)) = n$ it follows that $\mathrm{rank}(A) = \dim(\mathrm{range}(A)) = r$. This completes the proof. Q.E.D.

For a quadratic linear system with regular coefficient matrix A the Gauß-Jordan algorithm for computing the inverse A^{-1} is always applicable.

Example 2.6:

$$\begin{bmatrix} 1 & 2 & 1 \\ -3 & -5 & -1 \\ -7 & -12 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$

Exchange steps: The pivot elements are marked by $\boxed{\cdot}$.

x_1	x_2	x_3	
1	2	1	y_1
−3	−5	−1	y_2
−7	$\boxed{-12}$	−2	y_3

x_1	y_3	x_3	
−1/6	−1/6	$\boxed{2/3}$	y_1
−1/12	5/12	−1/6	y_2
−7/12	−1/12	−1/6	x_2

x_1	y_3	y_1	
1/4	1/4	3/2	x_3
$\boxed{-1/8}$	3/8	−1/4	y_2
−5/8	−1/8	−1/4	x_2

y_2	y_3	y_1	
−2	1	1	x_3
−8	3	−2	x_1
5	−2	1	x_2

inverse: $\begin{bmatrix} -2 & -8 & 3 \\ 1 & 5 & -2 \\ 1 & -2 & 1 \end{bmatrix}$

Lemma 2.2: *The inversion of a regular $n \times n$-matrix by simultaneous elimination or the Gauß-Jordan algorithmalgorithmus requires*

$$N_{Gauß\text{-}Jordan}(n) = n^3 + O(n^2) \;\; a.\ op. \tag{2.1.19}$$

Proof. (i) The $n-1$ steps of forward elimination at the matrix A require $\frac{1}{3}n^3 + O(n^2)$ a. op. The simultaneous treatment of the columns of the identity matrix requires additional $\frac{1}{6}n^3 + O(n^2)$ a. op. The backward elimination for generating the identity matrix on the left requires again

$$(n-1)n + (n-2)n + \ldots + n = \tfrac{1}{2}n(n-1)n = \tfrac{1}{2}n^3 + O(n^2)$$

multiplications and additions and subsequently n^2 divisions. Hence the total work count for computing the inverse is

$$N_{\text{inverse}} = \tfrac{1}{3}n^3 + \tfrac{1}{6}n^3 + \tfrac{1}{2}n^3 + O(n^2) = n^3 + O(n^2).$$

(ii) In the Gauß-Jordan algorithm the kth exchange step requires $2n+1$ divisions in pivot row and column and $(n-1)^2$ multiplications and additions for the update of the remaining submatrix. Hence, all together $n^2 + O(n)$ a. op. The computation of the inverse requires n exchange steps so that the total work count again becomes $n^3 + O(n^2)$ a. op. Q.E.D.

2.2 Special matrices

2.2.1 Band matrices

The application of Gaussian elimination for the solution of large linear systems of size $n > 10^4$ poses technical difficulties if the primary main memory of the computer is not large enough for storing the matrices occurring during the process (fill-in problem). In this case secondary (external) memory has to be used, which increases run-time because of slower data transfer. However, many large matrices occurring in practice have special structures, which allow for memory saving in the course of Gaussian elimination.

Definition 2.3: *A matrix $A \in \mathbb{R}^{n,n}$ is called "band matrix" of "band type" (m_l, m_r) with $0 \le m_l, m_r \le n-1$, if*

$$a_{jk} = 0, \quad for \quad k < j - m_l \;\; or \;\; k > j + m_r \quad (j,k = 1,\ldots,n),$$

i. e., the elements of A outside of the main diagonal and of $m_l + m_r$ secondary diagonals are zero. The quantity $m = m_l + m_r + 1$ is called the "band width" of A.

Example 2.7: We give some very simple examples of band matrices:

Typ $(n-1, 0)$: lower triangular matrix

Typ $(0, n-1)$: upper triangular matrix

Typ $(1,1)$: tridiagonal matrix

Example of a (16×16)-band matrix of band type $(4,4)$:

$$A = \left[\begin{array}{cccc} B & -I & & \\ -I & B & -I & \\ & -I & B & -I \\ & & -I & B \end{array}\right] \Bigg\} 16, \qquad B = \left[\begin{array}{cccc} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{array}\right] \Bigg\} 4, \quad I = \left[\begin{array}{cccc} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{array}\right] \Bigg\} 4.$$

Theorem 2.4 (Band matrices): *Let $A \in \mathbb{R}^{n\times n}$ be a band matrix of band type (m_l, m_r), for which Gaussian elimination can be applied without pivoting, i. e., without permutation of rows. Then, all reduced matrices are also band matrices of the same band type and the matrix factors L and R in the triangular decomposition of A are band matrices of type $(m_l, 0)$ and $(0, m_r)$, respectively. The work count for the computation of the LR decomposition $A = LR$ is*

$$N_{LR} = \tfrac{1}{3} n m_l m_r + O(n(m_l + m_r)) \quad \textit{a. op.} \tag{2.2.20}$$

Proof. The assertion follows by direct computation (exercise). Q.E.D.

In Gaussian elimination applied to a band matrix it suffices to store the "band" of the matrix. For $n \approx 10^5$ and $m \approx 10^2$ this makes Gaussian elimination feasible at all. For the small model matrix from above (finite difference discretization of the Poisson problem) this means a reduced memory requirement of $16 \times 9 = 144$ instead of $16 \times 16 = 256$ for the full matrix. How the symmetry of A can be exploited for further memory reduction will be discussed below.

An extreme storage saving is obtained for tridiagonal matrices

$$\left[\begin{array}{cccc} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & c_n & a_n \end{array}\right].$$

Here, the elements of the LR decomposition

$$L = \left[\begin{array}{cccc} 1 & & & \\ \gamma_2 & \ddots & & \\ & \ddots & 1 & \\ & & \gamma_n & 1 \end{array}\right], \qquad R = \left[\begin{array}{cccc} \alpha_1 & \beta_1 & & \\ & \ddots & \ddots & \\ & & \alpha_{n-1} & \beta_{n-1} \\ & & & \alpha_n \end{array}\right]$$

are simply be obtained by short recursion formulas (sometimes called "Thomas[4] algorithm"),

[4]Llewellyn Thomas (1903-1992): British physicist and applied mathematician; studied at Cambridge University, since 1929 prof. of physics at Ohio State University, after the war, 1946, staff member at Watson Scientific Computing Laboratory at Columbia University, since 1968 visiting professor at North Carolina State University until retirement; best known for his contributions to atomic physics, thesis (1927) "Contributions to the theory of the motion of electrified particles through matter and some effects of that motion"; his name is frequently attached to an efficient version of the Gaussian elimination method for tridiagonal matrices.

$$\begin{array}{lllllllllll} & & & & \alpha_1 & = & a_1 & , & \beta_1 & = & b_1\,, \\ i=2,\ldots,n-1: & \gamma_i & = & c_i/\alpha_{i-1} & , & \alpha_i & = & a_i-\gamma_i\beta_{i-1} & , & \beta_i & = & b_i\,, \\ & \gamma_n & = & c_n/\alpha_{n-1} & , & \alpha_n & = & a_n-\gamma_n\beta_{n-1} & . & & & \end{array}$$

For this only $3n-2$ storage places and $2n-2$ a. op. are needed.

Frequently the band matrices are also sparse, i. e., most elements within the band are zero. However, this property cannot be used within Gaussian elimination for storage reduction because during the elimination process the whole band is filled with non-zero entries.

It is essential for the result of Theorem 2.4 that the Gaussian elimination can be carried out without perturbation of rows, i. e., without pivoting, since otherwise the bandwidth would increase in the course of the algorithm. We will now consider two important classes of matrices, for which this is the case.

2.2.2 Diagonally dominant matrices

Definition 2.4: *A matrix $A=(a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ is called "diagonally dominant", if there holds*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| \leq |a_{jj}|\,, \quad j=1,\ldots,n. \tag{2.2.21}$$

Theorem 2.5 (Existence of LR decomposition): *Let the matrix $A\in\mathbb{R}^{n\times n}$ be regular and diagonally dominant. Then, the LR decomposition $A=LR$ exists and can be computed by Gaussian elimination without pivoting.*

Proof. Since A is regular and diagonally dominant necessarily $a_{11}\neq 0$. Consequently, the first elimination step $A:=A^{(0)}\to A^{(1)}$ can be done *without* (column) pivoting. The elements $a_{jk}^{(1)}$ are obtained by $a_{1k}^{(1)}=a_{1k},\ k=1,\ldots,n$, and

$$j=2,\ldots,n\,,\quad k=1,\ldots,n: \qquad a_{jk}^{(1)}=a_{jk}-q_{j1}a_{1k}\,,\quad q_{j1}=\frac{a_{j1}}{a_{11}}\,.$$

Hence, for $j=2,\ldots,n$, there holds

$$\begin{aligned} \sum_{k=2,k\neq j}^{n} |a_{jk}^{(1)}| &\leq \sum_{k=2,k\neq j}^{n} |a_{jk}| + |q_{j1}| \sum_{k=2,k\neq j}^{n} |a_{1k}| \\ &\leq \underbrace{\sum_{k=1,k\neq j}^{n} |a_{jk}|}_{\leq |a_{jj}|} - |a_{j1}| + \underbrace{|q_{j1}|}_{=\left|\frac{a_{j1}}{a_{11}}\right|} \underbrace{\sum_{k=2}^{n} |a_{1k}|}_{\leq |a_{11}|} - |q_{j1}||a_{1j}| \\ &\leq |a_{jj}| - |q_{j1}a_{1j}| \leq |a_{jj}-q_{j1}a_{1j}| = |a_{jj}^{(1)}|. \end{aligned}$$

The matrix $A^{(1)}=G_1A^{(0)}$ is regular and obviously again diagonally dominant. Consequently, $a_{22}^{(1)}\neq 0$. This property is maintained in the course of the elimination process, i. e., the elimination is possible without any row permutations. Q.E.D.

Remark 2.1: If in (2.2.21) for all $j \in \{1, \ldots, n\}$ the strict inequality holds, then the matrix A is called "strictly diagonally dominant" . The proof of Theorem 2.5 shows that for such matrices Gaussian elimination is applicable without pivoting, i. e., such a matrix is necessarily regular. The above model matrix is diagonally dominant but not *strictly* diagonally dominant. Its regularity will be shown later by other arguments based on a slightly more restrictive assumption.

2.2.3 Positive definite matrices

We recall that a (symmetric) matrix $A \in \mathbb{R}^{n\times n}$ with the property

$$(Ax, x)_2 > 0\,, \quad x \in \mathbb{R}^n \setminus \{0\},$$

is called "positive definite".

Theorem 2.6 (Existence of LR decomposition): *For positive definite matrices $A \in R^{n\times n}$ the Gaussian elimination algorithm can be applied without pivoting and all occurring pivot elements $a_{ii}^{(i)}$ are positive.*

Proof. For the (symmetric) positive matrix A there holds $a_{11} > 0$. The relation

$$a_{jk}^{(1)} = a_{jk} - \frac{a_{j1}}{a_{11}} a_{1k} = a_{kj} - \frac{a_{k1}}{a_{11}} a_{1j} = a_{kj}^{(1)}$$

for $j, k = 2, \ldots, n$ shows that the first elimination step yields an $(n-1)\times(n-1)$-matrix $\tilde{A}^{(1)} = (a_{jk}^{(1)})_{j,k=2,\ldots,n}$, which ia again symmetric. We have to show that it is also positive definite, i. e., $a_{22}^{(1)} > 0$. The elimination process can be continued with a positive pivot element and the assertion follows by induction. Let $\tilde{x} = (x_2, \ldots, x_n)^T \in \mathbb{R}^{n-1} \setminus \{0\}$ and $x = (x_1, \tilde{x})^T \in \mathbb{R}^n$ with

$$x_1 = -\frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k} x_k\,.$$

Then,

$$\begin{aligned} 0 < \sum_{j,k=1}^{n} a_{jk} x_j x_k &= \sum_{j,k=2}^{n} a_{jk} x_j x_k + 2x_1 \sum_{k=2}^{n} a_{1k} x_k + a_{11} x_1^2 \\ &\quad \underbrace{- \frac{1}{a_{11}} \sum_{j,k=2}^{n} a_{k1} a_{1j} x_k x_j + \frac{1}{a_{11}} \Big(\sum_{k=2}^{n} a_{1k} x_k \Big)^2}_{= 0 \ (a_{jk} = a_{kj})} \\ &= \sum_{j,k=2}^{n} \underbrace{\Big(a_{jk} - \frac{a_{k1} a_{1j}}{a_{11}} \Big)}_{= a_{jk}^{(1)}} x_j x_k + a_{11} \underbrace{\Big(x_1 + \frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k} x_k \Big)^2}_{= 0} \end{aligned}$$

and, consequently, $\tilde{x}^T \tilde{A}^{(1)} \tilde{x} > 0$, what was to be proven. Q.E.D.

For positive definite matrices an LR decomposition $A = LR$ exists with positive pivot elements $r_{ii} = a_{ii}^{(i)} > 0,\ i = 1, \ldots, n$. Since $A = A^T$ there also holds

$$A = A^T = (LR)^T = (LD\tilde{R})^T = \tilde{R}^T D L^T$$

with the matrices

$$\tilde{R} = \begin{bmatrix} 1 & r_{12}/r_{11} & \cdots & r_{1n}/r_{11} \\ & \ddots & \ddots & \vdots \\ & & 1 & r_{n-1,n}/r_{n-1,n-1} \\ 0 & & & 1 \end{bmatrix}, \qquad D = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{bmatrix}.$$

In virtue of the uniqueness of the LR decomposition it follows that

$$A = LR = \tilde{R}^T D L^T,$$

and, consequently, $L = \tilde{R}^T$ and $R = DL^T$. This proves the following theorem.

Theorem 2.7: *Positive definite matrices allow for a so-called "Cholesky[5] decomposition".*

$$A = LDL^T = \tilde{L}\tilde{L}^T \tag{2.2.22}$$

with the matrix $\tilde{L} := LD^{1/2}$. *For computing the Cholesky decomposition it suffices to compute the matrices* D *and* L. *This reduces the required work count to*

$$N_{Cholesky}(n) = n^3/6 + O(n^2) \quad \textit{a. op.} \tag{2.2.23}$$

The so-called "Cholesky method" for computing the decomposition matrix

$$\tilde{L} = \begin{bmatrix} \tilde{l}_{11} & & 0 \\ \vdots & \ddots & \\ \tilde{l}_{n1} & \cdots & \tilde{l}_{nn} \end{bmatrix}$$

starts from the relation $A = \tilde{L}\tilde{L}^T$, which can be viewed as a system of $n(n+1)/2$ equations for the quantities $\tilde{l}_{jk}$, $k \le j$. Multiplicating this out,

$$\begin{bmatrix} \tilde{l}_{11} & & 0 \\ \vdots & \ddots & \\ \tilde{l}_{n1} & \cdots & \tilde{l}_{nn} \end{bmatrix} \begin{bmatrix} \tilde{l}_{11} & \cdots & \tilde{l}_{n1} \\ & \ddots & \vdots \\ 0 & & \tilde{l}_{nn} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix},$$

yields in the first column of $\tilde{L}$:

$$\tilde{l}_{11}^2 = a_{11}\,, \quad \tilde{l}_{21}\,\tilde{l}_{11} = a_{21}\,, \quad \ldots\,, \quad \tilde{l}_{n1}\,\tilde{l}_{11} = a_{n1}\,,$$

[5] *Andrè Louis Cholesky (1975-1918): French mathematician; military career as engineer officer; contributions to numerical linear algebra, "Cholesky decomposition"; killed in battle shortly before the end of World War I, his discovery was published posthumously in "Bulletin Géodésique".*

from which, we obtain

$$\tilde{l}_{11} = \sqrt{a_{11}}\,, \quad j = 2, \ldots, n: \quad \tilde{l}_{j1} = \frac{a_{j1}}{\tilde{l}_{11}} = \frac{a_{j1}}{\sqrt{a_{11}}}. \tag{2.2.24}$$

Let now for some $i \in \{2, \cdots, n\}$ the elements $\tilde{l}_{jk}$, $k = 1, \ldots, i-1$, $j = k, \ldots, n$ be already computed. Then, from

$$\begin{aligned} \tilde{l}_{i1}^2 + \tilde{l}_{i2}^2 + \ldots + \tilde{l}_{ii}^2 = a_{ii}\,, \quad \tilde{l}_{ii} > 0\,, \\ \tilde{l}_{j1}\tilde{l}_{i1} + \tilde{l}_{j2}\tilde{l}_{i2} + \ldots + \tilde{l}_{ji}\tilde{l}_{ii} = a_{ji}, \end{aligned}$$

the next elements $\tilde{l}_{ii}$ and $\tilde{l}_{ji}$, $j = i+1, \ldots, n$ can be obtained,

$$\begin{aligned} \tilde{l}_{ii} &= \sqrt{a_{ii} - \tilde{l}_{i1}^2 - \tilde{l}_{i2}^2 - \ldots - \tilde{l}_{i,i-1}^2}\,, \\ \tilde{l}_{ji} &= \tilde{l}_{ii}^{-1}\big\{a_{ji} - \tilde{l}_{j1}\tilde{l}_{i1} - \tilde{l}_{j2}\tilde{l}_{i2} - \ldots - \tilde{l}_{j,i-1}\tilde{l}_{i,i-1}\big\}, \quad j = i+1, \ldots, n, \end{aligned}$$

Example 2.8: The 3×3-matrix

$$A = \begin{bmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{bmatrix}$$

has the following (uniquely determined) Cholesky decomposition $A = LDL = \tilde{L}\tilde{L}^T$:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -4 & 5 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 9 \end{bmatrix} \begin{bmatrix} 1 & 3 & -4 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{bmatrix} \begin{bmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{bmatrix}.$$

2.3 Irregular linear systems and QR decomposition

Let $A \in \mathbb{R}^{m \times n}$ be a not necessarily quadratic coefficient matrix and $b \in \mathbb{R}^m$ a right-hand side vector. We are mainly interested in the case $m > n$ (more equations than unknowns) but also allow the case $m \leq n$. We consider the linear system

$$Ax = b, \tag{2.3.25}$$

for $x \in \mathbb{R}^n$. In the following, we again seek a vector $\hat{x} \in \mathbb{R}^n$ with minimal defect norm $\|d\|_2 = \|b - A\hat{x}\|_2$, which coincides with the usual solution concept if $\operatorname{rank}(A) = \operatorname{rank}[A, b]$. In view of Theorem 1.4 such a generalized solution is characterized as solution of the "normal equation"

$$A^T Ax = A^T b. \tag{2.3.26}$$

In the rank-deficient case, $\operatorname{rank}(A) < n$, a particular solution $\hat{x}$ of the normal system is not unique, but of the general form $\hat{x} + y$ with any element $y \in \operatorname{kern}(A)$. In this case uniqueness is achieved by requiring the "least error-squares" solution to have minimal euclidian norm, $\|\hat{x}\|_2$.

We recall that the matrix A^TA is symmetric and positive semi-definite, and even positive definite if A has maximal rank $rank(A) = n$. In the latter case the normal equation can, in principle, be solved by the Cholesky algorithm for symmetric positive definite matrices. However, in general the matrix A^TA is rather ill-conditioned. In fact, for $m = n$, we have that

$$\mathrm{cond}_2(A^TA) \sim \mathrm{cond}_2(A)^2. \tag{2.3.27}$$

Example 2.9: Using 3-decimal arithmetic, we have

$$A = \begin{bmatrix} 1.07 & 1.10 \\ 1.07 & 1.11 \\ 1.07 & 1.15 \end{bmatrix} \quad \rightarrow \quad A^TA = \begin{bmatrix} 3.43 & 3.60 \\ 3.60 & 3.76 \end{bmatrix}.$$

But A^TA is *not* positive definite: $(-1,1)\cdot A^TA\cdot(-1,1)^T = -0.01$, i. e., in this case the Cholesky algorithm will not yield a solution.

We will now describe a method, by which the normal equation can be solved without explicitly forming the product A^TA. For later purposes, from now on, we admit complex matrices.

Theorem 2.8 (QR decomposition): *Let $A \in \mathbb{K}^{m\times n}$ be any rectangular matrix with $m \geq n$ and $\mathrm{rank}(A) = n$. Then, there exists a uniquely determined orthonormal matrix $Q \in \mathbb{K}^{m\times n}$ with the property*

$$\bar{Q}^TQ = I \quad (\mathbb{K} = \mathbb{C}), \qquad Q^TQ = I \quad (\mathbb{K} = \mathbb{R}), \tag{2.3.28}$$

and a uniquely determined upper triangular matrix $R \in \mathbb{K}^{n\times n}$ with real diagonal $r_{ii} > 0$, $i = 1,\ldots,n$, such that

$$A = QR. \tag{2.3.29}$$

Proof. (i) Existence: The matrix Q is generated by successive orthonormalization of the column vectors a_k, $k = 1,\ldots,n$, of A by the Gram-Schmidt algorithm:

$$q_1 \equiv \|a_1\|_2^{-1}a_1, \quad k = 2,\ldots,n: \quad \tilde{q}_k \equiv a_k - \sum_{i=1}^{k-1}(a_k,q_i)_2\, q_i\,, \quad q_k \equiv \|\tilde{q}_k\|_2^{-1}\,\tilde{q}_k.$$

Since by assumption $\mathrm{rank}(A) = n$ the n column vectors $\{a_1,\ldots,a_n\}$ are linearly independent and the orthonormalization process does not terminate before $k = n$. By construction the matrix $Q \equiv [q_1,\ldots,q_n]$ is orthonormal. Further, for $k = 1,\ldots,n$ there holds:

$$a_k = \tilde{q}_k + \sum_{i=1}^{k-1}(a_k,q_i)_2\, q_i = \|\tilde{q}_k\|_2\, q_k + \sum_{i=1}^{k-1}(a_k,q_i)_2\, q_i$$

and

$$a_k = \sum_{i=1}^{k} r_{ik}\, q_k\,, \qquad r_{kk} \equiv \|\tilde{q}_k\|_2 \in \mathbb{R}_+\,, \quad r_{ik} \equiv (a_k,q_i)_2\,.$$

Setting $r_{ik} \equiv 0$ for $i > k$ this is equivalent to the equation $A = QR$ with the upper triangular matrix $R = (r_{ik}) \in \mathbb{K}^{n\times n}$.

(ii) Uniqueness: For proving the uniqueness of the QR decomposition let $A = Q_1R_1$ and $A = Q_2R_2$ be two such decompositions. Since R_1 and R_2 are regular and $(\det(R_i) > 0)$ it follows that

$$\begin{array}{rcccl} Q & := & \bar{Q}_2^T Q_1 & = & R_2R_1^{-1} \quad \text{right upper triangular,} \\ \bar{Q}^T & = & \bar{Q}_1^T Q_2 & = & R_1R_2^{-1} \quad \text{right upper triangular.} \end{array}$$

Since $\bar{Q}^TQ = R_1R_2^{-1}R_2R_1^{-1} = I$ it follows that Q is *orthonormal* and *diagonal* with $|\lambda_i| = 1$. From $QR_1 = R_2$, we infer that $\lambda_i r_{ii}^1 = r_{ii}^2 > 0$ and, consequently, $\lambda_i \in \mathbb{R}$ and $\lambda_i = 1$. Hence, $Q = I$, i. e.,

$$R_1 = R_2\,, \qquad Q_1 = AR_1^{-1} = AR_2^{-1} = Q_2.$$

This completes the proof. Q.E.D.

In the case $\mathbb{K} = \mathbb{R}$, using the QR decomposition, the normal equation $A^TAx = A^Tb$ transforms into

$$A^TAx = R^TQ^TQRx = R^TRx = R^TQ^Tb,$$

and, consequently, in view of the regularity of R^T,

$$Rx = Q^Tb. \tag{2.3.30}$$

This triangular system can now be solved by backward substitution in $\mathcal{O}(n^2)$ arithmetic operations. Since

$$A^TA = R^TR \tag{2.3.31}$$

with the triangular matrix R, we are given a Cholesky decomposition of A^TA without explicit computation of the matrix product A^TA.

Example 2.10: The 3×3-matzrix

$$A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}$$

has the following uniquely determined QR decomposition

$$A = QR = \begin{bmatrix} 6/7 & -69/175 & -58/5 \\ 3/7 & 158/175 & 6/175 \\ -2/7 & 6/35 & -33/35 \end{bmatrix} \cdot \begin{bmatrix} 14 & 21 & -14 \\ 0 & 175 & -70 \\ 0 & 0 & 35 \end{bmatrix}.$$

2.3.1 Householder algorithm

The Gram-Schmidt algorithm used in the proof of Theorem 2.8 for orthonormalizing the column vectors of the matrix A is not suitable in practice because of its inherent instability. Due to strong round-off effects the orthonormality of the columns of Q is quickly lost already after

only few orthonormalization steps. A more stable algorithm for this purpose is the "Householder[6] algorithm", which is described below.

For any vector $v \in \mathbb{K}^m$ the "dyadic product" is defined as the matrix

$$v\bar{v}^T := \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix} [\bar{v}_1, \ldots, \bar{v}_m] = \begin{bmatrix} |v_1|^2 & v_1\bar{v}_2 & \cdots & v_1\bar{v}_m \\ \vdots & & & \\ v_m\bar{v}_1 & v_m\bar{v}_2 & \cdots & |v_m|^2 \end{bmatrix} \in \mathbb{K}^{m\times m},$$

(not to be confused with the "scalar product" $\bar{v}^T v = \|v\|_2^2$, which maps vectors to scalars).

Definition 2.5: *For a normalized vector $v \in \mathbb{K}^n$, $\|v\|_2 = 1$, the Matrix*

$$S = I - 2v\bar{v}^T \in \mathbb{K}^{m\times m}$$

is called "Householder transformation". Obviously $S = \bar{S}^T = S^{-1}$, i. e., S (and also $\bar{S}^T$) is hermitian and unitary. Further, the product of two (unitary) Householder transformations is again unitary.

For the geometric interpretation of the Householder transformation S, we restrict us to the real case, $\mathbb{K} = \mathbb{R}$, and for an arbitrary normed vector $v \in \mathbb{R}^2$, $\|v\|_2 = 1$, consider the basis $\{v, v^\perp\}$, where $v^T v^\perp = 0$. For an arbitrary vector $u = \alpha v + \beta v^\perp \in \mathbb{R}^2$ there holds

$$\begin{aligned} Su &= (I - 2vv^T)(\alpha v + \beta v^\perp) \\ &= \alpha v + \beta v^\perp - 2\alpha (v \underbrace{v^T) v}_{=1} - 2\beta (v \underbrace{v^T) v^\perp}_{=0} = -\alpha v + \beta v^\perp . \end{aligned}$$

Hence, the application of $S = I - 2vv^T$ to a vector u in the plane $\mathrm{span}\{v, u\}$ induces a reflection of u with respect to the orthogonal axis $\mathrm{span}\{v^\perp\}$.

Starting from a matrix $A \in \mathbb{K}^{m\times n}$ the Householder algorithm in n steps generates a sequence of matrices

$$A := A^{(0)} \to \cdots \to A^{(i-1)} \to \cdots \to A^{(n)} := \tilde{R},$$

where $A^{(i-1)}$ has the form

$$A^{(i-1)} = \left[\begin{array}{ccc|ccc} * & & & & & * \\ & \ddots & & & & \vdots \\ & & * & & & \\ \hline & & & * & \cdots & * \\ & 0 & & & & \\ & & & * & \cdots & * \end{array}\right] \begin{matrix} \\ \\ \\ i \\ \\ \\ \end{matrix}$$

$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i$$

[6] Alston Scott Householder (1904-1993): US-American mathematician; Director of Oak Ridge National Laboratory (1948-1969), thereafter prof. at the U of Tennessee; worked in mathematical biology, best known for his fundamental contributions to numerics, especially to numerical linear algebra.

In the ith step the Householder transformation $S_i \in \mathbb{K}^{m\times m}$ is determined such that

$$S_i A^{(i-1)} = A^{(i)}.$$

After n steps the result is

$$\tilde{R} = A^{(n)} = S_n S_{n-1} \cdots S_1 A =: \bar{\tilde{Q}}^T A,$$

where $\bar{\tilde{Q}} \in \mathbb{K}^{m\times m}$ as product of unitary matrices is also unitary and $\tilde{R} \in \mathbb{K}^{m\times n}$ has the form

$$\tilde{R} = \left[\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ \hline 0 & \cdots & 0 \end{array}\right] \begin{array}{l} \Big\} \text{n} \\ \Big\} \text{m}. \end{array}$$

This results in the representation

$$A = \bar{S}_1^T \cdots \bar{S}_n^T \tilde{R} = \tilde{Q}\tilde{R}.$$

From this, we obtain the desired QR decomposition of A simply by striking out the last $m-n$ columns in $\tilde{Q}$ and the last $m-n$ rows in $\tilde{R}$:

$$A = \underbrace{\left[\begin{array}{c} \\ Q \\ \\ \end{array}\right.}_{\text{n}} \underbrace{\left.\begin{array}{c} \\ * \\ \\ \end{array}\right]}_{\text{m - n}} \cdot \left[\begin{array}{c} R \\ \hline 0 \end{array}\right] \begin{array}{l} \Big\} \text{n} \\ \Big\} \text{m} - \text{n} \end{array} = QR.$$

We remark that here the diagonal elements of R do not need to be positive, i. e., the Householder algorithm does generally not yield the "uniquely determined" special QR decomposition given by Theorem 2.8.

Now, we describe the transformation process in more detail. Let a_k be the column vectors of the matrix A.

Step 1: S_1 is chosen such that $S_1 a_1 \in \operatorname{span}\{e_1\}$. The vector a_1 is reflected with respect to one of the axes $\operatorname{span}\{a_1 + \|a_1\|e_1\}$ or $\operatorname{span}\{a_1 - \|a_1\|e_1\}$) into the x_1-axis. The choice of the axis is oriented by $\operatorname{sgn}(a_{11})$ in order to minimize round-off errors. In case $a_{11} \geq 0$ this choice is

$$v_1 = \frac{a_1 + \|a_1\|_2 e_1}{\|a_1 + \|a_1\|_2 e_1\|_2}, \qquad v_1^{\perp} = \frac{a_1 - \|a_1\|_2 e_1}{\|a_1 - \|a_1\|_2 e_1\|_2}.$$

Then, the matrix $A^{(1)} = (I - 2v_1\bar{v}_1^T)A$ has the column vectors

$$a_1^{(1)} = -\|a_1\|_2 e_1, \quad a_k^{(1)} = a_k - 2(a_k, v_1)v_1, \quad k = 2, \ldots, n.$$

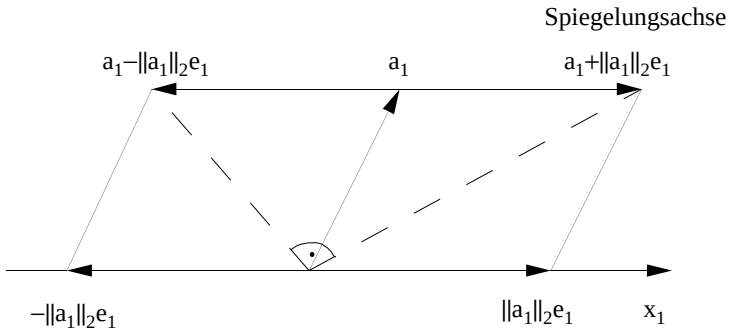


Figure 2.1: Scheme of the Householder transformation

Let now the transformed matrix $A^{(i-1)}$ be already computed.

***i*th step**: For S_i we make the following ansatz:

$$S_i = \underbrace{\left[\begin{array}{c|c} I & 0 \\ \hline 0 & I - 2\tilde{v}_i\bar{\tilde{v}}_i^T \end{array}\right]}_{\text{i - 1}} = I - 2v_i\bar{v}_i^T, \qquad v_i = \left.\left[\begin{array}{c} 0 \\ \vdots \\ 0 \\ \tilde{v}_i \end{array}\right]\right\} \begin{array}{l}\text{i - 1}\\ \\ \text{m}\end{array}$$

The application of the (unitary) matrix S_i to $A^{(i-1)}$ leaves the first $i-1$ rows and columns of $A^{(i-1)}$ unchanged. For the construction of v_i, we use the considerations of the 1st step for the submatrix:

$$\tilde{A}^{(i-1)} = \begin{bmatrix} \tilde{a}_{ii}^{(i-1)} & \cdots & \tilde{a}_{in}^{(i-1)} \\ \vdots & & \vdots \\ \tilde{a}_{mi}^{(i-1)} & \cdots & \tilde{a}_{mn}^{(i-1)} \end{bmatrix} = \left[\tilde{a}_i^{(i-1)}, \ldots, \tilde{a}_n^{(i-1)}\right].$$

It follows that

$$\tilde{v}_i = \frac{\tilde{a}_i^{(i-1)} - \|\tilde{a}_i^{(i-1)}\|_2\tilde{e}_i}{\| \quad \cdots \quad \|_2}, \quad \tilde{v}_i^{\perp} = \frac{\tilde{a}_i^{(i-1)} + \|\tilde{a}_i^{(i-1)}\|_2\tilde{e}_i}{\| \quad \cdots \quad \|_2},$$

and the matrix $A^{(i)}$ has the column vectors

$$\begin{aligned} a_k^{(i)} &= a_k^{(i-1)}, \quad k = 1, \ldots, i-1, \\ a_i^{(i)} &= (a_{1i}^{(i-1)}, \ \ldots \ , a_{i-1,i}^{(i-1)}, \ \|\tilde{a}_i^{(i-1)}\|, 0, \ldots, 0)^T, \\ a_k^{(i)} &= a_k^{(i-1)} - 2(\tilde{a}_k^{(i-1)}, \tilde{v}_i)v_i, \quad k = i+1, \ldots, n. \end{aligned}$$

Remark 2.2: For a quadratic matrix $A \in \mathbb{K}^{n\times n}$ the computation of the QR decomposition by the Householder algorithm costs about twice the work needed for the LR decomposition of A, i. e., $N_{QR} = \frac{2}{3}n^3 + \mathcal{O}(n^2)$ a. op.

2.4 Singular value decomposition

The methods for solving linear systems and equalization problems become numerically unreliable if the matrices are very ill-conditioned. It may happen that a theoretically regular matrix appears as singular for the (finite arithmetic) numerical computation or vice versa. The determination of the rank of a matrix cannot be accomplished with sufficient reliability by the LR or the QR decomposition. A more accurate approach for treating rank-deficient matrices uses the so-called "singular value decomposition (SVD)". This is a special orthogonal decomposition, which transforms the matrix from both sides. For more details, we refer to the literature, e. g., to the introductory textbook of Deuflhard & Hohmann [30].

Let $A \in \mathbb{K}^{m\times n}$ be given. Further let $Q \in \mathbb{K}^{m\times m}$ and $V \in \mathbb{K}^{n\times n}$ be orthonormal matrices. Then, the holds

$$\|QAZ\|_2 = \|A\|_2. \tag{2.4.32}$$

Hence this two-sided transformation does not change the conditioning of the matrix A. For suitable matrices Q and Z, we obtain precise information about the rank of A and the equalization problem can by accurately solved also for a rank deficient matrix. However, the numerically stable determination of such transformations is costly as will be seen below.

Theorem 2.9 (Singular value decomposition): *Let $A \in \mathbb{K}^{m\times n}$ be arbitrary real or complex. Then, there exist unitary matrices $V \in \mathbb{K}^{n\times n}$ and $U \in \mathbb{K}^{m\times m}$ such that*

$$A = U\Sigma\bar{V}^T, \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}, \quad p = \min(m, n), \tag{2.4.33}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$. Depending on whether $m \leq n$ or $m \geq n$ the matrix Σ has the form

$$\left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_m & \end{array}\right) \qquad \text{or} \qquad \left(\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ \hline & 0 & \end{array}\right).$$

Remark 2.3: The singular value decomposition $A = U\Sigma\bar{V}^T$ of a general matrix $A \in \mathbb{K}^{m\times n}$ is the natural generalization of the well-known decomposition

$$A = W\Lambda\bar{W}^T \tag{2.4.34}$$

of a square normal (and hence diagonalizable) matrix $A \in \mathbb{K}^{n\times n}$ where $\Lambda = \operatorname{diag}(\lambda_i)$, λ_i the eigenvalues of A, and $W = [w^1, \ldots, w^n]$, $\{w^1, \ldots, w^n\}$ an ONB of eigenvectors. It allows for a representation of the inverse of a general square regular matrix $A \in \mathbb{K}^{n\times n}$ in the form

$$A^{-1} = (U\Sigma\bar{V}^T)^{-1} = V^{-1}\Sigma^{-1}\bar{U}^T, \tag{2.4.35}$$

where the orthonormlity of U and V are used.

From (2.4.33), one sees that for the column vectors u^i, v^i of U, V there holds

$$Av^i = \sigma_i u^i, \quad \bar{A}^T u^i = \sigma_i v^i, \quad i = 1, \ldots, \min(m, n).$$

This implies that

$$\bar{A}^T A v^i = \sigma_i^2 v^i, \quad A\bar{A}^T u^i = \sigma_i^2 u^i,$$

which shows that the values σ_i, $i = 1, \ldots, \min(m,n)$, are the square roots of eigenvalues of the hermitian, positive semi-definite matrices $\bar{A}^T A \in \mathbb{K}^{n\times n}$ and $A\bar{A}^T \in \mathbb{K}^{m\times m}$ corresponding to the eigenvectors v^i and u^i, respectively. Hence, the σ_i are the "singular values" of the matrix A introduced in the preceding chapter. In the case $m \geq n$ the matrix $\bar{A}^T A \in \mathbb{K}^{n\times n}$ has the $p = n$ eigenvalues $\{\sigma_i^2, i = 1, \ldots, n\}$, while the matrix $A\bar{A}^T \in \mathbb{K}^{m\times m}$ has the m eigenvalues $\{\sigma_1^2, \ldots, \sigma_n^2, 0_{n+1}, \ldots, 0_m\}$. In the case $m \leq n$ the matrix $\bar{A}^T A \in \mathbb{K}^{n\times n}$ has the n eigenvalues $\{\sigma_i^2, \ldots, \sigma_m^2, 0_{m+1}, \ldots, 0_n\}$, while the matrix $A\bar{A}^T \in \mathbb{R}^{m\times m}$ has the $p = m$ eigenvalues $\{\sigma_1^2, \ldots, \sigma_m^2\}$. The existence of a decomposition (2.4.33) will be concluded by observing that $\bar{A}^T A$ is orthonormally diagonalizable,

$$\bar{Q}^T(\bar{A}^T A)Q = \mathrm{diag}(\sigma_i^2).$$

Proof of Theorem 2.9. We consider only the real case $\mathbb{K} = \mathbb{R}$.
(i) Case $m \geq n$ (overdetermined system): Let the eigenvalues of the symmetric, positive semi-definite matrix $A^T A \in \mathbb{R}^{n\times n}$ be ordered like $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$. Here, r is the rank of A and also of $A^T A$. Further, let $\{v^1, \ldots, v^n\}$ be a corresponding ONB of eigenvectors, $A^T A v^i = \lambda_i v^i$, such that the associated matrix $V := [v^1, \ldots, v^n]$ is unitary. We define the diagonal matrices $\Lambda := \mathrm{diag}(\lambda_i)$ and $\Sigma := \mathrm{diag}(\sigma_i)$ where $\sigma_i := \lambda_i^{1/2}$, $i = 1 \ldots, n$, are the "singular values" of A. In matrix notation there holds

$$AV = \Lambda V.$$

Next, we define the vectors $u^i := \sigma_i^{-1} A v^i \in \mathbb{R}^m$, $i = 1, \ldots, n$, which form an ONS in $\mathbb{R}^m$,

$$\begin{aligned}(u^i, u^j)_2 &= \sigma_i^{-1}\sigma_j^{-1}(Av^i, Av^j)_2 = \sigma_i^{-1}\sigma_j^{-1}(v^i, A^T A v^j)_2 \\ &= \sigma_i^{-1}\sigma_j^{-1}\lambda_j(v^i, v^j)_2 = \delta_{ij}, \quad i,j = 1, \ldots, n.\end{aligned}$$

The ONS $\{u^1, \ldots, u^n\}$ can be extended to an ONB $\{u^1, \ldots, u^m\}$ of $\mathbb{R}^m$ such that the associated matrix $U := [u^1, \ldots, u^m]$ is unitary. Then, in matrix notation there holds

$$A^T U = \Sigma^{-1} A^T A V = \Sigma^{-1}\Lambda V = \Sigma V, \qquad U^T A = \Sigma V^T, \qquad A = U\Sigma V^T.$$

(ii) Case $m \leq n$ (underdetermined system): We apply the result of (i) to the transposed matrix $A^T \in \mathbb{R}^{n\times m}$, obtaining

$$A^T = \tilde{U}\tilde{\Sigma}\tilde{V}^T, \qquad A = \tilde{V}\tilde{\Sigma}^T\tilde{U}^T.$$

Then, setting $U := \tilde{V}$, $V := \tilde{U}$, and observing that, in view of the above discussion, the eigenvalues of $(A^T)^T A^T = AA^T \in \mathbb{R}^{m\times m}$ are among those of $A^T A \in \mathbb{R}^{n\times n}$ besides $n - m$ zero eigenvalues. Hence, $\tilde{\Sigma}^T$ has the desired form. Q.E.D.

We now collect some important consequences of the decomposition (2.4.33). Suppose that the singular values are ordered like $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$, $p = \min(m,n)$. Then, there holds (proof exercise):

- $\mathrm{rank}(A) = r$,

- $\mathrm{kern}(A) = \mathrm{span}\{v^{r+1}, \ldots, v^n\}$,

- $\text{range}(A) = \text{span}\{u^1, \ldots, u^r\}$,
- $A = U_r \Sigma_r V_r^T \equiv \sum_{i=1}^r \sigma_i u^i v^{iT}$ (singular decomposition of A),
- $\|A\|_2 = \sigma_1 = \sigma_{\max}$,
- $\|A\|_F = (\sigma_1^2 + \cdots + \sigma_r^2)^{1/2}$ (Frobenius norm).

We now consider the problem of computing the "numerical rank" of a matrix. Let

$$\text{rank}(A, \varepsilon) = \min_{\|A-B\|_2 \le \varepsilon} \text{rank}(B)\,.$$

The matrix is called "numerically rank-deficient" if

$$\text{rank}(A, \varepsilon) < \min(m, n)\,, \qquad \varepsilon = \text{eps}\|A\|_2,$$

where eps is the "machine accuracy" (maximal relative round-off error). If the matrix elements come from experimental measurements then the parameter ε should be related to the measurement error. The concept of "numerically rank-deficient" has something in common with that of the ε-pseudospectrum discussed above.

Theorem 2.10 (Error estimate): *Let A, U, V, Σ be as in Theorem 2.9. If $k < r = \text{rank(A)}$ then in the truncated singular value decomposition,*

$$A_k = \sum_{i=1}^{k} \sigma_i u^i v^{iT},$$

there holds the estimate

$$\min_{\text{rank}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}\,.$$

This implies for $r_\varepsilon = \text{rank}(A, \varepsilon)$ the relation

$$\sigma_1 \ge \cdots \ge \sigma_{r_\varepsilon} > \varepsilon \ge \sigma_{r_\varepsilon+1} \ge \cdots \ge \sigma_p\,, \quad p = \min(m, n).$$

Proof. Since

$$U^T A_k V = \text{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$$

it follows that $\text{rank}(A_k) = k$. Further, we obtain

$$U^T (A - A_k) V = \text{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$$

and because of the orthonormality of U and V that

$$\|A - A_k\|_2 = \sigma_{k+1}.$$

It remains to show that for any other matrix B with rank k the following inequality holds

$$\|A - B\|_2 \ge \sigma_{k+1}.$$

To this end, we choose an ONB $\{x^1, \ldots, x^{n-k}\}$ of $\mathrm{kern}(B)$. For dimensional reasons there obviously holds

$$\mathrm{span}\{x^1, \ldots, x^{n-k}\} \cap \mathrm{span}\{v^1, \ldots, v^{k+1}\} \neq \emptyset.$$

Let z with $\|z\|_2 = 1$ be from this set. Then, there holds

$$Bz = 0\,, \qquad Az = \sum_{i=1}^{k+1} \sigma_i (v^{iT} z) u^i$$

and, consequently,

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 (v^{iT} z)^2 \geq \sigma_{k+1}^2.$$

Here, we have used that $z = \sum_{i=1}^{k+1} (v^{iT} z) vi$ and therefore

$$1 = \|z\|_2^2 = \sum_{i=1}^{k+1} (v^{iT} z)^2.$$

This completes the proof. Q.E.D.

With the aid of the singular value decomposition, one can also solve the equalization problem. In the following let again $m \geq n$. We have already seen that any minimal solution,

$$\|Ax - b\|_2 = \min!$$

necessarily solves the normal equation $A^T Ax = A^T b$. But this solution is unique only in the case of maximal $\mathrm{rank}(A) = n$, which may be numerically hard to verify. In this case $A^T A$ is invertible and there hold

$$x = (A^T A)^{-1} A^T b.$$

Now, knowing the (non-negative) eigenvalues λ_i, $i = 1, \ldots, n$, of $A^T A$ with corresponding ONB of eigenvectors $\{v^1, \ldots, v^n\}$ and setting $\Sigma = \mathrm{diag}(\sigma_i)$, $\sigma_i := \lambda_i^{1/2}$, $V = [v^1, \ldots, v^n]$, $u^i := \lambda_i^{-1/2} A v^i$, and $U := [u^1, \ldots, u^n]$, we have

$$(A^T A)^{-1} A^T = (V \Sigma^2 V^T)^{-1} A^T = V \Sigma^{-2} V^T A^T = V \Sigma^{-1} (AV)^T = V \Sigma^{-1} U^T.$$

This implies the solution representation

$$x = V \Sigma^{-1} U^T b = \sum_{i=1}^{n} \frac{u^{iT} b}{\sigma_i} v^i. \tag{2.4.36}$$

In the case $\mathrm{rank}(A) < n$ the normal equation has infinitely many solutions. Out of these solutions, one selects one with minimal euclidian norm, which is then uniquely determined. This particular solution is called "minimal solution" of the equalization problem. Using the singular value decomposition the solution formula (2.4.36) can be extended to this "irregular" situation.

Theorem 2.11 (Minimal solution): *Let $A = U \Sigma V^T$ be singular value decomposition of the*

matrix $A \in \mathbb{R}^{m\times n}$ *and let* $r = \operatorname{rank}(A)$. *Then,*

$$\bar{x} = \sum_{i=1}^{r} \frac{u^{iT}b}{\sigma_i} v^i$$

is the uniquely determined "minimal solution" of the normal equation. The corresponding least squares error satisfies

$$\rho^2 = \|A\bar{x} - b\|_2^2 = \sum_{i=r+1}^{m} (u^{iT}b)^2.$$

Proof. For any $x \in \mathbb{R}^n$ there holds

$$\|Ax - b\|_2^2 = \|AVV^Tx - b\|_2^2 = \|U^TAVV^Tx - U^Tb\|_2^2 = \|\Sigma V^Tx - U^Tb\|_2^2.$$

Setting $z = V^Tx$, we conclude

$$\|Ax - b\|_2^2 = \|\Sigma z - U^Tb\|_2^2 = \sum_{i=1}^{r}(\sigma_i z^i - u^{iT}b)^2 + \sum_{i=r+1}^{m} (u^{iT}b)^2.$$

Hence a minimal point necessarily satisfies

$$\sigma_i z^i = u^{iT}b, \quad i = 1, \ldots, r.$$

Among all z with this property $z^i = 0$, $i = r+1, \ldots, m$ has minimal euclidian norm. The identity for the least squares error is obvious. Q.E.D.

The uniquely determined minimal solution of the equalization problem has the following compact representation

$$\bar{x} = A^+b, \qquad \rho = \|(I - AA^+)b\|_2, \tag{2.4.37}$$

where

$$A^+ = V\Sigma^+U^T, \quad \Sigma^+ = \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_r^{-1}, 0, \ldots, 0) \in \mathbb{R}^{n\times m}.$$

The matrix

$$A^+ = V\Sigma^+U^T \tag{2.4.38}$$

is called "pseudo-inverse" of the matrix A (or "Penrose[7]inverse" (1955)). The pseudo-inverse is the unique solution of the matrix minimization problem

$$\min_{X\in\mathbb{R}^{n\times m}} \|AX - I\|_F,$$

[7]Roger Penrose (1931-): English mathematician; prof. at Birkbeck College in London (1964) and since 1973 prof. at the Univ. of Oxford; fundamental contributions to the theory of half-groups, to matrix calculus and to the theory of "tesselations" as well as in theoretical physics to cosmology, relativity and quantum mechanics.

with the Frobenius norm $\|\cdot\|_F$. Since the identity in (2.4.37) holds for all b it follows that

$$\begin{aligned} \operatorname{rank}(A) = n \quad &\Rightarrow \quad A^+ = (A^T A)^{-1} A^T, \\ \operatorname{rank}(A) = n = m \quad &\Rightarrow \quad A^+ = A^{-1}. \end{aligned}$$

In numerical practice the definition of the pseudo-inverse has to use the (suitably defined) numerical rank. The numerically stable computation of the singular value decomposition is rather costly. For details, we refer to the literature, e. g., the book of Golub & van Loan [33].

2.5 “Direct” determination of eigenvalues

In the following, we again consider general square matrices $A \in \mathbb{K}^{n\times n}$. The direct way of computing eigenvalues of A would be to follow the definition of what an eigenvalue is and to compute the zeros of the corresponding characteristic polynomial $\chi_A(z) = \det(zI - A)$ by a suitable method such as, e. g., the Newton method. However, the mathematical task of determining the zeros of a polynomial may be highly ill-conditioned if the polynomial is given in “monomial expansion”, although the original task of determining the eigenvalues of a matrix is mostly well-conditioned. This is another nice example of a mathematical problem the conditioning of which significantly depends on the choice of its formulation.

In general the eigenvalues cannot be computed via the characteristic polynomial. This is feasible only in special cases when the characteristic polynomial does not need to be explicitly built up, such as for tri-diagonal matrices or so-called “Hessenberg[8] matrices”.

Tridiagonal matrix

$$\begin{bmatrix} a_1 & b_1 & & \\ c_2 & \ddots & \ddots & \\ & \ddots & & b_{n-1} \\ & & c_n & a_n \end{bmatrix}$$

Hessenberg matrix

$$\begin{bmatrix} a_{11} & \cdots & & a_{1n} \\ a_{21} & \ddots & & \vdots \\ & \ddots & & a_{n-1,n} \\ 0 & & a_{n,n-1} & a_{nn} \end{bmatrix}$$

2.5.1 Reduction methods

We recall some properties related to the “similarity” of matrices. Two matrices $A, B \in \mathbb{C}^{n\times n}$ are “similar”, in symbols $A \sim B$, if with a regular matrix $T \in \mathbb{C}^{n\times n}$ there holds $A = T^{-1}BT$. In view of

$$\det(A - zI) = \det(T^{-1}[B - zI]T) = \det(T^{-1})\,\det(B - zI)\,\det(T) = \det(B - zI),$$

similar matrices A, B have the same characteristic polynomial and therefore also the same eigenvalues. For any eigenvalue λ of A with a corresponding eigenvector w there holds

$$Aw = T^{-1}BTw = \lambda w,$$

[8] Karl Hessenberg (1904-1959): German mathematicians; dissertation “Die Berechnung der Eigenwerte und Eigenl”osungen linearer Gleichungssysteme”, TU Darmstadt 1942.

i. e., Tw is an eigenvector of B corresponding to the same eigenvalue λ. Further, algebraic and geometric multiplicity of eigenvalues of similar matrices are the same. A "reduction method" reduces a given matrix $A \in \mathbb{C}^{n\times n}$ by a sequence of similarity transformations to a simply structured matrix for which the eigenvalue problem is then easier to solve,

$$A = A^{(0)} = T_1^{-1}A^{(1)}T_1 = Q\ldots = T_i^{-1}A^{(i)}T_i = \ldots. \tag{2.5.39}$$

In order to prepare for the following discussion of reduction methods, we recall (without proof) some basic results on matrix normal forms.

Theorem 2.12 (Jordan normal form): *Let the matrix $A \in \mathbb{C}^{n\times n}$ have the (mutually different) eigenvalues λ_i, $i = 1,\ldots,m$, with algebraic and geometric multiplicities σ_i and ρ_i, respectively. Then, there exist numbers $r_k^{(i)} \in \mathbb{N}$ $k = 1,\ldots,\rho_i$, $\sigma_i = r_1^{(i)} + \ldots + r_{\rho_i}^{(i)}$, such that A is similar to the Jordan normal form*

$$J_A = \begin{bmatrix} C_{r_1^{(1)}}(\lambda_1) & & & & & & \\ & \ddots & & & & 0 & \\ & & C_{r_{\rho_1}^{(1)}}(\lambda_1) & & & & \\ & & & \ddots & & & \\ & & & & C_{r_1^{(m)}}(\lambda_m) & & \\ & 0 & & & & \ddots & \\ & & & & & & C_{r_{\rho_m}^{(m)}}(\lambda_m) \end{bmatrix}.$$

Here, the numbers $r_k^{(i)}$ are up to their ordering uniquely determined.

The following theorem of Schur[9] concerns the case that in the similarity transformation only unitary matrices are allowed.

Theorem 2.13 (Schur normal form): *Let the matrix $A \in \mathbb{C}^{n\times n}$ have the eigenvalues λ_i, $i = 1,\ldots,n$ (counted accordingly to their algebraic multiplicities). Then, there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ such that*

$$\bar{U}^T AU = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}. \tag{2.5.40}$$

If $A \in \mathbb{C}^{n\times n}$ is hermitian, $A^T = \bar{A}$, so is also $\bar{U}^T AU$ hermitian. Hence, hermitian matrices $A \in \mathbb{C}^{n\times n}$ are "unitary similar" to a diagonal matrix $\bar{U}^T AU = diag(\lambda_i)$, i. e., "diagonalizable".

Lemma 2.3 (Diagonalization): *For any matrix $A \in \mathbb{C}^{n\times n}$ the following statements are equivalent:*

[9] Issai Schur (1875-1941): Russian-German mathematician; prof. in Bonn (1911-1916) and in Berlin (1916-1935), where he founded a famous mathematical school; because of his jewish origin persecuted he emigrated 1939 to Palestine; fundamental contributions especially to the representation theory of groups and to number theory.

(i) A is diagonalizable.

(ii) There exists an ONB in $\mathbb{C}^n$ of eigenvectors of A.

(iii) For all eigenvalues of A algebraic and geometric multiplicity coincide.

In general, the direct transformation of a given matrix into normal form in finitely many steps is possible only if all its eigenvectors are a priori known. Therefore, first one transforms the matrix in finitely many steps into a similar matrix of simpler structure (e. g., Hessenberg form) and afterwards applies other mostly iterative methods of the form

$$A = A^{(0)} \to A^{(1)} = T_1^{-1} A^{(0)} T_1 \to \dots A^{(m)} = T_m^{-1} A^{(m-1)} T_m.$$

Here, the transformation matrices T_i should be given explicitly in terms of the elements of $A^{(i-1)}$. Further, the eigenvalue problem of the matrix $A^{(i)} = T_i^{-1} A^{(i-1)} T_i$ should not be worse conditioned than that of $A^{(i-1)}$.

Let $\|\cdot\|$ be any natural matrix norm generated by a vector norm $\|\cdot\|$ on $\mathbb{C}^n$. For two similar matrices $B \sim A$ there holds

$$B = T^{-1} A T\,, \quad B + \delta B = T^{-1}(A + \delta A)\, T\,, \quad \delta A = T \delta B T^{-1},$$

and, therefore,

$$\|B\| \le \operatorname{cond}(T)\, \|A\|\,, \quad \|\delta A\| \le \operatorname{cond}(T)\, \|\delta B\|.$$

This implies that

$$\frac{\|\delta A\|}{\|A\|} \le \operatorname{cond}(T)^2\, \frac{\|\delta B\|}{\|B\|}. \tag{2.5.41}$$

Hence, for large $\operatorname{cond}(T) \gg 1$ even small perturbations in B may effect its eigenvalues significantly more than those in A. In order to guarantee the stability of the reduction approach, in view of

$$\operatorname{cond}(T) = \operatorname{cond}(T_1 \dots T_m) \le \operatorname{cond}(T_1) \cdot \ \dots \ \cdot \operatorname{cond}(T_m),$$

the transformation matrices T_i are to be chosen such that $\operatorname{cond}(T_i)$ does not become too large. This is especially achieved for the following three types of transformations:

a) Rotations (Givens transformation):

$$T = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos(\varphi) & & & -\sin(\varphi) & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & \sin(\varphi) & & & \cos(\varphi) & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{bmatrix} \Longrightarrow \operatorname{cond}_2(T) = 1.$$

b) Reflections (Householder transformation):

$$T = I - 2u\bar{u}^T \implies \mathrm{cond}_2(T) = 1.$$

The Householder transformations are *unitary* with spectral condition $\mathrm{cond}_2(T) = 1$.

c) Elimination

$$T = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{i+1,i} & 1 & \\ & & \vdots & & \ddots \\ & & l_{n,i} & & 1 \end{bmatrix}, \quad |l_{jk}| \le 1 \implies \mathrm{cond}_\infty(T) \le 4.$$

In the following, we consider only the eigenvalue problem of real matrices. The following theorem provides the basis of the so-called "Householder algorithm".

Theorem 2.14 (Hessenberg normal form): *To each matrix $A \in \mathbb{R}^{n\times n}$ there exists a sequence of Householder matrices T_i, $i = 1, \ldots, n-2$, such that TAT^T with $T = T_{n-2} \ldots T_1$ is a Hessenberg matrix . For symmetric A the transformed matrix TAT^T is tri-diagonal.*

Proof. Let $A = [a^1, \ldots, a^n]$ and a^k the column vectors of A. In the first step $u_1 = (0, u_{12}, \ldots, u_{1n})^T \in \mathbb{R}^n$, $\|u_1\|_2 = 1$, is determines such that with $T_1 = I - 2u_1u_1^T$ there holds $T_1a^1 \in \mathrm{span}\{e^1, e2\}$. Then,

$$A^{(1)} = T_1AT_1 = \underbrace{\left[\begin{array}{c|ccc} a_{11} & a_{12} & \ldots & a_{1n} \\ \hline //// & * & & \\ \hline 0 & & & \end{array}\right]}_{T_1A} \underbrace{\left[\begin{array}{c|cc} 1 & 0 & \ldots \\ \hline 0 & * & \\ \vdots & & \end{array}\right]}_{T_1^T} = \left[\begin{array}{c|c} a_{11} & * \\ \hline //// & \\ \hline 0 & \tilde{A}^{(1)} \end{array}\right].$$

In the next step, we apply the same procedure to the reduced matrix $\tilde{A}^{(1)}$. After $n-2$ steps, we obtain a matrix $A^{(n-2)}$ which has Hessenberg form. With A also $A^{(1)} = T_1AT_1$ is symmetric and then also $A^{(n-2)}$. The symmetric Hessenberg matrix $A^{(n-2)}$ is tri-diagonal. Q.E.D.

Remark 2.4: For a symmetric matrix $A \in \mathbb{R}^{n\times n}$ the Householder algorithm for reducing it to tri-diagonal form requires $\frac{2}{3}n^3 + O(n^2)$ a. op. and the reduction of a general matrix to Hessenberg form $\frac{5}{3}n^3 + O(n^2)$ a. op. For this purpose the alternative method of Wilkinson[10] using Gaussian elimination steps and row permutation is more efficient as it requires only half as many arithmetic operations. However, the row permutation destroys the possible symmetry of the original matrix. The oldest method for reducing a real symmetric matrix to tri-diagonal

[10] James Hardy Wilkinson (1919-1986): English mathematician; worked at National Physical Laboratory in London (since 1946); fundamental contributions to numerical linear algebra, especially to round-off error analysis; co-founder of the famous NAG software library (1970).

form is goes back to Givens[11] (1958). It uses (unitary) Givens rotation matrices. Since this algorithm requires twice as many arithmetic operations as the Householder algorithm it is not further discussed. For details, we refer to the literature, e. g., the textbook of Stoer & Bulirsch II [47].

2.5.2 Hyman's method

The classical method for computing the eigenvalues of a tri-diagonal or Hessenberg matrix is based on the characteristic polynomial without explicitly determining the coefficients in its monomial expansion. The method of Hyman[12] (1957) computes the characteristic polynomial $\chi_A(\cdot)$ of a Hessenberg matrix $A \in \mathbb{R}^{n\times n}$. Let as assume that the matrix A does not separate into two submatrices of Hessenberg form, i. e., $a_{j+1,j} \neq 0$, $j = 1, \ldots, n-1$. With a function $c(\cdot)$ still to be chosen, we consider the linear system

$$\begin{aligned}
(a_{11} - z)x_1 + a_{12}x_2 + \cdots + a_{1,n-1}x_{n-1} + a_{1n}x_n &= -c\,(z)\\
a_{21}x_1 + (a_{22} - z)x_2 + \cdots + a_{2,n-1}x_{n-1} + a_{2n}x_n &= 0\\
&\vdots\\
a_{n,n-1}x_{n-1} + (a_{nn} - z)x_n &= 0.
\end{aligned}$$

Setting $x_n = 1$ the values $x_{n-1}, \ldots, x_1$ and $c\,(z)$ can be successively determined. By Cramer's rule there holds

$$1 = x_n = \frac{(-1)^n c(z) a_{21}a_{32}\ldots a_{n,n-1}}{\det(A - zI)}.$$

Consequently, $c(z) = \text{const.}\,\det(A - zI)$, and we obtain a recursion formula for determining the characteristic polynomial $\chi_A(z) = \det(zI - A)$.

Let now $A \in \mathbb{R}^{n\times n}$ be a symmetric tri-diagonal matrix with entries $b_i \neq 0$, $i = 1, \ldots, n-1$:

$$A = \begin{bmatrix} a_1 & b_1 & & 0\\ b_1 & & \ddots & \\ & \ddots & \ddots & b_{n-1}\\ 0 & & b_{n-1} & a_n \end{bmatrix}.$$

For the computation of the characteristic polynomial $\chi_A(\cdot)$, we have the recursion formulas

$$p_0(z) = 1\,, \quad p_1(z) = a_1 - z, \qquad p_i(z) = (a_i - z)\,p_{i-1}(z) - b_{i-1}^2 p_{i-2}(z), \quad i = 2, \ldots, n.$$

The polynomials $p_i \in P_i$ are the ith principle minors of $\det(zI - A)$, i. e., $p_n = \chi_A$. To see

[11] James Wallace Givens, 1910-1993: US-American mathematician; worked at Oak Ridge National Laboratory; known by the named after him matrix transformation “Givens rotation” (“Computation of plane unitary rotations transforming a general matrix to triangular form”, SIAM J. Anal. Math. 6, 26-50, 1958).

[12] M. A. Hyman: Eigenvalues and eigenvectors of general matrices, Twelfth National Meeting A.C.M., Houston, Texas, 1957.

this, we expand the $(i+1)$th principle minor with respect to the $(i+1)$th column:

$$\left[\begin{array}{ccc|c|c|cc}
a_1 - z & b_1 & & & & & \\
b_1 & \ddots & \ddots & & & & \\
 & \ddots & & & & & \\
\cline{1-3}
 & & & b_{i-1} & & & \\
 & & b_{i-1} & a_i - z & b_i & & \\
\cline{1-4}
 & & & b_i & a_{i+1} - z & \ddots & \\
\cline{1-5}
 & & & & \ddots & \ddots & \\
\end{array}\right] = \underbrace{(a_{i+1} - z)p_i(z) - b_i^2 p_{i-1}(z)}_{=:\, p_{i+1}(z)} .$$

$$\qquad\qquad i-1 \qquad i \qquad i+1$$

Often it is useful to know the derivative $\chi_A'(\cdot)$ of $\chi_A(\cdot)$ (e. g., in using the Newton method for computing the zeros of $\chi_A(\cdot)$). This is achieved by the recursion formula

$$\begin{aligned}
q_0(z) &= 0\,, \quad q_1(z) = -1 \\
q_i(z) &= -p_{i-1}(z) + (a_i - z)q_{i-1}(z) - b_{i-1}^2 q_{i-2}(z)\,, \quad i = 2, \dots, n\,, \\
q_n(z) &= \chi_A'(z)\,.
\end{aligned}$$

If the zero λ of χ_A, i. e., an eigenvalue of A, has been determined a corresponding eigenvector $w\,(\lambda)$ is given by

$$w\,(z) = \begin{bmatrix} w_0\,(z) \\ \vdots \\ w_{n-1}(z) \end{bmatrix}, \qquad \begin{aligned} w_0(z) \;&\equiv\; 1 \quad (b_n := 1) \\ w_i(z) \;&:=\; \frac{(-1)^i p_i(z)}{b_1 \dots b_i}\,, \quad i = 1, \dots, n\,. \end{aligned} \tag{2.5.42}$$

For verifying this, we compute $(A - zI)\,w(z)$. For $i = 1, \dots, n-1$ $(b_0 := 0)$ there holds

$$\begin{aligned}
&b_{i-1}w_{i-2}(z) + a_i w_{i-1}(z) + b_i w_i(z) - z\,w_{i-1}(z) = \\
&= b_{i-1}(-1)^{i-2}\,\frac{p_{i-2}(z)}{b_1 \dots b_{i-2}} + a_i(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \dots b_{i-1}} + b_i(-1)^i\,\frac{p_i(z)}{b_1 \dots b_i} - z(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \dots b_{i-1}} \\
&= b_{i-1}^2(-1)^{i-2}\,\frac{p_{i-2}(z)}{b_1 \dots b_{i-1}} + a_i(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \dots b_{i-1}} + (-1)^i\,\frac{(a_i - z)\,p_{i-1}(z) - b_{i-1}^2 p_{i-2}(z)}{b_1 \dots b_{i-1}} \\
&\quad - z(-1)^{i-1}\,\frac{p_{i-1}(z)}{b_1 \dots b_{i-1}} = 0.
\end{aligned}$$

Further, for $i = n$ $(b_n := 1)$:

$$\begin{aligned}
b_{n-1}w_{n-2}(z) + a_n w_{n-1}(z) - z\,w_{n-1}(z) &= \\
&= b_{n-1}(-1)^{n-2}\,\frac{p_{n-2}(z)}{b_1 \dots b_{n-2}} + (a_n - z)(-1)^{n-1}\,\frac{p_{n-1}(z)}{b_1 \dots b_{n-1}} \\
&= -b_{n-1}^2(-1)^{n-1}\,\frac{p_{n-2}(z)}{b_1 \dots b_{n-1}} + (a_n - z)(-1)^{n-1}\,\frac{p_{n-1}(z)}{b_1 \dots b_{n-1}} \\
&= (-1)^{n-1}\,\frac{p_n(z)}{b_1 \dots b_{n-1}} = -w_n(z)\,.
\end{aligned}$$

Hence, we have

$$(A - zI)w(z) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -w_n(z) \end{bmatrix}. \tag{2.5.43}$$

For an eigenvalue λ of A there is $w_n(\lambda) = \text{const.}\, p_A(\lambda) = 0$, i. e., $(A - \lambda I)\, w(\lambda) = 0$.

2.5.3 Sturm's method

We will now describe a method for the determination of zeros of the characteristic polynomial χ_A of a real symmetric (irreducible) tridiagonal matrix $A \in \mathbb{R}^{n\times n}$. Differentiating in the identity (2.5.43) yields

$$[(A - zI)w(z)]' = -w(z) + (A - zI)w'(z) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -w_n'(z) \end{bmatrix}.$$

We set $z = \lambda$ with some eigenvalue λ of A and multiply by $-w(\lambda)$ to obtain

$$\begin{aligned} 0 \;&<\; \|w(\lambda)\|_2^2 - (\underbrace{[A - \lambda I]w(\lambda)}_{=0}, w'(\lambda)) \\ &=\; w_{n-1}(\lambda)w_n'(\lambda) = -\frac{p_{n-1}(\lambda)p_n'(\lambda)}{b_1^2 \dots b_{n-1}^2}. \end{aligned}$$

Consequently, $p_n'(\lambda) \neq 0$, i. e., there generally holds

(S1) *All zeros of* p_n *are simple.*

Further:

(S2) *For each zero* λ *of* p_n: $\quad p_{n-1}(\lambda)p_n'(\lambda) < 0.$

(S3) *For each real zero* ζ *of* p_{i-1}: $\quad p_i(\zeta)p_{i-2}(\zeta) < 0,\ i = 2, \dots, n;$
since in this case $p_i(\zeta) = -b_{i-1}^2 p_{i-2}(\zeta)$ and were $p_i(\zeta) = 0$ this would result in the contradiction

$$0 = p_i(\zeta) = p_{i-1}(\zeta) = p_{i-2}(\zeta) = \dots = p_0(\zeta) = 1.$$

Finally, there trivially holds:

(S4) $p_0 \neq 0$ *does not change sign.*

Definition 2.6: *A sequence of polynomials* $p = p_n, p_{n-1}, \dots, p_0$ *(or more general of continuous functions* $f_n, f_{n-1}, \dots, f_0$*) with the properties (S1)-(S4) is called a "Sturm*[13] *chain" of* p.

[13] *Jacques Charles Franois Sturm (1803-1855): French-Swiss mathematician; prof. at École Polytechnique in*

The preceding consideration has led us to the following result:

Theorem 2.15 (Sturm chain): *Let $A \in \mathbb{R}^{n\times n}$ be a symmetric, irreducible tri-diagonal matrix. Then, the principle minors $p_i(z)$ of the matrix $A - zI$ form a Sturm chain of the characteristic polynomial $\chi_A(z) = p_n(z)$ of A.*

The value of the existence of a Sturm chain of a polynomial p consists in the following result.

Theorem 2.16 (Bisection method): *Let p be a polynomial and $p = p_n, p_{n-1}, \ldots, p_0$ a corresponding Sturm chain. Then, the number of real zeros of p in an interval $[a, b]$ equals $N(b) - N(a)$, where $N(\zeta)$ is the number of sign changes in the chain $p_n(\zeta), \ldots, p_0(\zeta)$.*

Proof. We consider the number of sign changes $N(a)$ for increasing a. $N(a)$ remains constant as long as a does not pass a zero of one of the p_i. Let now a be a zero of one of the p_i. We distinguish two cases:

(i) Case $p_i(a) = 0$ for $i \neq n$: In this case $p_{i+1}(a) \neq 0$, $p_{i-1}(a) \neq 0$. Therefore, the sign of $p_j(a)$, $j \in \{i-1, i, i+1\}$ shows for sufficiently small $h > 0$ a behavior that is described by one of the following two tables:

	$a-h$	a	$a+h$
$i-1$	$-$	$-$	$-$
i	$+/-$	0	$-/+$
$i+1$	$+$	$+$	$+$

	$a-h$	a	$a+h$
$i-1$	$+$	$+$	$+$
i	$+/-$	0	$-/+$
$i+1$	$-$	$-$	$-$

In each case $N(a-h) = N(a) = N(a+h)$ and the number of sign changes does not change.

(ii) Case $p_n(a) = 0$: In this case the behavior of $p_j(a)$, $j \in \{n-1, n\}$, is described by one of the following two tables (because of (S2)):

	$a-h$	a	$a+h$
n	$-$	0	$+$
$n-1$	$-$	$-$	$-$

	$a-h$	a	$a+h$
n	$+$	0	$-$
$n-1$	$+$	$+$	$+$

Further, there holds $N(a-h) = N(a) = N(a+h) - 1$, i. e., passing a zero of p_n causes one more sign change. For $a < b$ and $h > 0$ sufficiently small the difference $N(b) - N(a) = N(b+h) - N(a-h)$ equals the number of zeros of p_n in the interval $[a-h, b+h]$. Since h can be chosen arbitrarily small the assertion follows. Q.E.D.

Theorem 2.15 suggests a simple bisection method for the approximation of roots of the characteristic polynomial χ_A of a symmetric, irreduzible tridiagonal matrix $A \in \mathbb{R}^{n\times n}$. Obviously, A has only real, simple eigenvalues

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n.$$

Paris since 1840; contributions to mathematical physics, differential equations, ("Sturm-Liouville problem") and differential geometry.

For $x \to -\infty$ the chain

$$\begin{aligned} &p_0(x) = 1\,, \quad p_1(x) = a_1 - x \\ &i = 2, \ldots, n: \quad p_i(x) = (a_i - x)p_{i-1}(x) - b_i^2 p_{i-2}(x)\,, \end{aligned}$$

has the sign distribution $+, \ldots, +$, which shows that $N(x) = 0$. Consequently, $N(\zeta)$ corresponds to the number of zeros λ of χ_A with $\lambda < \zeta$. For the eigenvalues λ_i of A it follows that

$$\lambda_i < \zeta \quad \Longleftrightarrow \quad N(\zeta) \geq i. \tag{2.5.44}$$

In order to determine the ith eigenvalue λ_i, one starts from an interval $[a_0, b_0]$ containing λ_i, i. e., $a_0 < \lambda_1 < \lambda_n < b_0$. Then the interval is bisected and it is tested using the Sturm sequence, which of the both new subintervals λ_i contains λ_i. Continuing this process for $t = 0, 1, 2, \ldots$, one obtains:

$$\mu_t := \frac{a_t + b_t}{2}, \qquad \begin{aligned} a_{t+1} &:= \begin{cases} a_t\,, & \text{for} \quad N(\mu_t) \geq i \\ \mu_t\,, & \text{for} \quad N(\mu_t) < i \end{cases} \\ b_{t+1} &:= \begin{cases} \mu_t\,, & \text{for} \quad N(\mu_t) \geq i \\ b_t\,, & \text{for} \quad N(\mu_t) < i \end{cases} \end{aligned} \tag{2.5.45}$$

By construction, we have $\lambda_i \in [a_{t+1}, b_{t+1}]$ and

$$[a_{t+1}, b_{t+1}] \subset [a_t, b_t]\,, \quad |a_{t+1} - b_{t+1}| = \tfrac{1}{2}|a_t - b_t|,$$

i. e., the points a_t converge monotonically increasing and b_t monotonically decreasing to λ_i. This algorithm is slow but very robust with respect to round-off perturbations and allows for the determination of any eigenvalue of A independently of the others.

2.6 Exercises

Exercise 2.1: a) Construct examples of real matrices, which are symmetric, diagonally dominant and regular but indefinite (i. e. neither positive definite, nor negative definite), and vice versa those, which are positive (or negative) definite but not diagonally dominant. This demonstrates that these two properties of matrices are independent of each other.

b) Show that a matrix $A \in \mathbb{K}^{n\times n}$ for which the conjugate transpose $\bar{A}^T$ is strictly diagonally dominant is regular.

c) Show that a strictly diagonally dominant real matrix, which is symmetric and has positive diagonal elements is positive definite.

Exercise 2.2: Let $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ be a symmetric, positive definite matrix. The Gaussian elimination algorithm (without pivoting) generates a sequence of matrices $A = A^{(0)} \to \ldots \to A^{(k)} \to \ldots \to A^{(n-1)} = R$, where $R = (r_{ij})_{i,j=1}^n$ is the resulting upper-right triangular matrix. Prove that the algorithm is "stable" in the following sense:

$$k = 1, \ldots, n-1: \qquad a_{ii}^{(k)} \leq a_{ii}^{(k-1)}, \;\; i = 1, \ldots, n, \quad \max_{1\leq i,j\leq n} |r_{ij}| \leq \max_{1\leq i,j\leq n} |a_{ij}|.$$

(Hint: Use the recursion formula and employ an inductive argument.)

Exercise 2.3: The "LR-decomposition" of a regular matrix $A \in \mathbb{R}^{n\times n}$ is the representation of A as a product $A = LR$ consisting of a lower-left triangular matrix L with normalized diagonal ($l_{ii} = 1, 1 \le i \le n$) and an upper-right triangular matrix R.

(i) Verify that the set of all (regular) lower-left triangular matrices $L \in \mathbb{R}^{n\times n}$, with normalized diagonal ($l_{ii} = 1$, $i = 1, \ldots, n$), as well as the set of all regular, upper-right triangular matrices $R \in \mathbb{R}^{n\times n}$ form groups with respect to matrix multiplication. Are these groups abelian?

(ii) Use the result of (i) to prove that if the LR-decomposition of a regular matrix $A \in \mathbb{R}^{n\times n}$ exists, it must be unique.

Exercise 2.4: Let $A \in \mathbb{R}^{n\times n}$ be a regular matrix that admits an "LR-decomposition". In the class it was stated that Gaussian elimination (without pivoting) has an algorithmic complexity of $\frac{1}{3}n^3 + O(n^2)$ a. op., and that in case of a symmetric matrix this reduces to $\frac{1}{6}n^3 + O(n^2)$ a. op. Hereby, an "a. op." (arithmetic operation) consists of exactly one multiplication (with addition) or of a division.

Question: What are the algorithmic complexities of these algorithms in case of a band matrix of type (m_l, m_r) with $m_l = m_r = m$? Give explicit numbers for the model matrix introduced in the lecture with $m = 10^2$, $n = m^2 = 10^4$, and $m = 10^4$, $n = m^2 = 10^8$, respectively.

Exercise 2.5: Consider the linear system $Ax = b$ where

$$\begin{bmatrix} 1 & 3 & -4 \\ 3 & 9 & -2 \\ 4 & 12 & -6 \\ 2 & 6 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

a) Investigate whether this system is solvable (with argument).

b) Determine the least-error squares solution of the system ("minimal solution").

c) Is this "solution" unique?

d) Are the matrices A^TA and AA^T (symmetric) positive definit?

3 Iterative Methods for Linear Algebraic Systems

In this chapter, we discuss *iterative* methods for solving linear systems. The underlying problem has the form

$$Ax = b, \tag{3.0.1}$$

with a *real* square matrix $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}$ and a vector $b = (b_j)_{j=1}^n \in \mathbb{R}^n$. Here, we concentrate on the higher-dimensional case $n \gg 10^3$, such that, besides arithmetical complexity, also storage requirement becomes an important issue. In practice, high-dimensional matrices usually have very special structure, e. g., band structure and extreme sparsity, which needs to be exploited by the solution algorithms. The most cost-intensive parts of the considered algorithms are simple matrix-vector multiplications $x \to Ax$. Most of the considered methods and results are also applicable in the case of matrices and right-hand sides with *complex* entries.

3.1 Fixed-point iteration and defect correction

For the construction of cheap iterative methods for solving problem (3.0.1), one rewrites it in form of an equivalent fixed-point problem,

$$Ax = b \quad \leftrightarrow \quad Cx = Cx - Ax + b \quad \leftrightarrow \quad x = (I - C^{-1}A)x + C^{-1}b,$$

with a suitable regular matrix $C \in \mathbb{R}^{n\times n}$, the so-called "preconditioner". Then, starting from some initial value x^0, one uses a simple fixed-point iteration,

$$x^t = \underbrace{(I - C^{-1}A)}_{=:\,B}\, x^{t-1} + \underbrace{C^{-1}b}_{=:\,c}, \quad t = 1, 2, \ldots . \tag{3.1.2}$$

Here, the matrix $B = I - C^{-1}A$ is called the "iteration matrix" of the fixed-point iteration. Its properties are decisive for the convergence of the method. In practice, such a fixed-point iteration is organized in form of a "defect correction" iteration, which essentially requires in each step only a matrix-vector multiplication and the solution of a linear system with the matrix C as coefficient matrix:

$$d^{t-1} = b - Ax^{t-1} \text{ (defect)}, \quad C\delta x^t = d^{t-1} \text{ (correction)}, \quad x^t = x^{t-1} + \delta x^t \text{ (update)}.$$

Example 3.1: The simplest method of this type is the *(damped) Richardson*[1] *method*, which for a suitable parameter $\theta \in (0, 2\lambda_{max}(A)^{-1}]$ uses the matrices

$$C = \theta^{-1}I, \qquad B = I - \theta A. \tag{3.1.3}$$

Starting from some initial value x^0 the iteration looks like

$$x^t = x^{t-1} + \theta(b - Ax^{t-1}), \quad t = 1, 2, \ldots . \tag{3.1.4}$$

[1] Lewis Fry Richardson (1881-1953): English mathematician and physicist; worked at several institutions in in England and Scotland; a typical "applied mathematician"; pioneered modeling and numerics in weather prediction.

In view of the Banach fixed-point theorem a sufficient criterion for the convergence of the fixed-point iteration (3.1.2) is the contraction property of the corresponding fixed-point mapping $g(x) := Bx + c$,

$$\|g(x) - g(y)\| = \|B(x-y)\| \le \|B\|\|x-y\|, \qquad \|B\| < 1,$$

in some vector norm $\|\cdot\|$. For a given iteration matrix B the property $\|B\| < 1$ may depend on the particular choice of the norm. Hence, it is desirable to characterize the convergence of this iteration in terms of norm-independent properties of B. For this, the appropriate quantity is the "spectral radius"

$$\mathrm{spr}\,(B) := \max\{\,|\lambda| : \lambda \in \sigma(B)\,\}.$$

Obviously, $\mathrm{spr}\,(B)$ is the radius of the smallest circle in $\mathbb{C}$ around the origin, which contains all eigenvalues of B. For any natural matrix norm $\|\cdot\|$, there holds

$$\mathrm{spr}\,(B) \le \|B\|. \tag{3.1.5}$$

For symmetric B, we even have

$$\mathrm{spr}\,(B) = \|B\|_2 = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Bx\|_2}{\|x\|_2}. \tag{3.1.6}$$

However, we note that $\mathrm{spr}\,(\cdot)$ does not define a norm on $\mathbb{R}^{n\times n}$ since the triangle inequality does not hold in general.

Theorem 3.1 (fixed-point iteration): *The fixed-point iteration (3.1.2) converges for any starting value x^0 if and only if*

$$\rho := \mathrm{spr}(B) < 1. \tag{3.1.7}$$

In case of convergence the limit is the uniquely determined fixed point x. The asymptotic convergence behavior with respect to any vector norm $\|\cdot\|$ is characterized by

$$\sup_{x^0\in\mathbb{R}^n} \limsup_{t\to\infty} \Big(\frac{\|x^t - x\|}{\|x^0 - x\|}\Big)^{1/t} = \rho. \tag{3.1.8}$$

Hence, the number of iteration steps necessary for an asymptotic error reduction by a small factor $TOL > 0$ is approximately given by

$$t(\varepsilon) \approx \frac{\ln(1/TOL)}{\ln(1/\rho)}. \tag{3.1.9}$$

Proof. Assuming the existence of a fixed point x, we introduce the notation $e^t := x^t - x$. Recalling that $x = Bx + c$, we find

$$e^t = x^t - x = Bx^{t-1} + c - (Bx + c) = Be^{t-1} = \cdots = B^t e^0.$$

(i) In case that $\mathrm{spr}\,(B) < 1$ in view of Lemma 3.1 below, there exists a vector norm $\|\cdot\|_{B,\varepsilon}$ depending on B and some $\varepsilon > 0$ chosen sufficiently small, such that the corresponding natural matrix norm $\|\cdot\|_{B,\varepsilon}$ satisfies

$$\|B\|_{B,\varepsilon} \leq \operatorname{spr}(B) + \varepsilon = \rho + \varepsilon < 1.$$

Consequently, by the Banach fixed-point theorem, there exists a unique fixed-point x and the fixed-point iteration converges for any starting value x^0:

$$\|e^t\|_{B,\varepsilon} = \|B^t e^0\|_{B,\varepsilon} \leq \|B^t\|_{B,\varepsilon}\, \|e^0\|_{B,\varepsilon} \leq \|B\|^t_{B,\varepsilon}\, \|e^0\|_{B,\varepsilon} \;\to\; 0.$$

In view of the norm equivalence in $\mathbb{R}^n$ this means convergence in any norm, $x^t \to x\,(t \to \infty)$.

(ii) Now, we assume convergence for any starting value x^0. Let λ be an eigenvalue of B such that $|\lambda| = \rho$ and $w \neq 0$ a corresponding eigenvector. Then, for the particular starting value $x^0 := x + w$, we obtain

$$\lambda^t e^0 = \lambda^t w = B^t w = B^t e^0 = e^t \to 0 \quad (t \to \infty).$$

This necessarily requires $\operatorname{spr}(B) = |\lambda| < 1$. As byproduct of this argument, we see that in this particular case

$$\Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} = \rho, \quad t \in \mathbb{N}.$$

(iii) For an arbitrary small $\varepsilon > 0$ let $\|\cdot\|_{B,\varepsilon}$ again be the above special norm for which $\|B\|_{B,\varepsilon} \leq \rho + \varepsilon$ holds. Then, by the norm equivalence for any other vector norm $\|\cdot\|$ there exist positive numbers $m = m(B,\varepsilon)$, $M = M(B,\varepsilon)$ such that

$$m\|x\| \leq \|x\|_{B,\varepsilon} \leq M\|x\|, \quad x \in \mathbb{R}^n.$$

Using this notation, we obtain

$$\|e^t\| \leq \frac{1}{m}\|e^t\|_{B,\varepsilon} = \frac{1}{m}\|B^t e^0\|_{B,\varepsilon} \leq \frac{1}{m}\|B\|^t_{B,\varepsilon}\|e^0\|_{B,\varepsilon} \leq \frac{M}{m}(\rho+\varepsilon)^t\|e^0\|,$$

and, consequently, observing that $\big(\frac{M}{m}\big)^{1/t} \to 1\ (t \to \infty)$:

$$\limsup_{t\to\infty}\ \Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} \leq \rho + \varepsilon.$$

Since $\varepsilon > 0$ can be chosen arbitrarily small and recalling the last identity in (ii), we obtain the asserted identity (3.1.8).

(iv) Finally requiring an error reduction by $TOL > 0$, we have to set

$$\frac{\|x^t - x\|}{\|x^0 - x\|} \leq (\rho+\varepsilon)^t \approx TOL, \quad t \geq t_\varepsilon,$$

from which we obtain

$$t(\varepsilon) \approx \frac{\ln(TOL)}{\ln(\rho)}.$$

This completes the proof. Q.E.D.

The spectral radius of the iteration matrix determines the general asymptotic convergence behavior of the fixed-point iteration. The relation (3.1.9) can be interpreted as follows: In case that $\rho = \operatorname{spr}(B) < 1$ the error obtained in the tth step (t sufficiently large) can be further

reduced by a factor 10^{-1}, i. e., gaining one additional decimal in accuracy, by

$$t_1 = -\frac{\ln(10)}{\ln(\rho)}$$

more iterations. For example, for $\rho \sim 0.99$, which is not at all unrealistic, we have $t_1 \sim 230$. For large systems with $n \gg 10^6$ this means substantial work even if each iteration step only requires $\mathcal{O}(n)$ arithmetic operations.

We have to provide the auxiliary lemma used in the proof of Theorem 3.1.

Lemma 3.1 (spectral radius): *For any matrix $B \in \mathbb{R}^{n\times n}$ and any small $\varepsilon > 0$ there exists a natural matrix norm $\|\cdot\|_{B,\varepsilon}$, such that*

$$\mathrm{spr}(B) \le \|B\|_{B,\varepsilon} \le \mathrm{spr}(B) + \varepsilon. \tag{3.1.10}$$

Proof. The matrix B is similar to an upper triangular matrix (e. g., its Jordan normal form),

$$B = T^{-1}RT\,, \quad R = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix},$$

with the eigenvalues of B on its main diagonal. Hence,

$$\mathrm{spr}\,(B) = \max_{1\le i\le n} |r_{ii}|.$$

For an arbitrary $\delta \in (0, 1]$, we set

$$S_\delta = \begin{bmatrix} 1 & & & 0 \\ & \delta & & \\ & & \ddots & \\ 0 & & & \delta^{n-1} \end{bmatrix}, \quad R_0 = \begin{bmatrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{bmatrix}, \quad Q_\delta = \begin{bmatrix} 0 & r_{12} & \delta r_{13} & \cdots & \delta^{n-2} r_{1n} \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & \delta r_{n-2,n} \\ & & & \ddots & r_{n-1,n} \\ & & & & 0 \end{bmatrix},$$

and, with this notation, have

$$R_\delta := S_\delta^{-1} R S_\delta = \begin{bmatrix} r_{11} & \delta r_{12} & \cdots & \delta^{n-1} r_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \delta r_{n-1,n} \\ 0 & & & r_{nn} \end{bmatrix} = R_0 + \delta Q_\delta.$$

In view of the regularity of $S_\delta^{-1}T$, a vector norm is defined by

$$\|x\|_\delta := \|S_\delta^{-1} T x\|_2\,, \quad x \in \mathbb{R}^n.$$

Then, observing $R = S_\delta R_\delta S_\delta^{-1}$, there holds

$$B = T^{-1}RT = T^{-1}S_\delta R_\delta S_\delta^{-1}T.$$

Hence for all $x \in \mathbb{R}^n$ and $y = S_\delta^{-1}Tx$:

$$\begin{aligned} \|Bx\|_\delta &= \|T^{-1}S_\delta R_\delta S_\delta^{-1}Tx\|_\delta = \|R_\delta y\|_2 \\ &\le \|R_0 y\|_2 + \delta\|Q_\delta y\|_2 \le \{\max_{1\le i\le n}|r_{ii}| + \delta\mu\}\,\|y\|_2 \\ &\le \{\mathrm{spr}\,(B) + \delta\mu\}\|x\|_\delta \end{aligned}$$

with the constant

$$\mu = \Big(\sum_{i,j=1}^{n} |r_{ij}|^2\Big)^{1/2}.$$

This implies

$$\|B\|_\delta = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Bx\|_\delta}{\|x\|_\delta} \le \mathrm{spr}\,(B) + \mu\delta,$$

and setting $\delta := \varepsilon/\mu$ the desired vector norm is given by $\|\cdot\|_{B,\varepsilon} := \|\cdot\|_\delta$. Q.E.D.

3.1.1 Stopping criteria

In using an iterative method, one needs "stopping criteria", which for some prescribed accuracy TOL terminates the iteration, in the ideal case, once this required accuracy is reached.

(i) **Strategy 1.** From the Banach fixed-point theorem, we have the general error estimate

$$\|x^t - x\| \le \frac{q}{1-q}\,\|x^t - x^{t-1}\|, \tag{3.1.11}$$

with the "contraction constant" $q = \|B\| < 1$. For a given error tolerance $TOL > 0$ the iteration could be stopped when

$$\frac{\|B\|}{1-\|B\|}\frac{\|x^t - x^{t-1}\|}{\|x^t\|} \le TOL. \tag{3.1.12}$$

The realization of this strategy requires an quantitatively correct estimate of the norm $\|B\|$ or of $\mathrm{spr}\,(B)$. That has to be generated from the computed iterates x^t, i. e., *a posteriori* in the course of the computation. In general the iteration matrix $B = I - C^{-1}A$ cannot be computed explicitly with acceptable work. Methods for estimating $\mathrm{spr}\,(B)$ will be considered in the chapter about the iterative solution of eigenvalue problems, below.

(ii) **Strategy 2.** Alternatively, one can evaluate the "residual" $\|Ax^t - b\|$. Observing that $e^t = x^t - x = A^{-1}(Ax^t - b)$ and $x = A^{-1}b$, it follows that

$$\|e^t\| \le \|A^{-1}\|\,\|Ax^t - b\|, \quad \frac{1}{\|b\|} \ge \frac{1}{\|A\|\,\|x\|},$$

and further

$$\frac{\|e^t\|}{\|x\|} \le \|A^{-1}\|\|A\|\frac{\|Ax^t - b\|}{\|b\|} = \mathrm{cond}(A)\,\frac{\|Ax^t - b\|}{\|b\|}.$$

This leads us to the stopping criterion

$$\mathrm{cond}(A)\,\frac{\|Ax^t - b\|}{\|b\|} \leq TOL. \tag{3.1.13}$$

The evaluation of this criterion requires an estimate of $\mathrm{cond}(A)$, which may be as costly as the solution of the equation $Ax = b$ itself. Using the spectral norm $\|\cdot\|_2$ the condition number is related to the singular values of A, i. e., the square roots of the eigenvalues of $A^T A$,

$$\mathrm{cond}_2(A) = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

Again generating accurate estimates of these eigenvalues may require more work than the solution of $Ax = b$.

This short discussion shows that designing useful stopping criteria for iterative methods is an not at all an easy task. However, in the context of linear systems originating from the "finite element discretization" ("FEM") of partial differential equations there are approaches based on the concept of "Galerkin orthogonality", which allow for a systematic balancing of iteration and discretization errors. In this way, practical stopping criteria can be designed, by which the iteration may be terminated once the level of the discretization error is reached. Here, the criterion is essentially the approximate solution's "violation of Galerkin orthogonality" (s. Meidner et al. [40] and Rannacher et al. [42] for more details).

3.1.2 Construction of iterative methods

The construction of concrete iterative methods for solving the linear system $Ax = b$ by defect correction requires the specification of the preconditioner C. For this task two particular goals have to be observed:

- $\mathrm{spr}\,(I - C^{-1}A)$ should be as small as possible.
- The correction equation $C\delta x^t = b - Ax^{t-1}$ should be solvable with $\mathcal{O}(n)$ a. op., requiring storage space not much exceeding that for storing the matrix A.

Unfortunately, these requirements contradict each other. The two extreme cases are:

$$\begin{aligned} C = A \quad &\Rightarrow \quad \mathrm{spr}\,(I - C^{-1}A) = 0 \\ C = \theta^{-1} I \quad &\Rightarrow \quad \mathrm{spr}\,(I - C^{-1}A) \sim 1. \end{aligned}$$

The simplest preconditioners are defined using the following natural additive decomposition of the matrix $A = L + D + R$:

$$D = \begin{bmatrix} a_{11} & & \cdots & 0 \\ & \ddots & & \\ & & \ddots & \\ 0 & \cdots & & a_{nn} \end{bmatrix} \quad L = \begin{bmatrix} 0 & & \cdots & 0 \\ a_{21} & \ddots & & \\ \vdots & \ddots & \ddots & \\ a_{n1} & \cdots & a_{n,n-1} & 0 \end{bmatrix} \quad R = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & a_{n-1,n} \\ 0 & \cdots & & 0 \end{bmatrix}.$$

Further, we assume that the main diagonal elements of A are nonzero, $a_{ii} \neq 0$.

1. *Jacobi[2] method ("Gesamtschrittverfahren" in German):*

$$C = D, \qquad B = -D^{-1}(L+R) =: J \quad (\text{"Jacobi iteration matrix"}). \tag{3.1.14}$$

The iteration of the Jacobi method reads

$$Dx^t = b - (L+R)x^{t-1}, \quad t = 1, 2, \dots , \tag{3.1.15}$$

or written component-wise:

$$a_{ii}x_i^t = b_i - \sum_{j=1}^{n} a_{ij}x_j^t, \quad i = 1, \dots, n,$$

2. *Gauß-Seidel method ("Einzelschrittverfahren" in German):*

$$C = D + L, \qquad B = -(D+L)^{-1}R =: H_1 \quad (\text{Gauß-Seidel iteration matrix}). \tag{3.1.16}$$

The iteration of the Gauß-Seidel method reads as follow:

$$(D+L)x^t = b - Rx^{t-1}, \quad t = 1, 2, \dots .$$

Writing this iteration componentwise,

$$a_{ii}x_i^t = b_i - \sum_{j<i} a_{ij}x_j^t - \sum_{j>i} a_{ij}x_j^{t-1}, \quad i = 1, \dots, n,$$

one sees that Jacobi and Gauß-Seidel method have exactly the same arithmetical complexity per iteration step and require the same amount of storage. However, since the latter method uses a better approximation of the matrix A as preconditioner it is expected to have an iteration matrix with smaller spectral radius, i. e., converges faster. It will be shown below that this is actually the case for certain classes of matrices A.

3. *SOR method ("Successive Over-Relaxation"):* $\omega \in (0, 2)$

$$C = \frac{1}{\omega}(D + \omega L), \qquad B = -(D+\omega L)^{-1}[(\omega - 1)D + \omega R]. \tag{3.1.17}$$

The SOR method is designed to accelerate the Gauß-Seidel method by introducing a "relaxation parameter" $\omega \in \mathbb{R}$, which can be optimized in order to minimize the spectral radius of the corresponding iteration matrix. Its iteration reads as follows:

$$(D+\omega L)x^t = \omega b - [(\omega - 1)D + \omega R]x^{t-1}, \quad t = 1, 2, \dots .$$

The arithmetical complexity is about that of Jacobi and Gauß-Seidel method. But the parameter ω can be optimized for a certain class of matrices resulting in a significantly faster convergence than that of the other two simple methods.

[2] Carl Gustav Jakob Jacobi (1804-1851): German mathematician; already as child highly gifted; worked in Königsberg and Berlin; contributions to many parts of mathematics: number theory, elliptic functions, partial differential equations, functional determinants, and theoretical mechanics.

4. *ILU method ("Incomplete LU Decomposition"):*

$$C = \tilde{L}\tilde{R}, \qquad B = I - \tilde{R}^{-1}\tilde{L}^{-1}A. \tag{3.1.18}$$

For a symmetric, positive definite matrix A the ILU method naturally becomes the ILL^T method ("Incomplete Cholesky[3] decomposition"). The ILU decomposition is obtained by the usual recursive process for the direct computation of the LU decomposition from the relation $LU = A$ by setting all matrix elements to zero, which correspond to index pairs $\{i, j\}$ for which $a_{ij} = 0$:

$$\begin{aligned} i = 1, ..., n: \quad & \tilde{r}_{il} = a_{il} - \sum_{k=1}^{i-1} \tilde{l}_{ik}\tilde{r}_{kl} \quad (l = 1, ..., n) \\ & \tilde{l}_{ii} = 1, \quad \tilde{l}_{ki} = \tilde{r}_{ii}^{-1}\Big\{a_{ki} - \sum_{l=1}^{i-1} \tilde{l}_{kl}\tilde{r}_{li}\Big\} \quad (k = i+1, ..., n) \\ & \tilde{l}_{ij} = 0, \ \tilde{r}_{ij} = 0, \ \text{for } a_{ij} = 0\,. \end{aligned}$$

If this process stops because some $\tilde{r}_{ii} = 0$, we set $\tilde{r}_{ii} := \delta > 0$ and continue. The iteration of the ILU method reads as follows:

$$\tilde{L}\tilde{R}x^t = (\tilde{L}\tilde{R} - A)x^{t-1} + b, \quad t = 1, 2, \dots\,.$$

Again this preconditioner is cheap, for sparse matrices $\mathcal{O}(n)$ a. op. per iteration step, but its convergence is difficult to analyze and will not be discussed further. However, in certain situations the ILU method plays an important role as a robust "smoothing iteration" within "multigrid methods" to be discussed below.

5. *ADI method ("Alternating-Direction Implicit Iteration"):*

$$\begin{aligned} C &= (A_x + \omega I)(A_y + \omega I), \\ B &= (A_y + \omega I)^{-1}(\omega I - A_x)(A_x + \omega I)^{-1}(\omega I - A_y). \end{aligned} \tag{3.1.19}$$

The ADI method can be applied to matrices A which originate from the discretization of certain elliptic partial differential equations, in which the contributions from the different spatial directions (x-direction and y-direction in 2D) are separated in the form $A = A_x + A_y$. A typical example is the central difference approximation of the Poisson equation described in Chapter 0.4.2. The iteration of the ADI method reads as follows

$$(A_x + \omega I)(A_y + \omega I)x^t = \big((A_x + \omega I)(A_y + \omega I) - A\big)x^{t-1} + b, \quad t = 1, 2, \dots\,.$$

Here, the matrices $A_x + \omega I$ and $A_y + \omega I$ are tri-diagonal, such that the second goal "solution efficiency" is achieved, while the full matrix A is five-diagonal. This method can be shown to converge for any choice of the parameter $\omega > 0$. For certain classes of matrices the optimal choice of ω leads to convergence, which is at least as fast as that of the optimal SOR method. We will not discuss this issue further since the range of applicability of the ADI method is rather limited.

[3] Andrè Louis Cholesky (1975-1918): French mathematician; military career; contribution to numerical linear algebra.

Remark 3.1 (block-versions of fixed-point iterations): Sometimes the coefficient matrix A has a regular block structure for special numberings of the unknowns (e. g., in the discretization of the Navier-Stokes equations when grouping the velocity and pressure unknowns together at each mesh point):

$$A = \begin{bmatrix} A_{11} & & \cdots & A_{1r} \\ & \ddots & & \\ \vdots & & \ddots & \vdots \\ A_{r1} & \cdots & & A_{rr} \end{bmatrix},$$

where the submatrices A_{ij} are of small dimension, $3-10$, such that the explicit inversion of the diagonal blocks A_{ii} is possible without spoiling the overall complexity of $\mathcal{O}(n)$ a. op. per iteration step.

3.1.3 Jacobi- and Gauß-Seidel methods

In the following, we will give a complete convergence analysis of Jacobi and Gauß-Seidel method. As already stated above, both methods have the same arithmetical cost (per iteration step) and require not much more storage as needed for storing the matrix A. This simplicity suggests that both methods may not be very fast, which will actually be seen below at the model matrix in Example (2.7) of Section 2.2.

Theorem 3.2 (strong row-sum criterion): *If the row sums or the column sums of the matrix $A \in \mathbb{R}^{n\times n}$ satisfy the condition (strict diagonal dominance)*

$$\sum_{k=1,k\neq j}^{n} |a_{jk}| < |a_{jj}| \quad or \quad \sum_{k=1,k\neq j}^{n} |a_{kj}| < |a_{jj}|, \quad j = 1,\ldots,n, \tag{3.1.20}$$

then, $\mathrm{spr}(J) < 1$ *and* $\mathrm{spr}(H_1) < 1$, *i. e., Jacobi and Gauß-Seidel method converge.*

Proof. First, assume that the matrix A is strictly diagonally dominant. Let $\lambda \in \sigma(J)$ and $\mu \in \sigma(H_1)$ with corresponding eigenvectors v and w, respectively. Then, noting that $a_{jj} \neq 0$, we have

$$\lambda v = Jv = -D^{-1}(L+R)v$$

and

$$\mu w = H_1 w = -(D+L)^{-1}Rw \quad \Leftrightarrow \quad \mu w = -D^{-1}(\mu L+R)w.$$

From this it follows that for $\|v\|_\infty = \|w\|_\infty = 1$ and using the strict diagonal dominance of A:

$$|\lambda| \leq \|D^{-1}(L+R)\|_\infty = \max_{j=1,\ldots,n} \Big\{ \frac{1}{|a_{jj}|} \sum_{k=1,k\neq j}^{n} |a_{jk}| \Big\} < 1.$$

Hence, $\mathrm{spr}\,(J) < 1$. Further,

$$|\mu| \leq \|D^{-1}(\mu L+R)\|_\infty \leq \max_{1\leq j\leq n} \Big\{ \frac{1}{|a_{jj}|} \Big[\sum_{k<j} |\mu|\,|a_{jk}| + \sum_{k>j} |a_{jk}| \Big] \Big\}.$$

For $|\mu| \geq 1$, we would obtain the contradiction

$$|\mu| \leq |\mu| \, \|D^{-1}(L+R)\|_\infty < |\mu|,$$

so that also $\mathrm{spr}\,(H_1) < 1$. If instead of A its transpose A^T is strictly diagonally dominant, we can argue analogously since, in view of $\lambda(\bar{A}^T) = \overline{\lambda(A)}$, the spectral radii of these two matrices coincide. Q.E.D.

Remark 3.2: We show an example of a non-symmetric matrix A, which satisfies the strong column- but not the strong row-sum criterion:

$$A = \begin{bmatrix} 4 & 4 & 1 \\ 2 & 5 & 3 \\ 1 & 0 & 5 \end{bmatrix}, \qquad A^T = \begin{bmatrix} 4 & 2 & 1 \\ 4 & 5 & 0 \\ 1 & 3 & 5 \end{bmatrix}.$$

Clearly, for symmetric matrices the two conditions are equivalent.

The strict diagonal dominance of A or A^T required in Theorem 3.2 is a too restrictive condition for the needs of many applications. In most cases only simple "diagonal dominance" is given as in the Example (2.7) of Section 2.2,

$$A = \left.\begin{bmatrix} B & -I_4 & & \\ -I_4 & B & -I_4 & \\ & -I_4 & B & -I_4 \\ & & -I_4 & B \end{bmatrix}\right\} 16, \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{bmatrix}\right\} 4$$

However, this matrix is *strictly* diagonally dominant in some of its rows, which together with an additional structural property of A can be used to guarantee convergence of Jacobi and Gauß-Seidel method.

Definition 3.1: *A matrix $A \in \mathbb{R}^{n\times n}$ is called "reducible", if there exists a permutation matrix P such that*

$$PAP^T = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix},$$

(simultaneous row and column permutation) with matrices $\tilde{A}_{11} \in \mathbb{R}^{p\times p}$, $\tilde{A}_{22} \in \mathbb{R}^{q\times q}$, $\tilde{A}_{21} \in \mathbb{R}^{q\times p}$, $p, q > 0$, $p + q = n$. It is called "irreducible" if it is not reducible.

For a reducible matrix A the linear system $Ax = b$ can be transformed into an equivalent system of the form $PAP^T y = Pb$, $x = P^T y$ which is decoupled into two separate parts such that it could be solved in two successive steps. The following lemma provides a criterion for the irreducibility of the matrix A, which can be used in concrete cases. For example, the above model matrix A is irreducible.

Lemma 3.2 (irreducibility): *A matrix $A \in \mathbb{R}^{n\times n}$ is irreducible if and only if the associated directed graph*

$$G(A) := \left\{ \textit{knots } P_1, ..., P_n \,,\ \textit{edges } \overline{P_j P_k} \Leftrightarrow a_{jk} \neq 0,\ j, k = 1, ..., n \right\}$$

is connected, i. e., for each pair of knots $\{P_j, P_k\}$ *there exists a directed connection between* P_j *and* P_k *.*

Proof. The reducibility of A can be formulated as follows: There exists a non-trivial decomposition $N_n = J \cup K$ of the index set $N_n = \{1, ..., n\}$, $J, K \neq \emptyset$, $J \cap K = \emptyset$ such that $a_{jk} = 0$ for all pairs $\{j, k\} \in J \times K$. Connectivity of the graph $G(A)$ now means that for any pair of indices $\{j, k\}$ there exists a chain of indices $i_1, \ldots, i_m \in \{1, \ldots, n\}$ such that

$$a_{ji_1} \neq 0\,,\ a_{i_1 i_2} \neq 0\,, \ldots\,, a_{i_{m-1} i_m} \neq 0\,,\ a_{i_m k} \neq 0.$$

From this, we conclude the asserted characterization (exercise). Q.E.D.

For irreducible matrices the condition in the strong row-sum criterion can be relaxed.

Theorem 3.3 (weak row-sum criterion): *Let the matrix* $A \in \mathbb{R}^{n\times n}$ *be irreducible and diagonally dominant,*

$$\sum_{k=1, k\neq j}^{n} |a_{jk}| \leq |a_{jj}| \quad j = 1, \ldots, n, \tag{3.1.21}$$

and let for at least one index $r \in \{1, \ldots, n\}$ *the corresponding row sum satisfy*

$$\sum_{k=1, k\neq r}^{n} |a_{rk}| < |a_{rr}|. \tag{3.1.22}$$

Then, A *is regular and* $\mathrm{spr}(J) < 1$ *and* $\mathrm{spr}(H_1) < 1$*, i. e., Jacobi and Gauß-Seidel method converge. An analogous criterion holds in terms of the column sums of* A*.*

Proof. (i) Because of the assumed irreducibility of the matrix A there necessarily holds

$$\sum_{k=1}^{n} |a_{jk}| \ > \ 0\,, \quad j = 1, \ldots, n\,,$$

and, consequently, by its diagonal dominance, $a_{jj} \neq 0$, $j = 1, \ldots, n$. Hence, Jacobi and Gauß-Seidel method are feasible. With the aid of the diagonal dominance, we conclude analogously as in the proof of Theorem 3.2 that

$$\mathrm{spr}\,(J) \leq 1\,, \quad \mathrm{spr}\,(H_1) \leq 1.$$

(ii) Suppose now that there is an eigenvalue $\lambda \in \sigma(J)$ with modulus $|\lambda| = 1$. Let $v \in \mathbb{C}^n$ be a corresponding eigenvector with a component v_s satisfying $|v_s| = \|v\|_\infty = 1$. There holds

$$|\lambda|\,|v_i| \leq |a_{ii}|^{-1} \sum_{k\neq i} |a_{ik}|\,|v_k|\,, \quad i = 1, \ldots, n. \tag{3.1.23}$$

By the assumed irreducibility of A in the sense of Lemma 3.2 there exist a chain of indices $i_1, \ldots, i_m$ such that $a_{si_1} \neq 0\,, \ldots, a_{i_m r} \neq 0$. Hence, by multiple use of the inequality (3.1.23),

we obtain the following contradiction (observe that $|\lambda| = 1$)

$$
\begin{aligned}
|v_r| = |\lambda v_r| &\le |a_{rr}|^{-1} \sum_{k \neq r} |a_{rk}| \, \|v\|_\infty < \|v\|_\infty, \\
|v_{i_m}| = |\lambda v_{i_m}| &\le |a_{i_m i_m}|^{-1} \Big\{ \sum_{k \neq i_m, r} |a_{i_m k}| \, \|v\|_\infty + \underbrace{|a_{i_m r}|}_{\neq 0} \, |v_r| \Big\} < \|v\|_\infty, \\
&\vdots \\
|v_{i_1}| = |\lambda v_{i_1}| &\le |a_{i_1 i_1}|^{-1} \Big\{ \sum_{k \neq i_1, i_2} |a_{i_1 k}| \, \|v\|_\infty + \underbrace{|a_{i_1 i_2}|}_{\neq 0} \, |v_{i_2}| \Big\} < \|v\|_\infty, \\
\|v\|_\infty = |\lambda v_s| &\le |a_{ss}|^{-1} \Big\{ \sum_{k \neq s, i_1} |a_{sk}| \, \|v\|_\infty + \underbrace{|a_{s i_1}|}_{\neq 0} \, |v_{i_1}| \Big\} < \|v\|_\infty.
\end{aligned}
$$

Consequently, there must hold $\operatorname{spr}(J) < 1$. Analogously, we also conclude $\operatorname{spr}(H_1) < 1$. In view of $A = D(I - J)$ the matrix A must be regular. Q.E.D.

3.2 Acceleration methods

For practical problems Jacobi and Gauß-Seidel method are usually much too slow. Therefore, one tries to improve their convergence by several tricks, two of which will be discussed below.

3.2.1 SOR method

The SOR method can be interpreted as combining the Gauß-Seidel method with an extra "relaxation step". Starting from a standard Gauß-Seidel step in the t-th iteration,

$$
\tilde{x}_j^t = \frac{1}{a_{jj}} \Big\{ b_j - \sum_{k<j} x_k^t - \sum_{k>j} x_k^{t-1} \Big\}
$$

the next iterate x_j^t is generated as a convex linear combination ("relaxation") of the form

$$
x_j^t = \omega \tilde{x}_j^t + (1 - \omega) \, x_j^{t-1},
$$

with a parameter $\omega \in (0, 2)$. For $\omega = 1$ this is just the Gauß-Seidel iteration. For $\omega < 1$, one speaks of "underrelaxation" and for $\omega > 1$ of "overrelaxation". The iteration matrix of the SOR methods is obtained from the relation

$$
x^t = \omega D^{-1} \{ b - L x^t - R x^{t-1} \} + (1 - \omega) x^{t-1}
$$

as

$$
H_\omega = -(D + \omega L)^{-1} \, [(\omega - 1) \, D + \omega R].
$$

Hence, the iteration reads

$$
x^t = H_\omega x^{t-1} + \omega (D + \omega L)^{-1} b, \tag{3.2.24}
$$

or in componentwise notation:

$$x_i^t = (1-\omega)x_i^{t-1} + \frac{\omega}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij}x_j^t - \sum_{j>i} a_{ij}x_j^{t-1}\Big), \quad i = 1,\ldots,n. \tag{3.2.25}$$

The following lemma shows that in the relaxation parameter has to be picked as $0 < \omega < 2$ if one wants to guarantee convergence.

Lemma 3.3 (relaxation): *For an arbitrary matrix $A \in \mathbb{R}^{n\times n}$ with regular D there holds*

$$\operatorname{spr}(H_\omega) \geq |\omega - 1|, \quad \omega \in \mathbb{R}. \tag{3.2.26}$$

Proof. We have

$$H_\omega = (D+\omega L)^{-1}[\,(1-\omega)\,D - \omega R\,] = (I+\omega \underbrace{D^{-1}L}_{=:\,L'})^{-1}\,\underbrace{D^{-1}D}_{=\,I}\,[\,(1-\omega)\,I - \omega \underbrace{D^{-1}R}_{=:\,R'}].$$

Then,

$$\det(H_\omega) = \underbrace{\det(I+\omega L')}_{=\,1}{}^{-1} \cdot \underbrace{\det((1-\omega)\,I - \omega R')}_{=\,(1-\omega)^n} = (1-\omega)^n.$$

Since $\det(H_\omega) = \prod_{i=1}^n \lambda_i$ ($\lambda_i \in H_\omega$) it follows that

$$\operatorname{spr}(H_\omega) = \max_{1\leq i\leq n} |\lambda_i| \geq \Big(\prod_{i=1}^n |\lambda_i|\Big)^{1/n} = |1-\omega|,$$

which proves the asserted estimate. Q.E.D.

For positive definite matrices the assertion of Lemma 3.3 can be reversed in a certain sense.

Theorem 3.4 (Theorem of Ostrowski-Reich): *For a positive definite matrix $A \in \mathbb{R}^{n\times n}$ there holds*

$$\operatorname{spr}(H_\omega) < 1, \quad \textit{for} \quad 0 < \omega < 2. \tag{3.2.27}$$

Hence, especially the Gauß-Seidel method ($\omega = 1$) is convergent. Its asymptotic convergence speed can be estimated by

$$\operatorname{spr}(H_1) \leq 1 - \frac{2}{\mu} + \frac{2}{\mu(\mu+1)}, \qquad \mu := \frac{\lambda_{\max}(D)}{\lambda_{\min}(A)}, \tag{3.2.28}$$

assuming the quantity $\mu \approx \operatorname{cond}_2(A)$ to be large.

Proof. (i) In view of the symmetry of A, we have $R = L^T$, i. e., $A = L + D + L^T$. Let $\lambda \in \sigma(H_\omega)$ be arbitrary for $0 < \omega < 2$, with some eigenvector $v \in \mathbb{R}^n$, i. e., $H_\omega v = \lambda v$. Thus, there holds

$$\big((1-\omega)\,D - \omega L^T\big)\,v = \lambda\,(D+\omega L)\,v$$

and

$$\omega\,(D+L^T)\,v = (1-\lambda)\,Dv - \lambda\omega Lv.$$

From this, we conclude that

$$\begin{aligned}\omega Av &= \omega\,(D+L^T)\,v + \omega Lv \\ &= (1-\lambda)\,Dv - \lambda\omega Lv + \omega Lv = (1-\lambda)\,Dv + \omega\,(1-\lambda)\,Lv,\end{aligned}$$

and

$$\begin{aligned}\lambda\omega Av &= \lambda\omega\,(D+L^T)\,v + \lambda\omega Lv \\ &= \lambda\omega\,(D+L^T)\,v + (1-\lambda)\,Dv - \omega\,(D+L^T)\,v \\ &= (\lambda-1)\omega(D+L^T)\,v + (1-\lambda)\,Dv = (1-\lambda)(1-\omega)\,Dv - (1-\lambda)\,\omega L^T v.\end{aligned}$$

Observing $v^T Lv = v^T L^T v$ implies

$$\begin{aligned}\omega v^T Av &= (1-\lambda)\,v^T Dv + \omega\,(1-\lambda)\,v^T Lv \\ \lambda\omega v^T Av &= (1-\lambda)(1-\omega)\,v^T Dv - (1-\lambda)\,\omega v^T Lv,\end{aligned}$$

and further by adding the two equations,

$$\omega\,(1+\lambda)\,v^T Av = (1-\lambda)\,(2-\omega)\,v^T Dv.$$

As with A also D is positive definite there holds $v^T Av > 0$, $v^T Dv > 0$. Consequently (observing $0 < \omega < 2$), $\lambda \neq \pm 1$, and it follows that

$$\mu := \frac{1+\lambda}{1-\lambda} = \frac{2-\omega}{\omega}\,\frac{v^T Dv}{v^T Av} > 0.$$

Resolving this for λ, we finally obtain qualitative estimate

$$|\lambda| = \left|\frac{\mu-1}{\mu+1}\right| < 1, \tag{3.2.29}$$

what was to be shown.

(ii) To derive the quantitative estimate (3.2.28), we rewrite (3.2.29) in the form

$$|\lambda| = \left|\frac{\mu-1}{\mu+1}\right| = \left|\frac{1-1/\mu}{1+1/\mu}\right| \leq 1 - \frac{2}{\mu} + \frac{2}{\mu(\mu+1)}\,,$$

where

$$\mu = \frac{v^T Dv}{v^T Av} \leq \frac{\max_{\|y\|_2=1} y^T Dy}{\min_{\|y\|_2=1} y^T Ay} \leq \frac{\lambda_{\max}(D)}{\lambda_{\min}(A)}.$$

This completes the proof. Q.E.D.

Remark 3.3: The estimate (3.2.28) for the convergence rate of the Gauß-Seidel method in the case of a symmetric, positive definite matrix A has an analogue for the Jacobi method,

$$\mathrm{spr}(J) \leq 1 - \frac{1}{\mu}, \tag{3.2.30}$$

where μ is defined as in (3.2.28). This is easily seen by considering any eigenvalue $\lambda \in \sigma(J)$

with corresponding normalized eigenvector v, $\|v\|_2 = 1$, satisfying

$$\lambda Dv = Dv - Av.$$

Multiplying by v and observing that A as well as D are positive definite, then yields

$$\lambda = 1 - \frac{v^T Av}{v^T Dv} \leq 1 - \frac{1}{\mu}.$$

Comparing this estimate with (3.2.28) and observing that $\mathrm{spr}(J)^2 = (1-\mu^{-1})^2 \approx 1 - 2\mu^{-1} \approx \mathrm{spr}(H_1)$, for $\mu \gg 1$, indicates that the Gauß-Seidel method may be almost twice as fast than the Jacobi method. That this is actually the case will be seen below for a certain class of matrices.

Definition 3.2: *A matrix $A \in \mathbb{R}^{n\times n}$ with the additive splitting $A = L + D + R$ is called "consistently ordered" if the eigenvalues of the matrices*

$$J(\alpha) = -D^{-1}\{\alpha L + \alpha^{-1} R\}, \quad \alpha \in \mathbb{C},$$

are independent of the parameter α, i. e., equal to the eigenvalues of the matrix $J = J(1)$.

The importance of this property lies in the fact that in this case there are explicit relations between the eigenvalues of J and those of H_ω.

Example 3.2: Though the condition of "consistent ordering" appears rather strange and restrictive, it is satisfied for a large class of matrices. Consider the model matrix in Subsection 0.4.2 of Chapter 0. Depending on the numbering of the mesh points matrices with different block structures are encountered.

(i) If the mesh points are numbered in a checker-board manner a block-tridiagonal matrix

$$A = \begin{bmatrix} D_1 & A_{12} & & \\ A_{21} & D_2 & \ddots & \\ & \ddots & \ddots & A_{r-1,r} \\ & & A_{r,r-1} & D_r \end{bmatrix},$$

occurs where the D_i are diagonal and regular. Such a matrix is consistently ordered, which is seen by applying a suitable similarity transformation,

$$T = \begin{bmatrix} I & & & \\ & \alpha I & & \\ & & \ddots & \\ & & & \alpha^{r-1} I \end{bmatrix}, \qquad \alpha D^{-1}L + z^{-1}D^{-1}R \;=\; T(D^{-1}L + D^{-1}R)T^{-1}.$$

and observing that similar matrices have the same eigenvalues.

(ii) If the mesh points are numbered in a row-wise manner a block-tridiagonal matrix

$$A = \begin{bmatrix} A_1 & D_{12} & & \\ D_{21} & A_2 & \ddots & \\ & \ddots & \ddots & D_{r-1,r} \\ & & D_{r,r-1} & A_r \end{bmatrix},$$

occurs where the A_i are tridiagonal and the D_{ij} diagonal. Such a matrix is consistently ordered, which is seen by first applying the same similarity transformation as above,

$$TAT^{-1} = \begin{bmatrix} A_1 & \alpha^{-1}D_{12} & & \\ \alpha D_{21} & A_2 & \ddots & \\ & \ddots & \ddots & \alpha^{-1}D_{r-1,r} \\ & & \alpha D_{r,r-1} & A_r \end{bmatrix},$$

and then a similarity transformation with the diagonal-block matrix $S = \mathrm{diag}\{S_1, \ldots, S_m\}$, where $S_i = \mathrm{diag}\{1, \alpha, \alpha^2, \ldots, \alpha^{r-1}\}$, $i = 1, \ldots, m$,

$$STAT^{-1}S^{-1} = \begin{bmatrix} S_1A_1S_1^{-1} & \alpha^{-1}D_{12} & & \\ \alpha D_{21} & S_2A_2S_2^{-1} & \ddots & \\ & \ddots & \ddots & \alpha^{-1}D_{r-1,r} \\ & & \alpha D_{r,r-1} & S_rA_rS_r^{-1} \end{bmatrix}.$$

Here, it has been used that the blocks D_{ij} are diagonal. Since the main-diagonal blocks are tri-diagonal, they split like $A_i = D_i + L_i + R_i$ and there holds $S_iA_iS_i^{-1} = D_i + \alpha L + \alpha^{-1}R$. This implies that the matrix A is consistently ordered.

Theorem 3.5 (optimal SOR method): *Let the matrix $A \in \mathbb{R}^{n\times n}$ be consistently ordered and $0 \le \omega \le 2$. Then, the eigenvalues $\mu \in \sigma(J)$ and $\lambda \in \sigma(H_\omega)$ are related by*

$$\lambda^{1/2}\omega\mu = \lambda + \omega - 1. \tag{3.2.31}$$

Proof. Let $\lambda, \mu \in \mathbb{C}$ two numbers, which satisfy equation (3.2.31). If $0 \neq \lambda \in \sigma(H_\omega)$ the relation $H_\omega v = \lambda v$ is equivalent to

$$\left((1-\omega)I - \omega D^{-1}R\right) v = \lambda(I + \omega D^{-1}L)v$$

and

$$(\lambda + \omega - 1)v = -\lambda^{1/2}\omega\left(\lambda^{1/2}D^{-1}L + \lambda^{-1/2}D^{-1}R\right) v = \lambda^{1/2}\omega J(\lambda^{1/2})\, v.$$

Thus, v is eigenvector of $J(\lambda^{1/2})$ corresponding to the eigenvalue

$$\mu = \frac{\lambda + \omega - 1}{\lambda^{1/2}\omega}.$$

Then, by the assumption on A also $\mu \in \sigma(J)$. In turn, for $\mu \in \sigma(J)$, by the same relation we see that $\lambda \in \sigma(H_\omega)$. Q.E.D.

As direct consequence of the above result, we see that for consistently ordered matrices the Gauß-Seidel matrix (case $\omega = 1$) either has spectral radius $\mathrm{spr}\,(H_1) = 0$ or there holds

$$\mathrm{spr}\,(H_1) = \mathrm{spr}\,(J)^2. \tag{3.2.32}$$

In case $\mathrm{spr}\,(J) < 1$ the Jacobi method converges. For reducing the error by the factor 10^{-1} the Gauß-Seidel method only needs half as many iterations than the Jacobi method and is therefore to be preferred. However, this does not necessarily hold in general since one can construct examples for which one or the other method converges or diverges.

For consistently ordered matrices from the identity (3.2.31), we can derive a formula for the "optimal" relaxation parameter ω_{opt} with $\mathrm{spr}\,(H_{\omega_{\mathrm{opt}}}) \le \mathrm{spr}\,(H_\omega)$, $\omega \in (0,2)$. If there holds $\rho := \mathrm{spr}\,(J) < 1$, then

$$\mathrm{spr}\,(H_\omega) = \begin{cases} \omega - 1 & , \quad \omega_{\mathrm{opt}} \le \omega \\ \frac{1}{4}\left(\rho\,\omega + \sqrt{\rho^2\omega^2 - 4(\omega-1)}\right)^2 & , \quad \omega \le \omega_{\mathrm{opt}}\,. \end{cases}$$

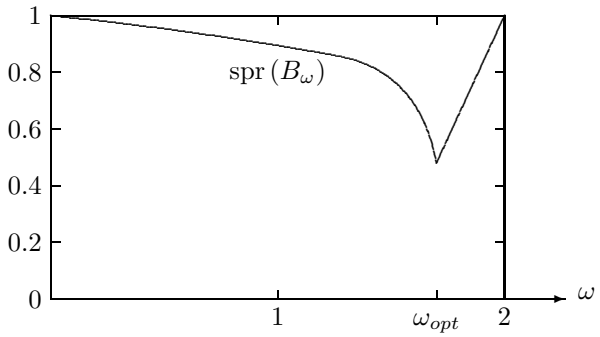


Figure 3.1: *Spectral radius of the SOR matrix H_ω as function of ω*

.

Then, there holds

$$\omega_{\mathrm{opt}} = \frac{2}{1+\sqrt{1-\rho^2}}, \qquad \mathrm{spr}\,(H_{\omega_{\mathrm{opt}}}) = \omega_{\mathrm{opt}} - 1 = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} < 1. \tag{3.2.33}$$

In general the exact value for $\mathrm{spr}\,(J)$ is not known. Since the left-sided derivative of the function $f(\omega) = \mathrm{spr}\,(H_\omega)$ for $\omega \to \omega_{\mathrm{opt}}$ is singular, in estimating ω_{opt} it is better to take a value slightly larger than the exact one. Using inclusion theorems for eigenvalues or simply the bound $\rho \le \|J\|_\infty$ one obtains estimates $\bar\rho \ge \rho$. In case $\bar\rho < 1$ this yields an upper bound $\bar\omega \ge \omega_{\mathrm{opt}}$

$$\bar\omega := \frac{2}{1+\sqrt{1-\bar\rho^2}} \ge \frac{2}{1+\sqrt{1-\rho^2}} = \omega_{\mathrm{opt}}$$

for which

$$\mathrm{spr}\,(H_{\bar\omega}) = \bar\omega - 1 = \frac{1-\sqrt{1-\bar\rho^2}}{1+\sqrt{1-\bar\rho^2}} < 1. \tag{3.2.34}$$

However, this consideration requires the formula (3.2.33) to hold true.

Example 3.3: To illustrate the possible improvement of convergence by optimal overrelaxation, we note that

$$\mathrm{spr}\,(H_1) = \mathrm{spr}\,(J)^2 = \begin{cases} 0.81 \\ 0.99 \end{cases} \quad \Rightarrow \quad \mathrm{spr}\,(H_{\omega_{\mathrm{opt}}}) = \begin{cases} 0.39 \\ 0.8 \end{cases}$$

This will be further discussed for the model matrix in Section 3.4, below.

3.2.2 Chebyshev acceleration

In the following, we discuss another method of convergence acceleration, termed "Chebyshev acceleration", which can be used in the case of a *symmetric* coefficient matrices A, for fixed-point iterations of the form

$$x^t = Bx^{t-1} + c, \quad t = 1, 2, \dots , \tag{3.2.35}$$

with *diagonalizable* iteration matrix B. First, we describe the general principle of this approach and then apply it to a symmetrized version of the SOR method. Suppose that the above fixed-point iteration converges to the solution $x \in \mathbb{R}^n$ of the linear system

$$Ax = b \quad \Leftrightarrow \quad x = Bx + c, \tag{3.2.36}$$

i. e., that $\mathrm{spr}\,(B) < 1$. The idea of Chebyshev acceleration is to construct linear combinations

$$y^t := \sum_{s=0}^{t} \gamma_s^t x^s, \quad t \geq 1, \tag{3.2.37}$$

with certain coefficients γ_s^t, such that the new sequence $(y^t)_{t\geq 0}$ converges faster to the fixed point x than the original sequence $(x^t)_{t\geq 0}$. Once the fixed-point has been reached, i. e., $x^t \approx x$, the new iterates should also be close to x. This imposes the consistency condition

$$\sum_{s=0}^{t} \gamma_s^t = 1. \tag{3.2.38}$$

Then, the corresponding error has the form

$$y^t - x = \sum_{s=0}^{t} \gamma_s^t (x^s - x) = \sum_{s=0}^{t} \gamma_s^t B^s (x^0 - x) = p_t(B)(x^0 - x), \tag{3.2.39}$$

with the polynomial $p_t \in P_t$ of degree t given by

$$p_t(z) = \sum_{s=0}^{t} \gamma_s^t z^s, \quad p_t(1) = 1. \tag{3.2.40}$$

This iteration may be viewed as one governed by a sequence of "iteration matrices" $p_t(B)$, $t = 1, 2, \dots$, and therefore, we may try to characterize its convergence by the spectral radius $\mathrm{spr}\,(p_t(B))$ as in the standard situation of a "stationary fixed-point iteration (i. e., one with

a fixed iteration matrix). This requires us to relate the eigenvalues of $p_t(B)$ to those of B,

$$\lambda(p_t(B)) = p_t(\lambda(B)). \tag{3.2.41}$$

This leads us to consider the following optimization problem

$$\operatorname{spr}(p_t(B)) = \min_{p\in P_t, p(1)=1} \max_{\lambda\in\sigma(B)} |p(\lambda)|. \tag{3.2.42}$$

Since the eigenvalues $\lambda \in \operatorname{spr}(B)$ are usually not known, rather the bound $\operatorname{spr}(B) \leq 1-\delta$ with some small $\delta > 0$ may be available, this optimization problem has to be relaxed to

$$\operatorname{spr}(p_t(B)) \leq \min_{p\in P_t, p(1)=1} \max_{|x|\leq 1-\delta} |p_t(x)|. \tag{3.2.43}$$

This optimization problem can be explicitly solved in the case $\sigma(B) \in \mathbb{R}$. Therefore, we make the following assumption.

Assumption 3.2.1: *The coefficient matrix $A = L + D + L^T$ is assumed to be symmetric and the iteration matrix B of the base iteration (3.2.35) to be similar to a symmetric matrix and, therefore, is diagonalizable with real eigenvalues,*

$$\sigma(B) \subset \mathbb{R}. \tag{3.2.44}$$

Remark 3.4: In general the iteration matrix B cannot be assumed to be symmetric and not even similar to a symmetric matrix (e. g., in the Gauß-Seidel method with $H_1 = -(D+L)^{-1}L^T$). But if this were the case (e. g., in the Richardson method with $B = I - \theta A$ or in the Jacobi method with $J = -D^{-1}(L + L^T)$) the analysis of the new sequence $(y_t)_{t\geq 0}$ may proceed as follows. Taking spectral-norms, we obtain

$$\|y^t - x\|_2 \leq \|p_t(B)\|_2 \|x^0 - x\|_2. \tag{3.2.45}$$

Hence, the convergence can be improved by choosing the polynomial p_t such the the norm $\|p_t(B)\|_2$ becomes minimal,

$$\frac{\|y^t - x\|_2}{\|x^0 - x\|_2} \leq \min_{p_t\in P_t, p_t(1)=1} \|p_t(B)\|_2 \ll \|B^t\|_2 \leq \|B\|_2^t. \tag{3.2.46}$$

Using the representation of the spectral norm, valid for symmetric matrices,

$$\|p_t(B)\|_2 = \max_{\lambda\in\sigma(B)} |p_t(\lambda)|. \tag{3.2.47}$$

and observing $\sigma(B) \in [-1+\delta, 1-\delta]$, for same small $\delta > 0$, the optimization problem takes the form

$$\min_{p_t\in P_t, p_t(1)=1} \max_{|x|\leq 1-\delta} |p_t(x)|. \tag{3.2.48}$$

The solution of the optimization problem (3.2.43) is given by the well-known Chebyshev polynomials (of the first kind), which are the orthogonal polynomials obtained by successively orthogonalizing (using the the Gram-Schmidt algorithm with exact arithmetic) the monomial

basis $\{1, x, x^2, \ldots, x^t\}$ with respect to the scalar product

$$(p, q) := \int_{-1}^{1} p(x)q(x) \frac{dx}{\sqrt{1-x^2}}, \quad p, q \in P_t,$$

defined on the function space $C[-1, 1]$. These polynomials, named $T_t \in P_t$, are usually normalized to satisfy $T_t(1) = 1$,

$$\int_{-1}^{1} T_t(x)T_s(x) \frac{dx}{\sqrt{1-x^2}} = \begin{cases} 0, & t \neq s, \\ \pi, & t = s = 0, \\ \pi/2, & t = s \neq 0. \end{cases}$$

They can be written in explicit form as (see, e. g., Stoer & Bulirsch [47] or Rannacher [1]):

$$T_t(x) = \begin{cases} (-1)^t \cosh(t \operatorname{arccosh}(-x)), & x \leq -1, \\ \cos(t \arccos(x)), & -1 \leq x \leq 1, \\ \cosh(t \operatorname{arccosh}(x)), & x \geq 1. \end{cases} \tag{3.2.49}$$

That the so defined functions are actually polynomials can be seen by induction. Further, there holds the three-term recurrence relation

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{t+1}(x) = 2xT_t(x) - T_{t-1}(x), \quad t \geq 1, \tag{3.2.50}$$

which allows the numerically stable computation and evaluation of the Chebyshev polynomials. Sometimes the following alternative global representation is useful:

$$T_t(x) = \tfrac{1}{2}\big([x + \sqrt{x^2-1}\,]^t + [x - \sqrt{x^2-1}\,]^t\big), \quad x \in \mathbb{R}. \tag{3.2.51}$$

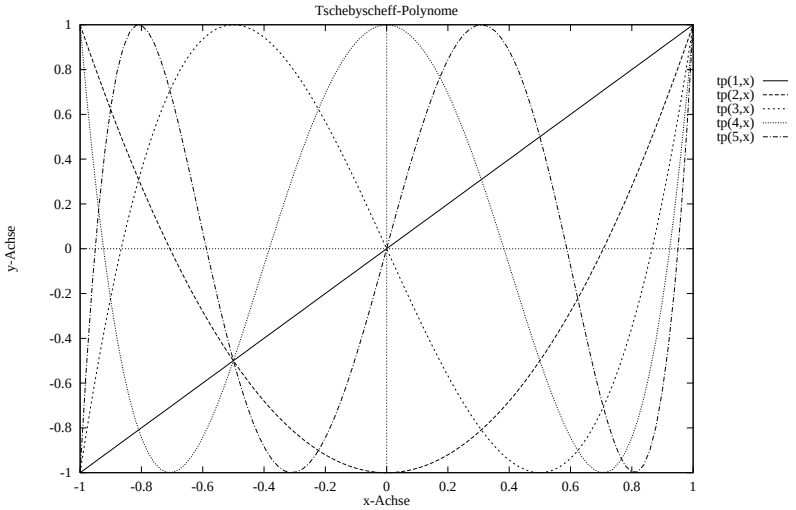


Figure 3.2: *Chebyshev polynomials* T_t, $t = 0, 1, \ldots, 5$.

With this notation, we have the following basic result.

Theorem 3.6 (Chebyshev polynomials): *Let $[a,b] \subset \mathbb{R}$ be a non-empty interval and let $c \in \mathbb{R}$ be any point outside this interval. Then, the minimum*

$$\min_{p \in P_t, p(c)=1} \max_{x \in [a,b]} |p(x)| \tag{3.2.52}$$

is attained by the uniquely determined polynomial

$$p(x) := C_t(x) = \frac{T_t(1 + 2\frac{x-b}{b-a})}{T_t(1 + 2\frac{c-b}{b-a})}, \quad x \in [a,b]. \tag{3.2.53}$$

Furthermore, for $a < b < c$ there holds

$$\min_{p \in P_t, p(c)=1} \max_{x \in [a,b]} |p(x)| = \frac{1}{T_t(1 + 2\frac{c-b}{b-a})} = \frac{2\gamma^t}{1+\gamma^{2t}} \leq 2\gamma^t, \tag{3.2.54}$$

where

$$\gamma := \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}, \quad \kappa := \frac{c-a}{c-b}.$$

Proof. (i) By affine transformation, which does not change the max-norm, we may restrict ourselves to the standard case $[a,b] = [-1,1]$ and $c \in \mathbb{R} \setminus [-1,1]$. Then, $C_t(x) = \tilde{C} T_t(x)$ with constant $\tilde{C} = T_t(c)^{-1}$. The Chebyshev polynomial $T_t(x) = \cos(t \arccos(x))$ attains the values ± 1 at the points $x_i = \cos(i\pi/t)$, $i = 0, \ldots, t$, and it alternates between 1 and -1, i. e., $T_t(x_i)$ and $T_t(x_{i+1})$ have opposite signs. Furthermore, $\max_{[-1,1]} |T_t| = 1$, implying $\max_{[-1,1]} |C_t| = |\tilde{C}|$.

(ii) Assume now the existence of $q \in P_t$ such that $\max_{[-1,1]} |q| < \max_{[-1,1]} |C_t| = |\tilde{C}|$ and $q(c) = 1$. Then, the polynomial $r = C_t - q$ changes sign t-times in the interval $[-1,1]$ since $\operatorname{sign} r(x_i) = \operatorname{sign} T_t(x_i)$, $i = 0, \ldots, t$. Thus, r has at least t zeros in $[-1,1]$. Additionally, $r(c) = 0$. Hence, $r \in P_t$ has at least $t+1$ zeros; thus, $r \equiv 0$, which leads to a contradiction.

(iii) By definition, there holds $|T_t(x)| \leq 1$, $x \in [-1,1]$. This implies that

$$\max_{x \in [a,b]} |C_t(x)| = \frac{1}{T_t(1 + 2\frac{c-b}{b-a})}.$$

The assertion then follows from the explicit representation of the T_t given above and some manipulations (for details see the proof of Theorem 3.11, below). Q.E.D.

Practical use of Chebyshev acceleration

We now assume $\sigma(B) \subset (-1,1)$, i. e., convergence of the primary iteration. Moreover, we assume that a parameter $\rho \in (-1,1)$ is known such that $\sigma(B) \subset [-\rho, \rho]$. With the parameters $a = -\rho$, $b = \rho$, and $c = 1$, we use the polynomials $p_t = C_t$ given in Theorem 3.6 in defining the secondary iteration (3.2.37). This results in the "Chebyshev-accelerated" iteration scheme. This is a consistent choice since $T_t(1) = 1$.

The naive evaluation of the secondary iterates (3.2.37) would require to store the whole convergence history of the base iteration $(x^t)_{t \geq 0}$, which may not be possible for large problems.

Fortunately, the three-term recurrence formula (3.2.50) for the Chebyshev polynomials carries over to the corresponding iterates $(y^t)_{t\geq 0}$, making the whole process feasible at all.

Since the T_t satisfy the three-term recurrence (3.2.50), and so do the polynomials $p = C_t$ from (3.2.53):

$$\mu_{t+1}p_{t+1}(x) = \frac{2x}{\rho}\mu_t p_t(x) - \mu_{t-1}p_{t-1}(x), \quad t \geq 1, \quad \mu_t = T_t(1/\rho), \tag{3.2.55}$$

with initial functions

$$p_0(x) \equiv 1, \quad p_1(x) = \frac{T_1(x/\rho)}{T_1(1/\rho)} = \frac{x/\rho}{1/\rho} = x,$$

i. e., $a_{0,0} = 1$ and $a_{1,0} = 0$, $a_{1,1} = 1$. We also observe the important relation

$$\mu_{t+1} = \frac{2}{\rho}\mu_t - \mu_{t-1}, \quad \mu_0 = 1, \ \ \mu_1 = 1/\rho. \tag{3.2.56}$$

which can be concluded from (3.2.55) observing that $p_t(1) = 1$. With these preparations, we can now implement the Chebyshev acceleration scheme. With the limit $x := \lim_{t\to\infty} x^t$, we obtain for the error $y^t - x = \tilde{e}^t = p_t(B)e^0$:

$$\begin{aligned} y^{t+1} &= x + \tilde{e}^{t+1} = x + p_{t+1}(B)e^0 = x + 2\frac{\mu_t}{\rho\mu_{t+1}}Bp_t(B)e^0 - \frac{\mu_{t-1}}{\mu_{t+1}}p_{t-1}(B)e^0 \\ &= x + 2\frac{\mu_t}{\rho\mu_{t+1}}B\tilde{e}^t - \frac{\mu_{t-1}}{\mu_{t+1}}\tilde{e}^{t-1} = x + 2\frac{\mu_t}{\rho\mu_{t+1}}B(y^t - x) - \frac{\mu_{t-1}}{\mu_{t+1}}(y^{t-1} - x) \\ &= 2\frac{\mu_t}{\rho\mu_{t+1}}By^t - \frac{\mu_{t-1}}{\mu_{t+1}}y^{t-1} + \frac{1}{\mu_{t+1}}\Big(\mu_{t+1} - \frac{2}{\rho}\mu_t B + \mu_{t-1}\Big)x. \end{aligned}$$

Now, using the fixed-point relation $x = Bx + c$ and the recurrence (3.2.56), we can remove the appearance of x in the above recurrence obtaining

$$y^{t+1} = 2\frac{\mu_t}{\rho\mu_{t+1}}By^t - \frac{\mu_{t-1}}{\mu_{t+1}}y^{t-1} + 2\frac{\mu_t}{\rho\mu_{t+1}}c, \quad y^0 = x^0, \ \ y^1 = x^1 = Bx^0 + c. \tag{3.2.57}$$

Hence, the use of Chebyshev acceleration for the primary iteration (3.2.35) consists in evaluating the three-term recurrences (3.2.56) and (3.2.57), which is of similar costs as the primary iteration (3.2.35) itself, in which the most costly step is the matrix-vector multiplication By^t.

In order to quantify the acceleration effect of this process, we write the secondary iteration in the form

$$y^t - x = \sum_{s=0}^{t} \gamma_s^t (x^s - x) = p_t(B)(x^0 - x),$$

where γ_s^t are the coefficients of the polynomial p_t. There holds

$$p_t(x) = C_t(x) = \frac{T_t(x/\rho)}{T_t(1/\rho)}.$$

By the estimate (3.2.54) of Theorem 3.6 it follows that

$$\operatorname{spr}(p_t(B)) = \max_{\lambda\in\sigma(B)} |p_t(x)| = \frac{2\gamma^t}{1+\gamma^{2t}} \leq 2\gamma^t, \quad \gamma := \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}}, \quad \kappa := \frac{1+\rho}{1-\rho}.$$

Hence, for the primary and the secondary iteration, we find the asymptotic error behavior

$$\limsup \Big(\frac{\|e^t\|}{\|e^0\|}\Big)^{1/t} = \mathrm{spr}(B) \le \rho = 1-\delta, \tag{3.2.58}$$

$$\limsup \Big(\frac{\|\tilde{e}^t\|}{\|e^0\|}\Big)^{1/t} \le \frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}} \le 1 - c'\sqrt{\delta}, \tag{3.2.59}$$

i. e., in the case $0 < \delta \ll 1$ by Chebyshev acceleration a significant improvement can be achieved for the convergence speed.

Application for accelerating the SOR method

We want to apply the concept of Chebyshev acceleration to the SOR method with the iteration Matrix (recalling that A is symmetric)

$$H_\omega = (D+\omega L)^{-1}\big((1-\omega)\,D - \omega L^T\big), \quad \omega \in (0,2).$$

However, it is not obvious whether this matrix is diagonalizable. Therefore, one introduces a symmetrized version of the SOR method, which is termed "SSOR method",

$$\begin{aligned}(D+\omega L)y^t &= [(1-\omega)D - \omega L^T]x^{t-1} + b,\\ (D+\omega L^T)x^t &= [(1-\omega)D - \omega L]y^t + b,\end{aligned}$$

or equivalently,

$$x^t = (D+\omega L^T)^{-1}[(1-\omega)\,D - \omega L](D+\omega L)^{-1}[(1-\omega)D - \omega L^T]x^{t-1} + b\big) + b\big), \tag{3.2.60}$$

with the iteration matrix

$$H_\omega^{\mathrm{SSOR}} := (D+\omega L^T)^{-1}\big((1-\omega)\,D - \omega L\big)(D+\omega L)^{-1}\big((1-\omega)D - \omega L^T\big).$$

The SSOR-iteration matrix is similar to a symmetric matrix, which is seen from the relation

$$\begin{aligned}(D+\omega L^T)H_\omega^{\mathrm{SSOR}}(D+\omega L^T)^{-1} &= [(1-\omega)\,D - \omega L](D+\omega L)^{-1}[(1-\omega)D - \omega L^T](D+\omega L^T)^{-1}\\ &= [(1-\omega)\,D - \omega L](D+\omega L)^{-1}(D+\omega L^T)^{-1}[(1-\omega)D - \omega L^T].\end{aligned}$$

The optimal relaxation parameter of the SSOR method is generally different from that of the SOR method.

Remark 3.5: In one step of the SSOR method the SOR loop is successively applied twice, once in the standard "forward" manner based on the splitting $A = (L+D) + L^T$ and then in "backward" form based on $A = L + (D+L^T)$. Hence, it is twice as expensive compared to the standard SOR method. But this higher cost is generally not compensated by faster convergence. Hence, the SSOR method is attractive mainly in connection with the Chebyshev acceleration as described above and not so much as a stand-alone solution method.

3.3 Descent methods

In the following, we consider a class of iterative methods, which are especially designed for linear systems with symmetric and positive definite coefficient matrices A, but can also be extended to more general situations. In this section, we use the abbreviated notation $(\cdot,\cdot) := (\cdot,\cdot)_2$ and $\|\cdot\| := \|\cdot\|_2$ for the euclidian scalar product and norm.

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite (and hence regular) matrix,

$$(Ax,y) = (x,Ay), \quad x,\, y \in \mathbb{R}^n, \qquad (Ax,x) > 0, \quad x \in \mathbb{R}^{n\times n} \setminus \{0\}. \tag{3.3.61}$$

This matrix generates the so-called "A-scalar product" and the corresponding "A-norm",

$$(x,y)_A := (Ax,y), \quad \|x\|_A := (Ax,x)^{1/2}, \quad x,\, y \in \mathbb{R}^n. \tag{3.3.62}$$

Accordingly, vectors with the property $(x,y)_A = 0$ are called "A-orthogonal". The positive definite matrix A has important properties. Its eigenvalues are real and positive $0 < \lambda := \lambda_1 \le \ldots \le \lambda_n =: \Lambda$ and there exists an ONB of eigenvectors $\{w_1,\ldots,w_n\}$. For its spectral radius and spectral condition number, there holds

$$\mathrm{spr}(A) = \Lambda\,, \quad \mathrm{cond}_2(A) = \frac{\Lambda}{\lambda}. \tag{3.3.63}$$

The basis for the descent methods discussed below is provided by the following theorem, which characterizes the solution of the linear system $Ax = b$ as the minimum of a quadratic functional.

Theorem 3.7 (Minimization property): *The matrix A be symmetric positive definite. The uniquely determined solution of the linear system $Ax = b$ is characterized by the property*

$$Q(x) < Q(y) \quad \forall\, y \in \mathbb{R}^n \setminus \{x\}, \qquad Q(y) := \tfrac{1}{2}(Ay,y)_2 - (b,y)_2. \tag{3.3.64}$$

Proof. Let $Ax = b$. Then, in view of the definiteness of A for $y \ne x$ there holds

$$\begin{aligned} Q(y) - Q(x) &= \tfrac{1}{2}\,\{\,(Ay,y) - 2(b,y) - (Ax,x) + 2(b,x)\,\} \\ &= \tfrac{1}{2}\,\{\,(Ay,y) - 2(Ax,y) + (Ax,x)\,\} = \tfrac{1}{2}\,(A[x-y],x-y) > 0. \end{aligned}$$

In turn, if $Q(x) < Q(y)$, for $x \ne y$, i. e., if x is a strict minimum of Q on $\mathbb{R}^n$, there must hold $\mathrm{grad}\, Q(x) = 0$. This means that (observe $a_{jk} = a_{kj}$)

$$\frac{\partial Q}{\partial x_i}(x) = \frac{1}{2}\,\frac{\partial}{\partial x_i}\sum_{j,k=1}^{n} a_{jk}x_jx_k - \frac{\partial}{\partial x_i}\sum_{k=1}^{n} b_kx_k = \sum_{k=1}^{n} a_{ik}x_k - b_i = 0, \quad i = 1,\ldots,n,$$

i. e., $Ax = b$. Q.E.D.

We note that the gradient of Q in a point $y \in \mathbb{R}^n$ is given by

$$\mathrm{grad}\; Q(y) = \tfrac{1}{2}\,(A + A^T)y - b = Ay - b. \tag{3.3.65}$$

This coincides with the "defect" of the point y with respect to the equation $Ax = b$ (negative "residual" $b - Ay$). The so-called "descent methods" determine, starting from some initial point

$x^{(0)} \in \mathbb{R}^n$, a sequence of iterates x^t, $t \geq 1$, by the prescription

$$x^{t+1} = x^t + \alpha_t r^t, \quad Q(x^{t+1}) = \min_{\alpha \in \mathbb{R}} Q(x^t + \alpha r^t). \tag{3.3.66}$$

Here, the "descent directions" r^t are a priori determined or adaptively chosen in the course of the iteration. The prescription for choosing the "step length" α_t is called "line search". In view of

$$\frac{d}{d\alpha} Q(x^t + \alpha r^t) = \operatorname{grad} Q(x^t + \alpha r^t) \cdot r^t = (Ax^t - b, r^t) + \alpha (Ar^t, r^t),$$

we obtain the formula

$$\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad g^t := Ax^t - b = \operatorname{grad} Q(x^t).$$

Definition 3.3: *The general descent method determines, starting from some initial point* $x^0 \in \mathbb{R}^n$, *a sequence of iterates* $x^t \in \mathbb{R}^n$, $t \geq 1$, *by the prescription*

$$(i)\ \textit{gradient}\ g^t = Ax^t - b, \quad (ii)\ \textit{descent direction}\ r^t,$$
$$(iii)\ \textit{step length}\ \alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad (iv)\ \textit{descent step}\ x^{t+1} = x^t + \alpha_t r^t.$$

Each descent step as described in the above definition requires two matrix-vector multiplications. By rewriting the algorithm in a slightly different way, one can save one of these multiplications at the price of additionally storing the vector Ar^t.

General descent algorithm:

Starting values:	$x^0 \in \mathbb{R}^n, \quad g^0 := Ax^0 - b.$
Iterate for $t \geq 0$:	descent direction r^t
	$\alpha_t = -\frac{(g^t, r^t)}{(Ar^t, r^t)}, \quad x^{t+1} = x^t + \alpha_t r^t, \quad g^{t+1} = g^t + \alpha_t A r^t.$

Using the notation $\|y\|_B := (By, y)^{1/2}$ there holds

$$2Q(y) = \|Ay - b\|^2_{A^{-1}} - \|b\|^2_{A^{-1}} = \|y - x\|^2_A - \|x\|^2_A, \tag{3.3.67}$$

i. e., the minimization of the functional $Q(\cdot)$ is equivalent to the minimization of the Defect norm $\|Ay - b\|_{A^{-1}}$ or the error norm $\|y - x\|_A$.

3.3.1 Gradient method

The various descent methods essentially differ by the choice of the descent directions r^t. One of the simplest a priori strategies uses in a cyclic way the cartesian coordinate direction $\{e^1, \ldots, e^n\}$. The resulting method is termed "coordinate relaxation" and is sometimes used in the context of *nonlinear* systems. For solving linear systems it is much too slow as it is in a certain sense equivalent to the Gauß-Seidel method (exercise). A more natural choice are the directions of steepest descent of $Q(\cdot)$ in the points x^t:

$$r^t = -\operatorname{grad} Q(x^t) = -g^t. \tag{3.3.68}$$

Definition 3.4: *The "gradient method" determines a sequence of iterates* $x^t \in \mathbb{R}^n$, $t \geq 0$, *by the prescription*

$$\begin{aligned} &\textit{Starting values:} && x^0 \in \mathbb{R}^n\,, \quad g^0 := Ax^0 - b. \\ &\textit{Iterate for } t \geq 0\textit{:} && \alpha_t = \tfrac{\|g^t\|^2}{(Ag^t, g^t)}, \quad x^{t+1} = x^t - \alpha_t g^t, \quad g^{t+1} = g^t - \alpha_t A g^t. \end{aligned}$$

In case that $(Ag^t, g^t) = 0$ for some $t \geq 0$ there must hold $g^t = 0$, i. e., the iteration can only terminate with $Ax^t = b$.

Theorem 3.8 (Gradient methods): *For a symmetric positive definite matrix* $A \in \mathbb{R}^{n\times n}$ *the gradient method converges for any starting point* $x^0 \in \mathbb{R}^n$ *to the solution of the linear system* $Ax = b$.

Proof. We introduce the "error functional"

$$E(y) := \|y - x\|_A^2 = (y - x, A[y-x])\,, \quad y \in \mathbb{R}^n,$$

and for abbreviation set $e^t := x^t - x$. With this notation there holds

$$\begin{aligned} \frac{E(x^t) - E(x^{t+1})}{E(x^t)} &= \frac{(e^t, Ae^t) - (e^{t+1}, Ae^{t+1})}{(e^t, Ae^t)} \\ &= \frac{(e^t, Ae^t) - (e^t - \alpha_t g^t, A[e^t - \alpha_t g^t])}{(e^t, Ae^t)} \\ &= \frac{2\alpha_t (e^t, Ag^t) - \alpha_t^2 (g^t, Ag^t)}{(e^t, Ae^t)} \end{aligned}$$

and consequently, because of $Ae^t = Ax^t - Ax = Ax^t - b = g^t$,

$$\frac{E(x^t) - E(x^{t+1})}{E(x^t)} = \frac{2\alpha_t \|g^t\|^2 - \alpha_t^2 (g^t, Ag^t)}{(g^t, A^{-1}g^t)} = \frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)}.$$

For the positive definite matrix A there holds

$$\lambda \|y\|^2 \leq (y, Ay) \leq \Lambda \|y\|^2\,, \quad \Lambda^{-1}\|y\|^2 \leq (y, A^{-1}y) \leq \lambda^{-1}\|y\|^2,$$

with $\lambda = \lambda_{\min}(A)$ and $\Lambda = \lambda_{\max}(A)$. In the case $x^t \neq x$, i. e., $E(x^t) \neq 0$ and $g^t \neq 0$, we conclude that

$$\frac{\|g^t\|^4}{(g^t, Ag^t)(g^t, A^{-1}g^t)} \geq \frac{\|g^t\|^4}{\Lambda\|g^t\|^2\, \lambda^{-1}\|g^t\|^2} = \frac{\lambda}{\Lambda},$$

and, consequently,

$$E(x^{t+1}) \leq \{\, 1 - \kappa^{-1} \,\}\, E(x^t), \quad \kappa := \mathrm{cond}_{\mathrm{nat}}(A).$$

Since $0 < 1 - 1/\kappa < 1$ for any $x^0 \in \mathbb{R}^n$ the error functional $E(x^t) \to 0$ $(t \to \infty)$, i. e., $x^t \to x$ $(t \to \infty)$. Q.E.D.

For the quantitative estimation of the speed of convergence of the gradient method, we need the following result of Kantorovich[4] .

[4]Leonid Vitalyevich Kantorovich (1912-1986): Russian Mathematician; Prof. at the U of Leningrad (1934-1960), at the Academy of Sciences (1961-1971) and at the U Moscow (1971-1976); fundamental contributions to

Lemma 3.4 (Lemma of Kantorovich): *For a symmetric, positive definite matrix* $A \in \mathbb{R}^n$ *with smallest and largest eigenvalues* λ *and* Λ*, respectively, there holds*

$$4\,\frac{\lambda\Lambda}{(\lambda+\Lambda)^2} \le \frac{\|y\|^4}{(y,Ay)(y,A^{-1}y)}\,, \quad y \in \mathbb{R}^n. \tag{3.3.69}$$

Proof. Let $\lambda = \lambda_1 \le \ldots \le \lambda_n = \Lambda$ be the eigenvalues of A and $\{w_1, \ldots, w_n\}$ a corresponding ONB of eigenvectors. An arbitrary vector $y \in \mathbb{R}^n$ admits an expansion $y = \sum_{i=1}^n y_i w_i$ with the coefficients $y_i = (y, w_i)$. Then,

$$\frac{\|y\|^4}{(y,Ay)(y,A^{-1}y)} = \frac{(\sum_{i=1}^n y_i^2)^2}{(\sum_{i=1}^n \lambda_i y_i^2)\,(\sum_{i=1}^n \lambda_i^{-1} y_i^2)} = \frac{1}{(\sum_{i=1}^n \lambda_i \zeta_i)\,(\sum_{i=1}^n \lambda_i^{-1}\zeta_i)} = \frac{\varphi(\zeta)}{\psi(\zeta)},$$

with the notation

$$\zeta = (\zeta_i)_{i=1,\ldots,n}\,, \quad \zeta_i = y_i^2 (\sum_{i=1}^n y_i^2)^{-1},$$

$$\psi(\zeta) = \sum_{i=1}^n \lambda_i^{-1}\zeta_i\,, \quad \varphi(\zeta) = (\sum_{i=1}^n \lambda_i \zeta_i)^{-1}.$$

Since the function $f(\lambda) = \lambda^{-1}$ is convex it follows from $0 \le \zeta_i \le 1$ and $\sum_{i=1}^n \zeta_i = 1$ that

$$\sum_{i=1}^n \lambda_i^{-1}\zeta_i \ge (\sum_{i=1}^n \lambda_i \zeta_i)^{-1}.$$

We set $g(\lambda) := (\lambda_1 + \lambda_n - \lambda)/(\lambda_1 \lambda_n)$.

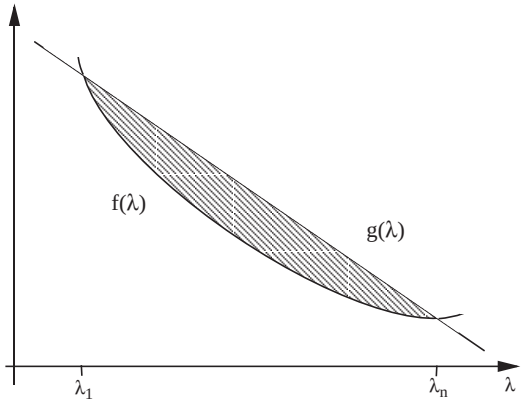


Figure 3.3: *Sketch to the proof of the Lemma of Kantorovich.*

Obviously, the graph of $\varphi(\zeta)$ lies, for all arguments ζ on the curve $f(\lambda)$, and that of $\psi(\zeta)$ between the curves $f(\lambda)$ and $g(\lambda)$ (shaded area). This implies that

$$\frac{\varphi(\zeta)}{\psi(\zeta)} \ge \min_{\lambda_1 \le \lambda \le \lambda_n} \frac{f(\lambda)}{g(\lambda)} = \frac{f([\lambda_1+\lambda_n]/2)}{g([\lambda_1+\lambda_n]/2)} = \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2},$$

which concludes the proof. Q.E.D.

linear optimization in Economy, to Functional Analysis and to Numerics (Theorem of Newton-Kantorovich).

Theorem 3.9 (Error estimate): *Let the matrix $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Then, for the gradient method the following error estimate holds:*

$$\|x^t - x\|_A \le \left(\tfrac{1-1/\kappa}{1+1/\kappa}\right)^t \|x^0 - x\|_A\,, \quad t \in \mathbb{N}, \tag{3.3.70}$$

with the spectral condition number $\kappa = \mathrm{cond}_2(A) = \Lambda/\lambda$ of A. For reducing the initial error by a factor ε the following number of iterations is required:

$$t(\varepsilon) \approx \tfrac{1}{2}\kappa\, \ln(1/\varepsilon). \tag{3.3.71}$$

Proof. (i) In the proof of Theorem 3.8 the following error identity was shown:

$$E(x^{t+1}) = \Big\{\, 1 - \frac{\|g^t\|^4}{(g^t, Ag^t)\,(g^t, A^{-1}g^t)} \,\Big\}\, E(x^t).$$

This together with the inequality (3.3.69) in the Lemma of Kantorovich yields

$$E(x^{t+1}) \le \Big\{\, 1 - 4\frac{\lambda\Lambda}{(\lambda+\Lambda)^2} \,\Big\}\, E(x^t) = \Big(\,\frac{\lambda-\Lambda}{\lambda+\Lambda}\,\Big)^2 E(x^t).$$

From this, we conclude by successive use of the recurrence that

$$\|x^t - x\|_A^2 \le \Big(\,\frac{\lambda-\Lambda}{\lambda+\Lambda}\,\Big)^{2t} \|x^0 - x\|_A^2\,, \quad t \in \mathbb{N},$$

Which proves the asserted estimate (3.3.70).

(ii) To prove (3.3.71), we take the logarithm on both sides of the relations

$$\Big(\frac{1-1/\kappa}{1+1/\kappa}\Big)^{t(\varepsilon)} = \Big(\frac{\kappa-1}{\kappa+1}\Big)^{t(\varepsilon)} < \varepsilon, \qquad \Big(\frac{\kappa+1}{\kappa-1}\Big)^{t(\varepsilon)} > \frac{1}{\varepsilon},$$

obtaining

$$t(\varepsilon) > \ln\Big(\frac{1}{\varepsilon}\Big) \ln\Big(\frac{\kappa+1}{\kappa-1}\Big)^{-1}.$$

Since

$$\ln\frac{x+1}{x-1} = 2\left\{\frac{1}{x} + \frac{1}{3}\frac{1}{x^3} + \frac{1}{5}\frac{1}{x^5} + \ldots\right\} \ge \frac{2}{x}$$

this is satisfied for $t(\varepsilon) \ge \frac{1}{2}\kappa\, \ln(1/\varepsilon)$. Q.E.D.

The relation

$$(g^{t+1}, g^t) = (g^{(t)} - \alpha_t A g^t, g^t) = \|g^t\|^2 - \alpha_t (Ag^t, g^t) = 0 \tag{3.3.72}$$

shows that the descent directions $r^t = -g^t$ used in the gradient method in consecutive steps are orthogonal to each other, while g^{t+2} may be far away form being orthogonal to g^t. This leads to strong oscillations in the convergence behavior of the gradient method especially for matrices A with large condition number, i. e., $\lambda \ll \Lambda$. In the two-dimensional case $n = 2$ this effect can be illustrated by the contour lines of the functional $Q(\cdot)$, which are eccentric ellipses, leading to a zickzack path of the gradient iteration (see Fig. 3.3.1).

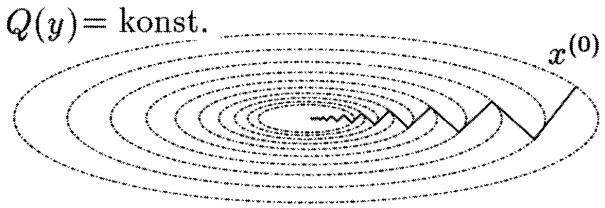


Figure 3.4: *Oscillatory convergence of the gradient method*

3.3.2 Conjugate gradient method (CG method)

The gradient method utilizes the particular structure of the functional $Q(\cdot)$, i. e., the distribution of the eigenvalues of the matrix A, only locally from one iterate x^t to the next x^{t+1}. It seems more appropriate to utilize the already obtained information about the global structure of $Q(\cdot)$ in determining the descent directions, e. g., by choosing the descent directions mutually orthogonal. This is the basic idea of the "conjugate gradient method" ("CG method") of Hestenes[5] and Stiefel[6] (1992), which successively generates a sequence of descent directions d^t which are mutually "A-orthogonal", i. e., orthogonal with respect to the scalar product $(\cdot,\cdot)_A$.

For developing the CG method, we start from the ansatz

$$B_t := \text{span}\{d^0, \cdots, d^{t-1}\} \tag{3.3.73}$$

with a set of linearly independent vectors d^i and seek to determine the iterates in the form

$$x^t = x^0 + \sum_{i=0}^{t-1} \alpha_i d^i \ \in \ x^0 + B_t, \tag{3.3.74}$$

such that

$$Q(x^t) = \min_{y \in x^0 + B_t} Q(y) \quad \Leftrightarrow \quad \|Ax^t - b\|_{A^{-1}} = \min_{y \in x^0 + B_t} \|Ay - b\|_{A^{-1}}. \tag{3.3.75}$$

Setting the derivatives of $Q(\cdot)$ with respect to the α_i to zero, we see that this is equivalent to solving the so-called "Galerkin[7]equations":

$$(Ax^t - b, d^j) = 0\,, \quad j = 0, \ldots, t-1, \tag{3.3.76}$$

or in compact form: $Ax^t - b = g^t \perp B_t$. Inserting the above ansatz for x^t into this orthogonality

[5]Magnus R. Hestenes (1906-1991): US-American mathematician; worked at the National Bureau of Standards (NBS) and the University of California at Los Angeles (UCLA); contributions to optimization and control theory and to numerical linear algebra.

[6]Eduard Stiefel (1909-1978): Swiss mathematician; since 1943 prof. for applied mathematics at the ETH Zurich; important contributions to topology, groupe theory, numerical linear algebra (CG method), approximation theory and celestrian mechanics.

[7]Boris Grigorievich Galerkin (1871-1945): Russian civil engineer and mathematician; prof. in St. Petersburg; contributions to structural mechanics especially plate bending theory.

condition, we obtain a regular linear system for the coefficients α_i, $i = 0, \ldots, t-1$,

$$\sum_{i=1}^{n} \alpha_i (Ad^i, d^j) = (b, d^j) - (Ax^0, d^j), \quad j = 0, \ldots, t-1. \tag{3.3.77}$$

Remark 3.6: We note that (3.3.76) does not depend on the symmetry of the matrix A. Starting from this relation one may construct CG-like methods for linear systems with asymmetric and even indefinite coefficient matrices. Such methods are generally termed "projection methods". Methods of this type will be discussed in more detail below.

Recall that the Galerkin equations (3.3.76) are equivalent to minimizing the defect norm $\|Ax^t - b\|_{A^{-1}}$ or the error norm $\|x^t - x\|_A$ on $x^0 + B_t$. Natural choices for the spaces B_t are the so-called Krylov[8]spaces

$$B_t = K_t(d^0; A) := \text{span}\{d^0, Ad^0, \ldots, A^{t-1}d^0\}, \tag{3.3.78}$$

with some vector d^0, e. g., the (negative) initial defect $d^0 = b - Ax^0$ of an arbitrary vector x^0. This is motivated by the observation that from $A^t d^0 \in K_t(d^0; A)$, we necessarily obtain

$$-g^t = b - Ax^t = d^0 + A(x^0 - x^t) \in d^0 + AK_t(d^0; A) \in K_t(d^0; A).$$

Because $g^t \perp K_t(d^0; A)$, this implies $g^t = 0$ by construction.

Now the CG method constructs a sequence of descent directions, which form an A-orthogonal basis of the Krylov spaces $K_t(d^0; A)$. We proceed in an inductive way: Starting from an arbitrary point x^0 with (negative) defect $d^0 = b - Ax^0$ let iterates x^i and corresponding descent directions $d^i (i = 0, ..., t-1)$ already been determined such that $\{d^0, ..., d^{t-1}\}$ is an A-orthogonal basis of $K_t(d^0; A)$. For the construction of the next descent direction $d^t \in K_{t+1}(d^0; A)$ with the property $d^t \perp_A K_t(d^0; A)$ we make the ansatz

$$d^t = -g^t + \sum_{j=0}^{t-1} \beta_j^{t-1} d^j \ \in \ K_{t+1}(d^0; A). \tag{3.3.79}$$

Here, we can assume that $g^t = Ax^t - b \notin K_t(d^0; A)$ as otherwise $g^t = 0$ and, consequently, $x^t = x$. Then, for $i = 0, ..., t-1$ there holds

$$(d^t, Ad^i) = (-g^t, Ad^i) + \sum_{j=0}^{t-1} \beta_j^{t-1} (d^j, Ad^i) = (-g^t + \beta_i^{t-1} d^i, Ad^i). \tag{3.3.80}$$

For $i < t-1$, we have $(g^t, Ad^i) = 0$ since $Ad^i \in K_t(d^0; A)$ and, consequently, $\beta_i^{t-1} = 0$. For $i = t-1$, the condition

$$0 = (-g^t, Ad^{t-1}) + \beta_{t-1}^{t-1} (d^{t-1}, Ad^{t-1}) \tag{3.3.81}$$

[8]Aleksei Nikolaevich Krylov (1863-1945): Russian mathematician; prof. at the Sov. Academy of Sciences in St. Petersburg; contributions to Fourier analysis and differential equations, applications in ship building.

leads us to the formulas

$$\beta_{t-1} := \beta_{t-1}^{t-1} = \frac{(g^t, Ad^{t-1})}{(d^{t-1}, Ad^{t-1})}\,, \quad d^t = -g^t + \beta_{t-1} d^{t-1}. \tag{3.3.82}$$

The next iterates x^{t+1} and $g^{t+1} = Ax^{t+1} - b$ are then determined by

$$\alpha_t = -\frac{(g^t, d^t)}{(d^t, Ad^t)}\,, \quad x^{t+1} = x^t + \alpha_t d^t\,, \quad g^{t+1} = g^t + \alpha_t A d^t. \tag{3.3.83}$$

These are the recurrence equations of the classical CG method. By construction there holds

$$(d^t, Ad^i) = (g^t, d^i) = 0, \quad i \le t-1\,, \quad (g^t, g^{t-1}) = 0. \tag{3.3.84}$$

From this, we conclude that

$$\|g^t\|^2 = (d^t - \beta_{t-1} d^{t-1}, -g^{t+1} + \alpha_t A d^t) = \alpha_t (d^t, Ad^t), \tag{3.3.85}$$

$$\|g^{t+1}\|^2 = (g^t + \alpha_t A d^t, g^{t+1}) = \alpha_t (Ad^t, g^{t+1}). \tag{3.3.86}$$

This allows for the following simplifications in the above formulas:

$$\alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad \beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \tag{3.3.87}$$

as long as the iteration does not terminate with $g^t = 0$.

Definition 3.5: *The CG method determines a sequence of iterates $x^t \in \mathbb{R}^n$, $t \ge 0$, by the prescription*

Starting values: $x^0 \in \mathbb{R}^n\,, \quad d^0 = -g^0 = b - Ax^0,$

Iterate for $t \ge 0$: $\alpha_t = \frac{\|g^t\|^2}{(d^t, Ad^t)}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t A d^t,$
$\beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t.$

By construction the CG method generates a sequence of descent directions d^t, which are automatically A-orthogonal. This implies that the vectors $d^0, \ldots, d^t$ are linearly independent and that therefore $\operatorname{span}\{d^0, \ldots, d^{n-1}\} = \mathbb{R}^n$. We formulate the properties of the CG method derived so far in the following theorem.

Theorem 3.10 (CG method): *Let the matrix $A \in \mathbb{R}^{n\times n}$ be symmetric positive definite. Then, (assuming round-off free arithmetic) the CG method terminates for any starting vector $x^0 \in \mathbb{R}^n$ after at most n steps at $x^n = x$. In each step there holds*

$$Q(x^t) = \min_{y \in x^0 + B_t} Q(y), \tag{3.3.88}$$

and, equivalently,

$$\|x^t - x\|_A = \|Ax^t - b\|_{A^{-1}} = \min_{y \in x^0 + B_t} \|Ay - b\|_{A^{-1}} = \min_{y \in x^0 + B_t} \|y - x\|_A, \tag{3.3.89}$$

where $B_t := \operatorname{span}\{d^0, \ldots, d^{t-1}\}$.

In view of the result of Theorem 3.10 the CG method formally belongs to the class of "direct" methods. In practice, however, it is used like an iterative method, since:

1. Because of round-off errors the descent directions d^t are not exactly A-orthogonal such that the iteration does not terminate.
2. For large matrices one obtains accurate approximations already after $t \ll n$ iterations.

As preparation for the main theorem about the convergence of the CG method, we provide the following auxiliary lemma.

Lemma 3.5 (Polynomial norm bounds): *Let A be a symmetric positive definite matrix with eigenvalues $\sigma \subset [a,b]$. Then, for any polynomial $p \in P_t$, $p(0) = 1$ there holds*

$$\|x^t - x\|_A \leq M \, \|x^0 - x\|_A, \quad M := \sup_{\mu \in [a,b]} |p(\mu)|. \tag{3.3.90}$$

Proof. Observing the relation

$$\|x^t - x\|_A = \min_{y \in x^0 + B_t} \|y - x\|_A,$$

$$B_t = \operatorname{span}\{d^0, \ldots, d^{t-1}\} = \operatorname{span}\{A^0 g^{(0)}, \ldots, A^{t-1} g^0\},$$

we find

$$\|x^t - x\|_A = \min_{p \in P_{t-1}} \|x^0 - x + p(A) g^0\|_A.$$

Since $g^0 = Ax^0 - b = A(x^0 - x)$ it follows

$$\begin{aligned} \|x^t - x\|_A &= \min_{p \in P_{t-1}} \|[I + Ap(A)](x^0 - x)\|_A \\ &\leq \min_{p \in P_{t-1}} \|I + Ap(A)\|_A \, \|x^0 - x\|_A \\ &\leq \min_{p \in P_t, p(0)=1} \|p(A)\|_A \, \|x^0 - x\|_A, \end{aligned}$$

with the natural matrix norm $\|\cdot\|_A$ generated from the A-norm $\|\cdot\|_A$. Let $0 < \lambda \leq \ldots \lambda_n$ be the eigenvalues and $\{w^1, \ldots, w^n\}$ a corresponding ONS of eigenvectors of the symmetric, positive definite matrix A. Then, for arbitrary $y \in \mathbb{R}^n$ there holds

$$y = \sum_{i=1}^{n} \gamma_i w_i \,, \quad \gamma_i = (y, w_i),$$

and, consequently,

$$\|p(A)y\|_A^2 = \sum_{i=1}^{n} \lambda_i p(\lambda_i)^2 \gamma_i^2 \leq M^2 \sum_{i=1}^{n} \lambda_i \gamma_i^2 = M^2 \, \|y\|_A^2.$$

This implies

$$\|p(A)\|_A = \sup_{y \in \mathbb{R}^n, \, y \neq 0} \frac{\|p(A)y\|_A}{\|y\|_A} \leq M,$$

which completes the proof. Q.E.D.

As a consequence of Lemma 3.5, we obtain the following a priori error estimate.

Theorem 3.11 (CG convergence): *Let A be a symmetric positive definite matrix. Then, for the CG method there holds the error estimate*

$$\|x^t - x\|_A \leq 2\Big(\tfrac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}\Big)^t \|x^0 - x\|_A\,, \quad t \in \mathbb{N}\,, \tag{3.3.91}$$

with the spectral condition number $\kappa = \mathrm{cond}_2(A) = \Lambda/\lambda$ of A. For reducing the initial error by a factor ε the following number of iteration is required:

$$t(\varepsilon) \approx \tfrac{1}{2}\sqrt{\kappa}\,\ln(2/\varepsilon). \tag{3.3.92}$$

Proof. (i) Setting $[a,b] := [\lambda,\Lambda]$ in Lemma 3.5, we obtain

$$\|x^t - x\|_A \leq \min_{p\in P_t,\, p(0)=1} \Big\{ \sup_{\lambda\leq\mu\leq\Lambda} |p(\mu)| \Big\}\, \|x^0 - x\|_A.$$

This yields the assertion if we can show that

$$\min_{p\in P_t,\, p(0)=1} \Big\{ \sup_{\lambda\leq\mu\leq\Lambda} |p(\mu)| \Big\} \leq 2\Big(\frac{1-\sqrt{\lambda/\Lambda}}{1+\sqrt{\lambda/\Lambda}}\Big)^t.$$

This is again a problem of approximation theory with respect to the max-norm (Chebyshev approximation), which can be solved using the Chebyshev polynomials described above in Subsection 3.2.2. The solution $p_t \in P_t$ is give by

$$p_t(\mu) = T_t\Big(\frac{\Lambda+\lambda-2\mu}{\Lambda-\lambda}\Big)\; T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big)^{-1},$$

with the t-th Chebyshev polynomial T_t on $[-1,1]$. There holds

$$\sup_{\lambda\leq\mu\leq\Lambda} p_t(\mu) = T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big)^{-1}.$$

From the representation

$$T_t(\mu) = \tfrac{1}{2}\left[\big(\mu+\sqrt{\mu^2-1}\big)^t + \big(\mu-\sqrt{\mu^2-1}\big)^t\right], \quad \mu\in[-1,1],$$

for the Chebyshev polynomials and the identity

$$\frac{\kappa+1}{\kappa-1} + \sqrt{\Big(\frac{\kappa+1}{\kappa-1}\Big)^2 - 1} = \frac{\kappa+1}{\kappa-1} + \frac{2\sqrt{\kappa}}{\kappa-1} = \frac{(\sqrt{\kappa}+1)^2}{\kappa-1} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1},$$

we obtain the estimate

$$T_t\Big(\frac{\Lambda+\lambda}{\Lambda-\lambda}\Big) = T_t\Big(\frac{\kappa+1}{\kappa-1}\Big) = \frac{1}{2}\left[\Big(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\Big)^t + \Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^t\right] \geq \frac{1}{2}\Big(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\Big)^t.$$

Hence,

$$\sup_{\lambda\leq\mu\leq\Lambda} p_t(\mu) \leq 2\Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^t,$$

which implies (3.3.91).

(ii) For deriving (3.3.92), we require

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{t(\varepsilon)} \leq \varepsilon,$$

and, equivalently,

$$t(\varepsilon) > \ln\left(\frac{2}{\varepsilon}\right) \ln\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^{-1}.$$

Since

$$\ln\frac{x+1}{x-1} = 2\left\{\frac{1}{x} + \frac{1}{3}\frac{1}{x^3} + \frac{1}{5}\frac{1}{x^5} + \ldots\right\} \geq \frac{2}{x},$$

this is satisfied for $t(\varepsilon) \geq \frac{1}{2}\sqrt{\kappa}\,\ln(2/\varepsilon)$. Q.E.D.

Since $\kappa = \mathrm{cond}_{\mathrm{nat}}(A) > 1$, we have $\sqrt{\kappa} < \kappa$. Observing that the function $f(\lambda) = (1-\lambda^{-1})\,(1+\lambda^{-1})^{-1}$ is strictly monotonically increasing for $\lambda > 0$ $(f'(\lambda) > 0)$, there holds

$$\frac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}} < \frac{1-1/\kappa}{1+1/\kappa},$$

i. e., the CG method should converge faster than the gradient method. This actually the case in practice. Both methods converge the faster the smaller the condition number is. In case $\Lambda \gg \lambda$, which is frequently the case in practice, also the CG method is too slow. An acceleration can be achieved by so-called "preconditioning", which will be described below.

3.3.3 Generalized CG methods and Krylov space methods

For solving a general linear system $Ax = b$, with regular but not necessarily positive definite matrix $A \in \mathbb{R}^n$, by the CG method, one may consider the equivalent system

$$A^T A x = A^T b \tag{3.3.93}$$

with the symmetric, positive definite matrix $A^T A$. Applied to this system the CG method takes the following form:

Starting values: $x^0 \in \mathbb{R}^n\,, \quad d^0 = A^T(b - Ax^0) = -g^0,$

for $t \geq 0$: $\alpha_t = \frac{\|g^t\|^2}{\|Ad^t\|^2}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad g^{t+1} = g^t + \alpha_t A^T A d^t,$

$\beta_t = \frac{\|g^{t+1}\|^2}{\|g^t\|^2}, \quad d^{t+1} = -g^{t+1} + \beta_t d^t.$

This approach is referred to as CRS method ("Conjugate Residual Squared") of P. Sonneveld, 1989). The convergence speed is characterized by $\mathrm{cond}_2(A^T A)$. The whole method is equivalent to minimizing the functional

$$Q(y) := \tfrac{1}{2}\,(A^T A y, y) - (A^T b, y) = \tfrac{1}{2}\|Ay - b\|^2 - \tfrac{1}{2}\,\|b\|^2. \tag{3.3.94}$$

Since $\mathrm{cond}_2(A^T A) \approx \mathrm{cond}_2(A)^2$ the convergence of this variant of the CG method may be rather slow. However, its realization does not require the explicit evaluation of the matrix product $A^T A$ but only the computation of the matrix-vector products $z = Ay$ and $A^T z$.

On the basis of the formulation(3.3.75) the standard CG method is limited to linear systems with symmetric, positive definite matrices. But starting from the (in this case equivalent) Galerkin formulation (3.3.76) the method becomes meaningful also for more general matrices. In fact, in this way one can derive effective generalizations of the CG method also for nonsymmetric and even indefinite matrices. These modified CG methods are based on the Galerkin equations (3.3.76) and differ in the choices of "ansatz spaces" K_t and "test spaces" K_t^*,

$$x^t \in x^0 + K_t : \quad (Ax^t - b, y) = 0 \quad \forall\, y \in K_t^*. \tag{3.3.95}$$

Here, one usually uses the Krylov spaces

$$K_t = \text{span}\{d^0, Ad^0, ..., A^{t-1}d^0\},$$

combined with the test spaces $K_t^* = K_t$, or

$$K_t^* = \text{span}\{d^0, A^T d^0, ..., (A^T)^{t-1}d^0\}.$$

This leads to the general class of "Krylov space methods". Most popular representatives are the following methods, which share one or the other property with the normal CG method but generally do not allow for a similarly complete error analysis.

1. GMRES with or without restart ("Generalized Minimal Residual") of Y. Saad and M. H. Schultz, 1986): $K_t = \text{span}\{d^0, Ad^0, ..., A^{t-1}d^0\} = K_t^*$,

$$\|Ax^t - b\| = \min_{y \in x^0 + K_t} \|Ay - b\|. \tag{3.3.96}$$

   Since this method minimizes the residual over spaces of increasing dimension as the normal CG method also the GMRES methods yields the exact solution after at most n steps. However, for general nonsymmetric matrices the iterates x^t cannot be obtained by a simple tree-term recurrence as in the normal CG method. It uses a full recurrence, which results in high storage requirements. Therefore, to limit the costs the GMRES method is stopped after a certain number of steps, say k steps, and then restarted with x^k as new starting vector. The latter variant is denoted by "GMRES(k)" method.

2. BiCG and BiCGstab ("Biconjugate Gradient Stabilized" of H. A. Van der Vorst, 1992: $K_t = \text{span}\{d^0, Ad^0, ..., A^{t-1}d^0\}$, $K_t^* = \text{span}\{d^0, A^T d^0, ..., (A^T)^{t-1}d^0\}$,

$$x^t \in x^0 + K_t : \quad (Ax^t - b, y) = 0, \quad \forall y \in K_t^*. \tag{3.3.97}$$

   In the BiCG method the iterates x^t are obtained by a three-term recurrence but for an unsymmetric matrix the residual minimization property gets lost and the method may not even converge. Additional stability is provided in the "BiCGstab" method.

Both methods, GMRES(k) and BiCGstab, are especially designed for unsymmetric but definite matrices. They have there different pros and cons and are both not universally applicable. One can construct matrices for which one or the other of the methods does not work. The methods for the practical computation of the iterates x^t in the Krylov spaces K_t are closely related to the Lanczos and Arnoldi algorithms used for solving the corresponding eigenvalue problems, discussed in Chapter 4, below.

3.3.4 Preconditioning (PCG methods)

The error estimate (3.3.91) for the CG method indicates a particularly good convergence if the condition number of the matrix A is close to one. Hence, in case of large $\mathrm{cond}_2(A) \gg 1$, one uses "preconditioning", i. e., the system $Ax = b$ is transformed into an equivalent one, $\tilde{A}\tilde{x} = \tilde{b}$ with a better conditioned matrix $\tilde{A}$. To this end, let C be a symmetric, positive definite matrix, which is explicitly given in product form

$$C = KK^T, \tag{3.3.98}$$

with a regular matrix K. Then, the system $Ax = b$ can equivalently be written in the form

$$\underbrace{K^{-1}A\,(K^T)^{-1}}_{\tilde{A}}\ \underbrace{K^T x}_{\tilde{x}} = \underbrace{K^{-1}b}_{\tilde{b}}. \tag{3.3.99}$$

Then, the CG method is formally applied to the transformed system $\tilde{A}\tilde{x} = \tilde{b}$, while it is hoped that $\mathrm{cond}_2(\tilde{A}) \ll \mathrm{cond}_2(A)$ for an appropriate choice of C. The relation

$$(K^T)^{-1}\tilde{A}K^T = (K^T)^{-1}K^{-1}A(K^T)^{-1}K^T = C^{-1}A \tag{3.3.100}$$

shows that for $C \equiv A$ the matrix $\tilde{A}$ is similar to I, and therefore $\mathrm{cond}_2(\tilde{A}) = \mathrm{cond}_2(I) = 1$. Consequently, one chooses $C = KK^T$ such that C^{-1} is a good approximation of A^{-1}.

The CG method for the transformed system $\tilde{A}\tilde{x} = \tilde{b}$ can then be written in terms of the quantities A, b and x as so-called "PCG method" ("Preconditioned" CG method) as follows:

Starting value: $x^0 \in \mathbb{R}^N$, $d^0 = r^0 = b - Ax^0$, $\mathbf{C\rho^0 = r^0}$,

for $t \geq 0$:
$$\alpha_t = \frac{\langle r^t, \rho^t\rangle}{\langle Ad^t, d^t\rangle}, \quad x^{t+1} = x^t + \alpha_t d^t, \quad r^{t+1} = r^t - \alpha_t A d^t, \quad \mathbf{C\rho^{t+1} = r^{t+1}},$$
$$\beta_t = \frac{\langle r^{t+1}, \rho^{t+1}\rangle}{\langle r^t, \rho^t\rangle}, \quad d^{t+1} = r^{t+1} + \beta_t d^t.$$

Compared to the normal CG method the PCG iteration in each step additionally requires the solution of the system $C\rho^{t+1} = r^{t+1}$, which is easily accomplished using the decomposition $C = KK^T$. In order to preserve the work complexity $\mathcal{O}(n)$ a. op. in each step the triangular matrix K should have a sparsity pattern similar to that of the lower triangular part L of A. This condition is satisfied by the following popular preconditioners:

1) Diagonal preconditioning (scaling): $C := D = D^{1/2}D^{1/2}$.
The scaling ensures that the elements of A are brought to approximately the same size, especially with $\tilde{a}_{ii} = 1$. This reduces the condition number since

$$\mathrm{cond}_2(A) \geq \frac{\max_{1\leq i\leq n}\ a_{ii}}{\min_{1\leq i\leq n}\ a_{ii}}. \tag{3.3.101}$$

Example: The matrix $A = \mathrm{diag}\{\lambda_1 = ... = \lambda_{n-1} = 1, \lambda_n = 10^k\}$ has the condition number $\mathrm{cond}_2(A) = 10^k$, while the scaled matrix $\tilde{A} = D^{-1/2}AD^{-1/2}$ has the optimal condition number $\mathrm{cond}_2(\tilde{A}) = 1$.

2) SSOR preconditioning: We choose

$$\begin{aligned} C &:= (D+L)D^{-1}(D+L^T) = D+L+L^T+LD^{-1}L^T \\ &= (\underbrace{D^{1/2}+LD^{-1/2}}_{K})(\underbrace{D^{1/2}+D^{-1/2}L^T}_{K^T}), \end{aligned}$$

or, more generally, involving a relaxation parameter $\omega \in (0,2)$,

$$\begin{aligned} C &:= \frac{1}{2-\omega}\Big(\frac{1}{\omega}D+L\Big)\Big(\frac{1}{\omega}D\Big)^{-1}\Big(\frac{1}{\omega}D+L^T\Big) \\ &= \underbrace{\frac{1}{\sqrt{(2-\omega)\omega}}(D^{1/2}+\omega LD^{-1/2})}_{K}\ \underbrace{\frac{1}{\sqrt{(2-\omega)\omega}}(D^{1/2}+\omega D^{-1/2}L^T)}_{K^T}. \end{aligned}$$

Obviously, the triangular matrix K has the same sparsity pattern as L. Each step of the preconditioned iteration costs about twice as much work as the basic CG method. For an optimal choice of the relaxation parameter ω (not easy to determine) there holds

$$\mathrm{cond}_2(\tilde{A}) = \sqrt{\mathrm{cond}_2(A)}.$$

3) ICCG preconditioning (Incomplete Cholesky Conjugate Gradient): The symmetric, positive definite matrix A has a Cholesky decomposition $A = LL^T$ with an lower triangular matrix $L = (l_{ij})_{i,j=1}^n$. The elements of L are successively determined by the recurrence formulas

$$l_{ii} = \Big(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2\Big)^{1/2}, \quad i = 1,\dots,n, \qquad l_{ji} = \frac{1}{l_{ii}}\Big(a_{ji} - \sum_{k=1}^{i-1} l_{jk}l_{ik}\Big), \quad j = i+1,\dots,n.$$

The matrix L generally has nonzero elements in the whole band of A, which requires much more memory than A itself. This can be avoided by performing (such as in the ILU approach discussed in Subsection 3.1.2) only an "incomplete" Cholesky decomposition where within the elimination process some of the l_{ji} are set to zero, e. g., those for which $a_{ji} = 0$. This results in an incomplete decomposition

$$A = \tilde{L}\tilde{L}^T + E \tag{3.3.102}$$

with a lower triangular matrix $\tilde{L} = (\tilde{l}_{ij})_{i,j=1}^n$, which has a similar sparsity pattern as A. In this case, one speaks of the "$ICCG(0)$ variant". In case of a band matrix A, one may allow the elements of $\tilde{L}$ to be nonzero in further p off-diagonals resulting in the so-called "$ICCG(p)$ variant" of the $ICCG$ method, which is hoped to provide a better approximation $C^{-1} \approx A^{-1}$ for increasing p. Then, for preconditioning the matrix

$$C = KK^T := \tilde{L}\tilde{L}^T \tag{3.3.103}$$

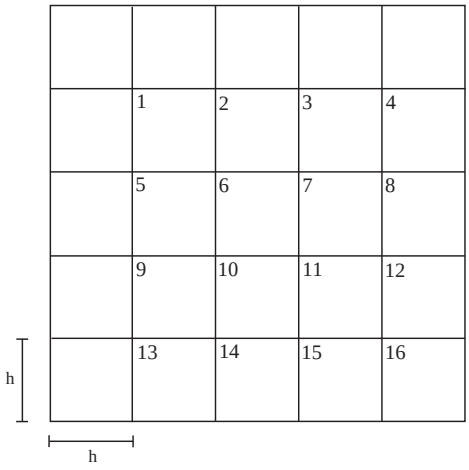
is used. Although, there is no full theoretical justification yet for the success of the $ICCG$ preconditioning practical tests show a significant improvement in the convergence behavior. This may be due to the fact that, though the condition number is not necessarily decreased, the eigenvalues of the corresponding transformed matrix $\tilde{A}$ cluster more around $\lambda = 1$.

3.4 A model problem

At the end of the discussion of simple iterative methods for solving linear systems $Ax = b$, we will determine their convergence rates for the model situation already described in Section 0.4.2 of Chapter 0. We consider the so-called "1-st boundary value problem of the Laplace operator"

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2}(x,y) - \frac{\partial^2 u}{\partial y^2}(x,y) &= f(x,y) \quad \text{for } (x,y) \in \Omega \\ u(x,y) &= 0 \quad \text{for } (x,y) \in \partial\Omega, \end{aligned} \tag{3.4.104}$$

on the unit square $\Omega = (0,1) \times (0,1) \subset \mathbb{R}^2$. For solving this problem the domain Ω is covered by a uniform mesh as shown in Fig. 3.4.



$h = \frac{1}{m+1}$ mesh size

$n = m^2$ number of unknown mesh values

Figure 3.5: *Mesh for the discretization of the model problem*

The "interior" mesh points are numbered row-wise. On this mesh the second derivatives in the differential equation (3.4.104) are approximated by second-order central difference quotients leading to the following difference equations for the mesh unknowns $U(x,y) \approx u(x,y)$:

$$-h^{-2}\big\{U(x+h,y) - 2U(x,y) + U(x-h,y) + U(x,y+h) - 2U(x,y) + U(x,y-h)\big\} = f(x,y).$$

Observing the boundary condition $u(x,y) = 0$ for $(x,y) \in \partial\Omega$ this set of difference equations is equivalent to the linear system

$$Ax = b, \tag{3.4.105}$$

for the vector $x \in \mathbb{R}^n$ of unknown mesh values $x_i \approx u(P_i)$, P_i interior mesh point. The matrix A has the already known form

$$A = \begin{bmatrix} B & -I & & \\ -I & B & -I & \\ & -I & B & \ddots \\ & & \ddots & \ddots \end{bmatrix} \Bigg\} n \qquad B = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix} \Bigg\} m$$

with the $m\times m$-unit matrix I. The right-hand side is given by $b = h^2(f(P_1),\dots,f(P_n))^T$. The matrix A has several special properties:

- "sparse band matrix" with bandwidth $2m+1$;
- "irreducible" and "strongly diagonally dominant";
- "symmetric" and "positive definite";
- "consistently ordered";
- "of nonnegative type" ("M-matrix"): $a_{ii} > 0,\ a_{ij} \le 0,\ i \ne j$.
  The importance of this last property will be illustrated in an exercise.

For this matrix eigenvalues and eigenvectors can be explicitly determined $(h = 1/(m+1))$:

$$\lambda_{kl} = 4 - 2\,(\cos[kh\pi] + \cos[lh\pi]), \quad w^{kl} = (\sin[ikh\pi]\ \sin[jlh\pi])_{i,j=1,\dots,m}, \quad k,l = 1,\dots,m,$$

i. e., $Aw^{kl} = \lambda_{kl} w^{kl}$. Hence for $h \ll 1$, we have

$$\Lambda := \lambda_{\max} = 4 - 4\,\cos\,(1-h)\pi \;\approx\; 8,$$
$$\lambda := \lambda_{\min} = 4 - 4\,\cos\,(h\pi) = 4 - 4\,(1 - \frac{\pi^2}{2}\,h^2 + O(h^4)) \;\approx\; 2\pi^2 h^2,$$

and consequently

$$\kappa \;:=\; \mathrm{cond}_2(A) \;\approx\; \frac{4}{\pi^2 h^2}. \tag{3.4.106}$$

Then, the eigenvalues of the Jacobi iteration matrix $J = -D^{-1}(L+R)$ are given by

$$\mu_{kl}(J) = \tfrac{1}{2}\,(\cos[kh\pi] + \cos[lh\pi]), \quad k,l = 1,\dots,m.$$

Hence,

$$\rho := \mathrm{spr}\,(J) = \mu_{\max}(J) = \cos[h\pi] = 1 - \frac{\pi^2}{2}\,h^2 + O(h^4). \tag{3.4.107}$$

For the iteration matrices of the Gauß-Seidel and the optimal SOR iteration matrices, H_1 and $H_{\omega_{\mathrm{opt}}}$, respectively, there holds

$$\mathrm{spr}\,(H_1) = \rho^2 = 1 - \pi^2 h^2 + O(h^4), \tag{3.4.108}$$
$$\mathrm{spr}\,(H_{\omega_{\mathrm{opt}}}) = \frac{1-\sqrt{1-\rho^2}}{1+\sqrt{1-\rho^2}} = \frac{1-\pi h + O(h^2)}{1+\pi h + O(h^2)} = 1 - 2\,\pi h + O(h^2)\,. \tag{3.4.109}$$

Comparison of convergence speed

Now, we make a comparison of the convergence speed of the various iterative methods considered above. The reduction of the initial error $\|x^{(0)} - x\|_2$ in a fixed-point iteration by the factor $\varepsilon \ll 1$ requires about $T(\varepsilon)$ iterations,

$$T(\varepsilon) \;\approx\; \frac{\ln(1/\varepsilon)}{\ln(1/\rho)}, \quad \rho = \mathrm{spr}\,(B),\ \ B = I - C^{-1}A \quad \text{iteration matrix.} \tag{3.4.110}$$

Using the above formulas, we obtain:

$$T_J(\varepsilon) \approx -\frac{\ln(1/\varepsilon)}{\ln(1-\frac{\pi^2}{2}h^2)} \approx 2\frac{\ln(1/\varepsilon)}{\pi^2 h^2} = \frac{2}{\pi^2} n \ln(1/\varepsilon),$$
$$T_{GS}(\varepsilon) \approx -\frac{\ln(1/\varepsilon)}{\ln(1-\pi^2 h^2)} \approx \frac{\ln(1/\varepsilon)}{\pi^2 h^2} = \frac{1}{\pi^2} n \ln(1/\varepsilon),$$
$$T_{SOR}(\varepsilon) \approx -\frac{\ln(1/\varepsilon)}{\ln(1-2\pi h)} \approx \frac{\ln(1/\varepsilon)}{2\pi h} = \frac{1}{2\pi}\sqrt{n} \ln(1/\varepsilon).$$

The gradient method and the CG method require for the reduction of the initial error $\|x^0 - x\|_2$ by the factor $\varepsilon \ll 1$ the following numbers of iterations:

$$T_G(\varepsilon) = \frac{1}{2}\kappa \ln(2/\varepsilon) \approx \frac{2}{\pi^2 h^2} \ln(1/\varepsilon) \approx \frac{2}{\pi^2} n \ln(1/\varepsilon),$$
$$T_{CG}(\varepsilon) = \frac{1}{2}\sqrt{\kappa} \ln(2/\varepsilon) \approx \frac{1}{\pi h} \ln(2/\varepsilon) \approx \frac{1}{\pi}\sqrt{n} \ln(2/\varepsilon).$$

We see that the Jacobi method and the gradient method converge with about the same speed. The CG method is only half as fast as the (optimal) SOR method, but it does not require the determination of an optimal parameter (while the SOR method does not require the matrix A to be symmetric). The Jacobi method with Chebyshev acceleration is as fast as the "optimal" SOR method but also does not require the determination of an optimal parameter.

For the special right-hand side function $f(x,y) = 2\pi^2 \sin(\pi x)\sin(\pi y)$ the exact solution of the boundary value problem is given by

$$u(x,y) = \sin(\pi x)\sin(\pi y). \tag{3.4.111}$$

The error caused by the finite difference discretization considered above can be estimates as follows:

$$\max_{P_i} |u(P_i) - x_i| \le \frac{\pi^4}{12} h^2 + O(h^4). \tag{3.4.112}$$

Hence, for achieving a relative accuracy of $TOL = 10^{-3}$ (three decimals) a mesh size

$$h \approx \frac{\sqrt{12}}{\pi^2} 10^{-3/2} \approx 10^{-2},$$

is required. This results in $n \approx 10^4$ unknowns. In this case, we obtain for the above spectral radii, conditions numbers and numbers of iterations required for error reduction by $\varepsilon = 10^{-4}$ (including a safety factor of $1/10$) the following values $(\ln(1/\varepsilon) \sim 10)$:

$$\begin{aligned} \operatorname{spr}(J) &\approx 0,9995 & T_{\mathrm{J}} &\approx 20.000 \\ \operatorname{spr}(H_1) &\approx 0,999 & T_{\mathrm{GS}} &\approx 10.000 \\ \operatorname{spr}(H_{\omega *}) &\approx 0,9372 & T_{\mathrm{SOR}} &\approx 160 \\ \operatorname{cond}_2(A) &\approx 5.000 & T_G &\approx 20.000, \quad T_{\mathrm{CG}} \approx 340 \end{aligned}$$

For the comparison of the various solution methods, we also have to take into account the work in each iteration step. For the number "OP" of "a. op." (1 multiplication + 1 addition) per

iteration step there holds:

$$\mathrm{OP_J} \approx \mathrm{OP_{H_1}} \approx \mathrm{OP_{H_\omega}} \approx 6\,\mathrm{n}\,,$$
$$\mathrm{OP_G} \approx \mathrm{OP_{CG}} \approx 10\,\mathrm{n}\,.$$

As final result, we see that the computation of the approximate solution of the boundary value model problem (3.4.104) with a prescribed accuracy TOL by the Jacobi method, the Gauß-Seidel method and the gradient method requires $\mathcal{O}(n^2)$ a. op. In this case a direct method such as the Cholesky algorithm requires $\mathcal{O}(n^2) = \mathcal{O}(m^2 n)$ a. op. but significantly more storage space. The (optimal) SOR method and the CG method only require $\mathcal{O}(n^{3/2})$ a. op.

For the model problem with $n = 10^4$, we have the following total work "TW" required for the solution of the system (3.4.105) to discreetization accuracy $\varepsilon = 10^{-4}$:

$$\begin{aligned}
\mathrm{TW}_J(\mathrm{TOL}) &\approx 4 \cdot 3n^2 \approx 1,2 \cdot 10^9\,\text{a. op.}\,,\\
\mathrm{TW}_{GS}(\mathrm{TOL}) &\approx 4 \cdot 1,5n^2 \approx 6 \cdot 10^8\,\text{a. op.}\,,\\
\mathrm{TW}_{SOR}(\mathrm{TOL}) &\approx 4 \cdot 2n^{3/2} \approx 8 \cdot 10^6\,\text{a. op.}\,,\\
\mathrm{TW}_{CG}(\mathrm{TOL}) &\approx 4 \cdot 10n^{3/2} \approx 4 \cdot 10^7\,\text{a. op.}\,.
\end{aligned}$$

Remark 3.7: Using an appropriate preconditioning, e. g., the ILU preconditioning, in the CG method the work count can be reduced to $\mathcal{O}(n^{5/4})$. The same complexity can be achieved by Chebyshev acceleration of the (optimal) SOR method. Later, we will discuss a more sophisticated iterative method based on the "multi-level concept", which has optimal solution complexity $\mathcal{O}(n)$. For such a multigrid ("MG") method, we can expect work counts like $\mathrm{TW}_{MG} \approx 4 \cdot 25n \approx 10^6$ a. op..

Remark 3.8: For the 3-dimensional version of the above model problem, we have

$$\lambda_{max} \approx 12h^{-2}, \quad \lambda_{min} \approx 3\pi^2, \quad \kappa \approx \frac{8}{3\pi^2 h^2},$$

and consequently the same estimates for ρ_J, ρ_{GS} and ρ_{SOR} as well as for the iteration numbers T_J, T_{GS}, T_{SOR}, T_{CG}, as in the 2-dimensional case. In this case the total work per iteration step is $\mathrm{OP_J}$, $\mathrm{OP_{GS}}$, $\mathrm{OP_{SOR}} \approx 8\,\mathrm{N}$, $\mathrm{OP_{CG}} \approx 12\,\mathrm{N}$. Hence, the resulting total work amounts to

$$\begin{aligned}
\mathrm{TW}_J(\mathrm{TOL}) &\approx 4 \cdot 4n^2 \approx 1,6 \cdot 10^{13}\,\text{a. op.}\,,\\
\mathrm{TW}_{GS}(\mathrm{TOL}) &\approx 4 \cdot 2n^2 \approx 8 \cdot 10^{12}\,\text{a. op.}\,,\\
\mathrm{TW}_{SOR}(\mathrm{TOL}) &\approx 4 \cdot 3n^{3/2} \approx 1,2 \cdot 10^{10}\,\text{a. op.}\,,\\
\mathrm{TW}_{CG}(\mathrm{TOL}) &\approx 4 \cdot 12n^{3/2} \approx 4,8 \cdot 10^{10}\,\text{a. op.},
\end{aligned}$$

while that for the multigrid method increases only insignificantly to $\mathrm{TW}_{MG} \approx 4 \cdot 50n \approx 2 \cdot 10^8$ a. op..

Remark 3.9: For the interpretation of the above work counts, we have to consider the computing power of present computer cores, e. g., 200 MFlops (200 millionen "floating-point" operationen per Sekunde) of a standard desktop computer. Here, the solution of the 3-dimensional model problem by the optimal SOR method takes about $1,5$ minutes while the multigrid method only needs 1 second.

3.5 Exercises

Exercise 3.1: Investigate the convergence of the fixed-point iteration $x^t = Bx^{t-1} + c$ with an arbitrary starting value $x^0 \in \mathbb{R}^3$ for the following matrices

$$(i) \quad B = \begin{bmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.7 \\ 0.1 & 0.1 & 0.1 \end{bmatrix}, \qquad (ii) \quad B = \begin{bmatrix} 0 & 0.5 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

What are the limits of the iterates in case of convergence? (Hint: The eigenvalues of the matrices B are to be estimated. This can be done via appropriate matrix norms or also via the determinants.

Exercise 3.2: The linear system

$$\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

is to be solved by the Jacobi and the Gauß-Seidel method. How many iterations are approximately (asymptotically) required for reducing the initial error $\|x^0 - x\|_2$ by the factor 10^{-6}? (Hint: Use the error estimate presented in class.)

Exercise 3.3: Show that the two definitions of "irreducibility" of a matrix $A \in \mathbb{R}^{n\times n}$ given in class are equivalent.

Hint: Use the fact that the definition of "reducibility" of the system $Ax = b$, i. e., the existence of simultaneous row and column permutations resulting in

$$PAP^T = \tilde{A} = \begin{bmatrix} \tilde{A}_{11} & 0 \\ 0 & \tilde{A}_{22} \end{bmatrix}, \quad \tilde{A}_{11} \in \mathbb{R}^{p\times p}, \quad \tilde{A}_{22} \in \mathbb{R}^{q\times q}, \quad n = p + q,$$

is equivalent to the existence of a non-trivial index partitioning $\{J, K\}$ of $N_n = \{1, \ldots, n\}$, $J \cup K = N_n$, $J \cap K = \emptyset$, such that $a_{jk} = 0$ for $j \in J$, $k \in K$.

Exercise 3.4: For the computation of the inverse A^{-1} of a regular matrix $A \in \mathbb{R}^{n\times n}$ the following two fixed-point iterations are considered:

$a)$ $X_t = X_{t-1}(I - AC) + C, \quad t = 1, 2, \ldots, \quad C \in \mathbb{R}^{n\times n}$ a regular "preconditioner",
$b)$ $X_t = X_{t-1}(2I - AX_{t-1}), \quad t = 1, 2, \ldots$.

Give criteria for the convergence of these iterations. How would for this task (computation of a matrix inverse) the Newton iteration look like?

Exercise 3.5: Let B be an arbitrary $n \times n$-matrix, and let p by a polynomial. Show that

$$\sigma(p(B)) = p(\sigma(B)),$$

i. e., for any $\lambda \in \sigma(p(B))$ there exists a $\mu \in \sigma(B)$ such that $\lambda = p(\mu)$ and vice versa. (Hint: Recall the Schur or the Jordan normal form.)

Exercise 3.6: The method of Chebyshev acceleration can be applied to any convergent fixed-point iteration

$$x^t = Bx^{t-1} + c, \quad t = 1, 2, \ldots,$$

with symmetric iteration matrix B. Here, the symmetry of B guarantees the relation $\|p(B)\|_2 = \mathrm{spr}(p(B)) = \max_{\lambda\in\sigma(B)} |p(\lambda|$ for any polynomial $p \in P_k$, which is crucial for the analysis of the acceleration effect. In class this has been carried out for the SSOR (**Symmetric** Successive Over-Relaxation) method. Repeat the steps of this analysis for the Jacobi method for solving the linear system $Ax = b$ with symmetric matrix $A \in \mathbb{R}^{n\times n}$.

Exercise 3.7: Consider the following symmetric "saddle point system"

$$\begin{bmatrix} A & B \\ B^T & O \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix},$$

with a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a not necessarily quadratic matrix $B \in \mathbb{R}^{m\times n}$, $m \le n$. The coefficient matrix cannot be positive definite since some of its main diagonal elements are zero. Most of the iterative methods discussed in class can directly be applied for this system.

(i) Can the damped Richardson method,

$$\begin{bmatrix} x^t \\ y^t \end{bmatrix} = \left(\begin{bmatrix} I & O \\ O & I \end{bmatrix} - \theta \begin{bmatrix} A & B \\ B^T & O \end{bmatrix} \right) \begin{bmatrix} x^{t-1} \\ y^{t-1} \end{bmatrix} + \theta \begin{bmatrix} b \\ c \end{bmatrix},$$

be made convergent in this case for appropriately chosen damping parameter θ? (Hint: Investigate whether the coefficient matrix may have positive AND negative eigenvalues.)

(ii) A classical approach to solving this saddle-point system is based on the equivalent "Schur-complement formulation":

$$B^T A^{-1} B y = B^T A^{-1} b - c, \quad x = b - A^{-1} B y,$$

in which the solution component y can be computed independently of x. The matrix $B^T A^{-1} B$ is called the "Schur complement" of A in in the full block matrix. Show that the matrix $B^T A^{-1} B$ is symmetric and positive semi-definite and even positive definite if B has maximal rank. Hence the symmetrized Gauß-Seidel method with Chebyshev acceleration may be applied to this reduced system for y. Formulate this iteration in a most efficient way. What are the most work intensive steps?

Exercise 3.8: The general "descent method" for the iterative solution of a linear system $Ax = b$ with symmetric positive definite matrix $A \in \mathbb{R}^{N\times N}$ has the form

$$\begin{aligned} \text{starting value:} \quad & x^0 \in \mathbb{R}^n, \quad g^0 := Ax^0 - b\,, \\ \text{for } t \ge 0: \quad & \text{descent direction} \quad r^t\,, \\ & \alpha_t = -\frac{(g^t, r^t)_2}{(Ar^t, r^t)_2}, \quad x^{t+1} = x^t + \alpha_t r^t, \quad g^{t+1} = g^t - \alpha_t A r^t\,. \end{aligned}$$

The so-called "Coordinate Relaxation" uses descent directions r^t, which are obtained by cycling through the cartesian unit vectors $\{e^1, \ldots, e^n\}$. Show that a full n-cycle of this method is equivalent to one step of the Gauß-Seidel iteration

$$\hat{x}^1 = D^{-1}b - D^{-1}(L\hat{x}^1 + Rx^0).$$

Exercise 3.9: The minimal squared-defect solution of an overdetermined linear system $Ax = b$ is characterized as solution of the normal equation

$$A^T Ax = A^T b.$$

The square matrix $A^T A$ is symmetric and also positive definite, provided A has full rank. Formulate the CG method for solving the normal equation without explicitly computing the matrix product $A^T A$. How many matrix-vector products with A are necessary per iteration (compared to the CG method applied to $Ax = b$)? Relate the convergence speed of this iteration to the singular values of the matrix A.

Exercise 3.10: For solving a linear system $Ax = b$ with symmetric positive definite coefficient matrix A one may use the Gauß-Seidel, the (optimal) SOR method, the gradient method, or the CG methods. Recall the estimates for the asymptotic convergence speed of these iterations expressed in terms of the spectral condition number $\kappa = \mathrm{cond}_2(A)$ and compare the corresponding performance results.

In order to derive convergence estimates for the Gauß-Seidel and (optimal) SOR method, assume that A is consistently ordered and that the spectral radius of the Jacobi iteration matrix is given by

$$\mathrm{spr}\,(J) = 1 - \frac{1}{\kappa}.$$

Discuss the pros and cons of the considered methods.

Exercise 3.11: Consider the following symmetric "saddle point system" from Exercise 7.4

$$\begin{bmatrix} A & B \\ B^T & O \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix},$$

with a symmetric positive definite matrix $A \in \mathbb{R}^{n\times n}$ and a not necessarily quadratic matrix $B \in \mathbb{R}^{n\times m}$, $m \leq n$ with full rank. The coefficient matrix cannot be positive definite since some of its main diagonal elements are zero.

A classical approach of solving this saddle-point system is based on the equivalent "Schur-complement formulation":

$$B^T A^{-1} By = B^T A^{-1} b - c, \quad x = A^{-1}b - A^{-1}By,$$

in which the solution component y can be computed independently of x. The matrix $B^T A^{-1} B$ is called the "Schur complement" of A in the full block matrix.

In Exercise 7.4 it was shown that a symmetric variant of the Gauß-Seidel method with Chebyshev acceleration can be applied to this system. However, this approach suffers from the severe

drawback that $B^T A^{-1} B$ has to be explicitly known in order to construct the decomposition

$$B^T A^{-1} B = L + D + R.$$

Prove that, in contrast, the CG method applied to the Schur complement method does not suffer from this defect, i. e. that an explicit construction of A^{-1} can be avoided. Formulate the CG algorithm for above Schur complement and explain how to efficiently treat the explicit occurrence of A^{-1} in the algorithm.

Exercise 3.12: For the gradient method and the CG method for a symmetric, positive definite matrix A there hold the error estimates

$$\|x^t_{\text{grad}} - x\|_A \leq \left(\tfrac{1-1/\kappa}{1+1/\kappa}\right)^t \|x^0_{\text{grad}} - x\|_A,$$

$$\|x^t_{\text{cg}} - x\|_A \leq 2\left(\tfrac{1-1/\sqrt{\kappa}}{1+1/\sqrt{\kappa}}\right)^t \|x^0_{\text{cg}} - x\|_A,$$

with the condition number $\kappa := \text{cond}_2(A) = \lambda_{\max}/\lambda_{\min}$. Show that for reducing the initial error by a factor ε the following numbers of iteration are required:

$$t_{\text{grad}}(\varepsilon) \approx \tfrac{1}{2}\kappa \ln(1/\varepsilon), \qquad t_{\text{cg}}(\varepsilon) \approx \tfrac{1}{2}\sqrt{\kappa} \ln(2/\varepsilon).$$

Exercise 3.13: The SSOR preconditioning of the CG method for a symmetric, positive definite matrix A with the usual additive decomposition $A = L+D+L^T$ uses the parameter dependent matrix

$$C := \frac{1}{2-\omega}\Big(\frac{1}{\omega}D + L\Big)\Big(\frac{1}{\omega}D\Big)^{-1}\Big(\frac{1}{\omega}D + L^T\Big), \quad \omega \in (0,2).$$

Write this matrix in the form $C = KK^T$ with a regular, lower-triangular matrix K and explain why C^{-1} may be viewed as an approximation to A^{-1}.

Exercise 3.14: In the lecture, we formulated the sequence of iterates $\{x^t\}_{t\geq 1}$ of the CG-method formally as the solution x^t of the optimization problem

$$Q(x^t) = \min_{y \in x^0 + K_t(d^0;A)} Q(y) \quad \leftrightarrow \quad \|Ax^t - b\|_{A^{-1}} = \min_{y \in x^0 + K_t(d^0;A)} \|Ax^t - b\|_{A^{-1}},$$

with the Krylow spaces $K_t(d^0;A) = \text{span}\{d^0, Ad^0, \cdots, A^{t-1}d^0\}$.

The so called "Generalized minimal residual method" (GMRES), instead, formally constructs a sequence of iterates $\{x^t_{\text{gmres}}\}_{t\geq 1}$ by

$$\|Ax^t_{\text{gmres}} - b\|_2 = \min_{y \in x^0 + K_t(d^0;A)} \|Ay - b\|_2.$$

i) Prove that the GMRES method allows for an error inequality similar to the one that was derived for the CG method:

$$\|Ax^t_{\text{gmres}} - b\|_2 \leq \min_{p \in P_t, p(0)=1} \|p(A)\|_2 \, \|Ax^0 - b\|_2,$$

where P_t denotes the space of polynomials up to order t.

ii) Prove that in case of A being a symmetric, positive definite matrix, this leads to the same asymptotic convergence rate as for the CG method.

iii) Show that the result obtained in (i) can also be applied to the case of A being similar to a diagonal matrix $D = \mathrm{diag}_i(\lambda_i) \in \mathbb{C}^{n\times n}$, i. e.

$$A = TDT^{-1},$$

with a regular matrix T: In this case there holds

$$\|x^t_{\mathrm{gmres}} - x\|_2 \leq \kappa_2(T) \min_{p\in P_t, p(0)=1} \max_i |p(\lambda_i)| \, \|x^0 - x\|_2.$$

What makes this result rather cumbersome in contrast to the case of a symmetric, positive matrix discussed in (ii)?

Remark: The advantage of the GMRES method lies in the fact that it is, in principle, applicable to any regular matrix A. However, good convergence estimates for the general case are hard to prove.

Exercise 3.15: Consider the model matrix $A \in \mathbb{R}^{n\times n}$, $n = m^2$, originating from the 5-point discretization of the Poisson problem on the unit square,

$$A = \left.\begin{bmatrix} B & -I & & \\ -I & B & -I & \\ & -I & B & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} n \qquad B = \left.\begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & \ddots \\ & & \ddots & \ddots \end{bmatrix}\right\} m$$

Determine the number of arithmetic operations (a. op. := 1 additions + 1 multiplication or 1 devision) in terms of the dimension n required in each step of (i) the CG method and (ii) the PCG method with SSOR preconditioning.

Exercise 3.16: The (regular) model matrix $A = (a_{ij})_{i,j=1}^n$ in Exercise 9.3 possesses another important property (of "'nonnegative type"' or a regular "'Z-matrix"'):

$$a_{ii} > 0, \quad a_{ij} \leq 0, \;\; i \neq j.$$

Show that the inverse $A^{-1} = (a_{ij}^{(-1)})_{i,j=1}^n$ has nonnegative elements $a_{ij}^{(-1)} \geq 0$, i. e., A is a so-called "'M-matrix"' ("'(inverse) monotone"' matrix). This implies that the solution x of a linear system $Ax = b$ with nonnegative right-hand side b, $b_i \geq 0$, is also nonnegative $x_i \geq 0$. (Hint: consider the Jacobi matrix $J = -D^{-1}(L + R)$ and the representation of the inverse $(I - J)^{-1}$ as a Neumann series.)

Exercise 3.17: Repeat the analysis of the convergence properties of the various solution methods for the 3-dimensional version of the model problem considered in class. The underlying

boundary value problem has the form

$$-\Big(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\Big)u(x,y,z)=f(x,y,z),\quad (x,y,z)\in\Omega=(0,1)^3\in\mathbb{R}^3,$$
$$u(x,y,z)=0,\quad (x,y,z)\in\partial\Omega,$$

and the corresponding difference approximation (so-called "7-point approximation") at interior mesh points $(x,y,z)\in\{P_{ijk},\, i,j,k=1,\ldots,m\}$, reads

$$-h^{-2}\big(U(x\pm h,y,z)+U(x,y\pm h,z)+U(x,y,z\pm h)-6U(x,y,z)\big)=f(x,y,z).$$

Using again row-wise numbering of the mesh points the resulting linear system for the mesh values $U_{ijk}\approx u(P_{ijk})$ takes the form

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

In this case the corresponding eigenvalues and eigenvectors are explicitly given by

$$\lambda_{ijk}=6-2\,\big(\cos[ih\pi]+\cos[jh\pi]+\cos[kh\pi]\big),\quad i,j,k=0,\ldots,m,$$
$$w^{ijk}=\big(\sin[pih\pi]\,\sin[qjh\pi]\,\sin[rkh\pi]\big)_{p,q,r=1}^{m}.$$

For the exact solution $u(x,y,z)=\sin(\pi x)\sin(\pi y)\sin(\pi z)$ there holds the error estimate

$$\max_{\Omega}|U_{ijk})-u(P_{ijk})|\le\frac{\pi^4}{8}\,h^2+O(h^4),$$

which induces the mesh size $h=10^{-2}$ in order to guarantee a desired relative discretization accuracy of $TOL=10^{-3}$.

a) Determine formulas for the condition number $\mathrm{cond}_2(A)$ and the spectral radius $\mathrm{spr}(J)$ in terms of the mesh size h.

b) Give the number of iterations of the Jacobi, Gauß-Seidel and optimal SOR method as well as the gradient and CG method approximately needed for reducing the initial error to size $\varepsilon=10^{-4}$ (including a safety factor of $1/10$).

c) Determine the cost, i. e., number of a. op., of each iteration step of the considered methods and compare the total efficiency of these methods.

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: *Die "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

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

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 nth 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 assume that the eigenvalue with largest modulus is 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*

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

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

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

[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.*

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 convergence 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

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

Here, the convergence of the eigenvalue approximations is twice as fast as in the nonhermitian 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.

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

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

4.1.2 The "Inverse Iteration"

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

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

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

[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.

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 assume 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

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

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

$$Az = \lambda z, \quad \lambda = h^2 \mu, \tag{4.1.19}$$

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 (5.2.67). 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

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

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 [27]), 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. [26] and Sleijpen & Van der Vorst [45]

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 arithmetical 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

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

[3]Heinz Rutishauser (1918-1970): Swiss mathematician and computer scientist; since 1962 prof. at the 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.

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

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

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 the Householder method. 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

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

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

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

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

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 $\Lambda = \mathrm{diag}(\lambda_i)$. 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

$$\begin{aligned} \underbrace{[Q^{(1)}\dots Q^{(t)}]}_{=P^{(t)}}\,\underbrace{[R^{(t)}\dots R^{(1)}]}_{=:S^{(t)}} &= \underbrace{[Q^{(1)}\dots Q^{(t-1)}]}_{=P^{(t-1)}}\,A^{(t)}\,\underbrace{[R^{(t-1)}\dots R^{(1)}]}_{=:S^{(t-1)}} \\ &= P^{(t-1)}\,[P^{(t-1)T}\,A\,P^{(t-1)}]S^{(t-1)} = A\,P^{(t-1)}S^{(t-1)}, \end{aligned}$$

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

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

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 Section3.4). For a more detailed discussion, we refer to the literature, e. g., Deuflhard & Hohmann [30], Stör & Bulirsch[47], Golub & Loan[33], and Parlett [41].

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),

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

(see Wang & Gragg [53]).

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 $2m + 1 \ll n = m^2$ (s. 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 of Golub & van Loan [33]. 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 $k \le n$. 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[51]). 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 dis-

[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", an eigenvalue algorithm used in numerical linear algebra, 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; Ph.D. thesis (1921) on relativity theory; assistant to Albert Einstein during the period of 192829; contributions to exact solutions of the Einstein field equation; discovery of the fast Fourier transform (FFT, 1940) but credited to him; worked in Washington DC at the U.S. National Bureau of Standards after 1949; invented the 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 an outstanding physics teacher and author of many classical text books.

cretization 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 $K_m = \mathrm{span}\{q^1, \ldots, q^m\}$ 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

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

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 $q^i = (q^i_j)_{j=1}^n$ the Galerkin eigenvalue problem (4.3.37) is written in the form

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

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 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 a 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 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 $\{q^1, \ldots, q^{t-1}\}$:

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

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

$$(q^t, \tilde{q}^t)_2 = \|\tilde{q}^t\|_2 = \big(\tilde{q}^t, Aq^{t-1} - \textstyle\sum_{j=1}^{t-1} \mathrm{proj}_{q^j}(Aq^{t-1})\big)_2 = (\tilde{q}^t, Aq^{t-1})_2.$$

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

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

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

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

This algorithm gives the same result as the original formula (4.3.43) in exact arithmetic 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:*

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

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) there holds ("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

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

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 $H^{(m)} = \bar{Q}^{(m)T}AQ^{(m)}$ 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 $\|q^{m+1}\|_2 = 1$,

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

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 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 arithmetical 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 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 [23].

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 proved 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 pseudospectrum

As an application of the Krylov space methods described so far, we discuss the computation of the pseudospectrum 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 [19], Trefethen [18], and Gerecht et al. [32]. The following lemma collects some useful facts on the pseudospectra of matrices.

Lemma 4.4: *(i) The* ε*-pseudospectrum 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*-pseudospectrum* $\sigma_\varepsilon(T)$ *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-pseudospectrum $\sigma_\varepsilon(T)$ of a matrix $T \in \mathbb{C}^{n\times n}$, which can be taken as starting point for the computation of pseudospectra (see Trefethen [18] and Trefethen & Embree [19]). 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 pseudospectra". This approach is rather expensive since in order to obtain precise information of the ε-pseudospectrum a really large number of random matrices are needed. It cannot be used for higher-dimensional matrices.

Remark 4.11: The determination of pseudospectra 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

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

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 comping 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 $K_m = \operatorname{span}\{q, Tq, .., T^{m-1}q\}$. 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 pseudospectrum 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,

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

For the reliable approximation of the pseudospectrum 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-pseudospectrum 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 pseudospectra. The pseudospectra of $H^{(m)}$ may then be computed using the approach described in Section 4.2.2. By Lemma 1.17 the pseudospectrum 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 pseudospectra 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 pseudospectrum. Let $D := \{z \in \mathbb{C} |\, \{\mathrm{Re}\, z, z\} \in [a_r, b_r] \times [a_i, b_i]\}$ for certain values $a_r < b_r$ and $a_i < b_i$. To determine the pseudospectrum 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-pseudospectrum.

(i) Computation of the pseudospectra $\sigma_\varepsilon(H^{(m)-1})$: 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 pseudospectra $\sigma_\varepsilon(H^{(m)})$: 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 $\Sigma = \mathrm{diag}\{\sigma_i,\, i = 1, \ldots, n\}$. 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 $\mathcal{O}(m^2)$, we limit the dimension of the Krylov space by $m \leq 200$.

Choice of parameters and accuracy issues

The described algorithm for computing the pseudospectrum 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 pseudospectrum of the differential operator A. First, h is refined and m is increased until no significant change in the boundaries of the ε-pseudospectrum 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 [18])

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

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 pseudospectra, 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-pseudospectra are computed for $\varepsilon = 10^{-1}, 10^{-2}, ..., 10^{-10}$ leading to the results shown in Figure 4.1. We observe that all eigenvalues have negative real part but also that the corresponding pseudospectra 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 pseudospectrum of the matrix $H_h^{(m)}$.

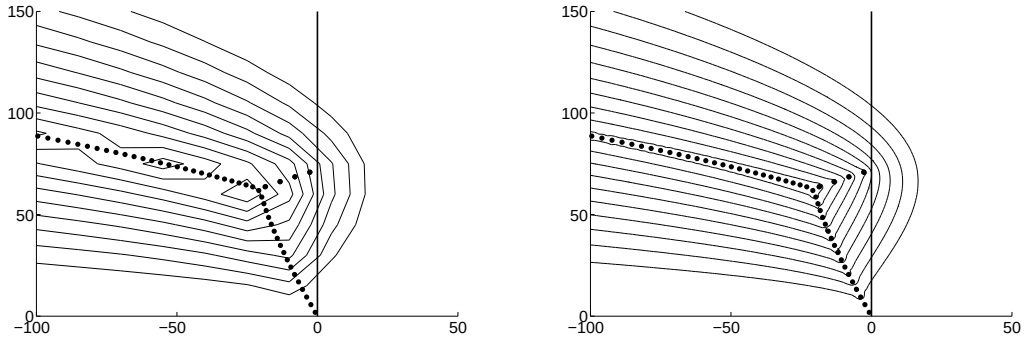


Figure 4.1: *Approximate eigenvalues and pseudospectra of the operator A computed from those of the inverse matrix A_h^{-1} on a 10×10 grid (left) and on a 100×100 grid (right): dots represent eigenvalues and the lines the boundaries of the ε-pseudospectra for $\varepsilon = 10^{-1}, ..., 10^{-10}$.*

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 $v = (v_1, v_2)$:

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

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.

For testing the accuracy of the proposed method, we compare the quality of the pseudospectra 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 Figure 4.2 and Figure 4.3 indicate that the choice $h = 2^{-7}$ and $m = 100$ is sufficient for the present example.

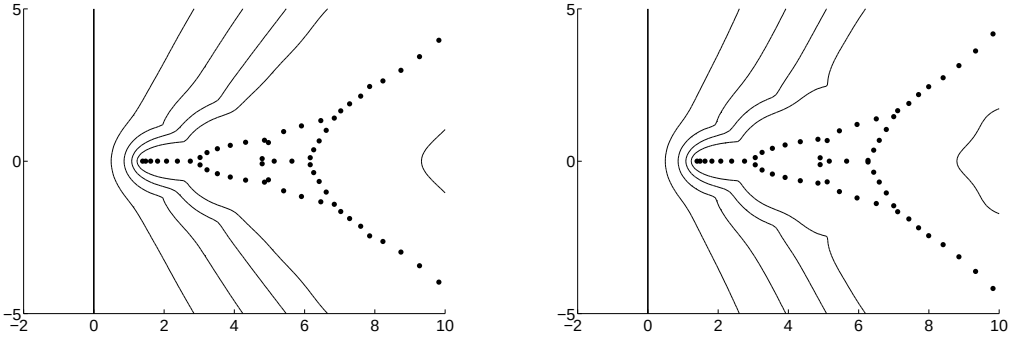


Figure 4.2: *Computed pseudospectra of the linearized Burgers operator with Dirichlet inflow condition for* $\nu = 0.01$ *and* $h = 2^{-7}$ *(left) and* $h = 2^{-8}$ *(right) computed by the Arnoldi method with* $m = 100$. *The dots represent eigenvalues and the lines the boundaries of the* ε*-pseudospectra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

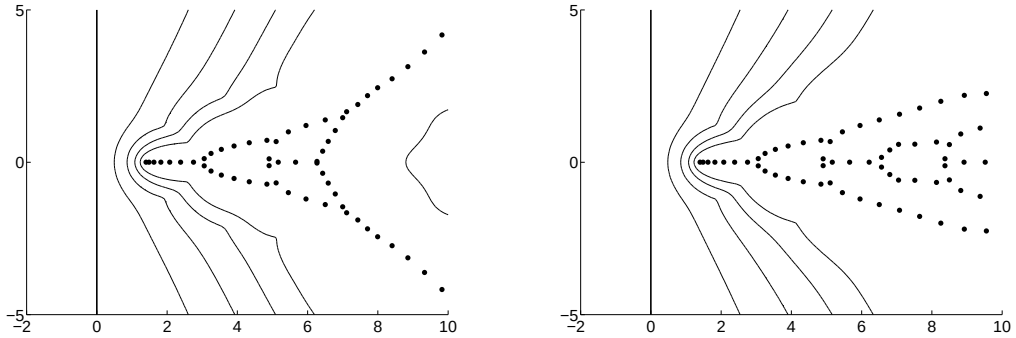


Figure 4.3: *Computed pseudospectra of the linearized Burgers operator with Dirichlet inflow condition for* $\nu = 0.01$ *and* $h = 2^{-8}$ *computed by the Arnoldi method with* $m = 100$ *(left) and* $m = 200$ *(right). The dots represent eigenvalues and the lines the boundaries of the* ε*-pseudospectra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

Now, we turn to Neumann inflow conditions. In this particular case the first eigenvalues 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 pseudospectra on the size of the viscosity, $0.001 \leq \nu \leq 0.01$. 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 $\mathrm{Re}\lambda \leq 6$ and corresponding ε-pseudospectra with $\varepsilon \geq 10^{-4}$ are reliably computed. The results are shown in Figure 4.4. 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 Figure 4.4, we see that the 0.1-pseudospectrum reaches into the negative complex half-plane indicating instability for such perturbations. This effect is even more pronounced for $\nu = 0.001$ with $\lambda^N_{\mathrm{crit}} \approx 0.0098$.

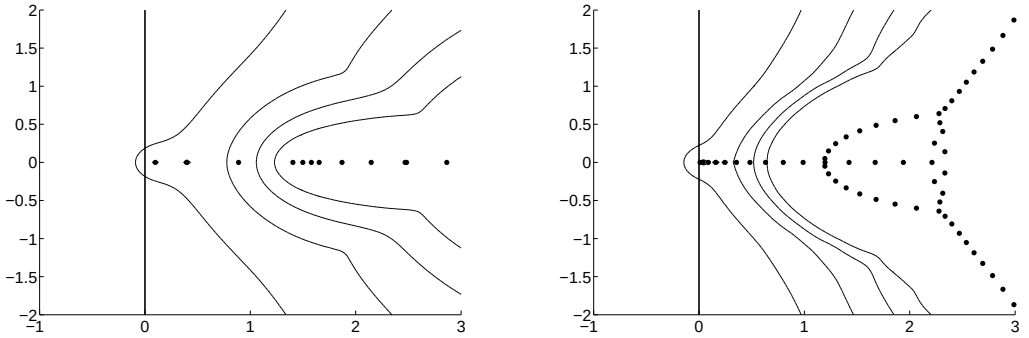


Figure 4.4: *Computed pseudospectra of the linearized (around Couette flow) Burger operator with Neumann inflow conditions for* $\nu = 0.01$ *(left) and* $\nu = 0.001$ *(right): The dots represent eigenvalues and the lines the boundaries of the* ε*-pseudospectra for* $\varepsilon = 10^{-1}, \ldots, 10^{-4}$.

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 [62]). The geometry data are as follows: channel domain $\Omega := (0.00\text{m}, 2.2\text{m}) \times (0.00\text{m}, 0.41\text{m})$, diameter of circle $D := 0.10\text{m}$, center of circle at $a := (0.20\text{m}, 0.20\text{m})$ (slightly nonsymmetric position). The Reynolds number is defined in terms of the diameter D and the maximum inflow velocity $\bar{U} = \max|v^{\text{in}}| = 0.3\text{m/s}$ (parabolic profile), $\text{Re} = \bar{U}^2 D/\nu$. The boundary conditions are $v|_{\Gamma_{\text{rigid}}} = 0$, $v|_{\Gamma_{\text{in}}} = v^{\text{in}}$, $\nu\partial_n v - np|_{\Gamma_{\text{out}}} = 0$. The viscosity is chosen such that the Reynolds number is small enough, $20 \le \text{Re} \le 40$, to guarantee stationarity of the base flow as shown in Figure 4.5. Already for $\text{Re} = 60$ the flow is nonstationary (time periodic).

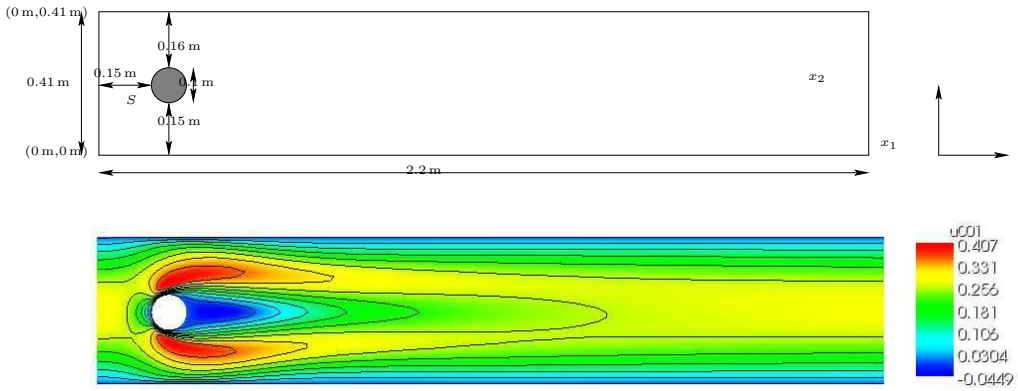


Figure 4.5: Configuration of the "channel flow" benchmark and x_1-component of the velocity for $\text{Re} = 40$.

We want to investigate the stability of the computed base flow for several Reynolds numbers in the range $20 \le \text{Re} \le 60$ and inflow conditions imposed on the admissible perturbations, Dirichlet or Neumann ("free"), by determining the corresponding critical eigenvalues and pseudospectrum. 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 pseudospectra of the critical eigenvalues for $\text{Re} = 40$ and $\text{Re} = 60$ are shown in Figure 4.3.2. 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 $\text{Re} = 40$ the relevant 10^{-2}-pseudospectrum 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 $40 \le \text{Re} \le 60$. 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 pseudospectrum.

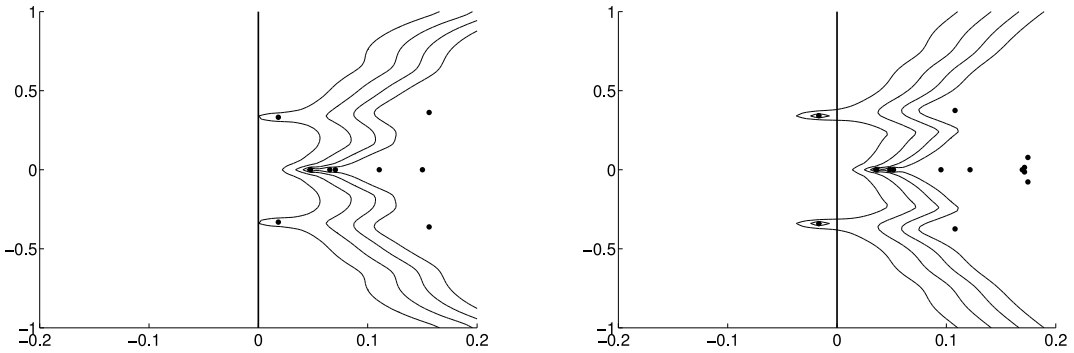


Figure 4.6: Computed pseudospectra of the linearized Navier-Stokes operator ("channel flow" benchmark) for different Reynolds numbers, $\mathrm{Re} = 40$ (left) and $\mathrm{Re} = 60$ (right), with **Dirichlet** inflow condition: The dots represent eigenvalues and the lines the boundaries of ε-pseudospectra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

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 pseudospectrum. The results are shown in Figure 4.3.2. Though, for $\mathrm{Re} = 40$ the real part of the most critical (positive) eigenvalue is rather small, the corresponding 10^{-2}-pseudospectrum reaches only a little into the negative complex half-plane.

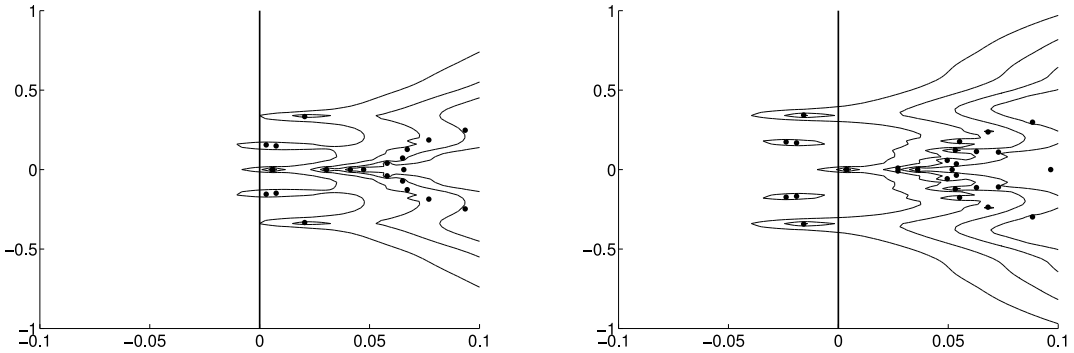


Figure 4.7: Computed pseudospectra of the linearized Navier-Stokes operator ("channel flow") with **Neumann** inflow conditions for different Reynolds numbers, $\mathrm{Re} = 40$ (left) and $\mathrm{Re} = 60$ (right): The dots represent eigenvalues and the lines the boundaries of the ε-pseudospectra 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 Exercise 10.2.

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 $K_m := \text{span}\{q, Aq, \ldots, A^{m-1}q\}$. 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 class, 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: Formulate the "Inverse Iteration" of Wielandt and the "Lanczos algorithm" (combined with the QR method) for computing the smallest eigenvalue of a symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$. Compare the arithmetic work (# of a. op.) of these two approaches for performing 100 iterations in both cases. How do the two methods compare if not only the smallest but the *ten* smallest eigenvalues are to be computed?

Exercise 4.9: Consider the model eigenvalue problem of Exercise 10.1, 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

$$\lambda_{ijk}^h = h^{-2}\big\{6 - 2\big(\cos[ih\pi] + \cos[jh\pi] + \cos[kh\pi]\big)\big\}, \quad i,j,k = 0, \ldots, m,$$
$$w_h^{ijk} = \big(\sin[pih\pi]\, \sin[qjh\pi]\, \sin[rkh\pi]\big)_{p,q,r=1}^m.$$

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

$$|\lambda_{ijk} - \lambda_{ijk}^h| \le c \lambda_{ijk} h^2,$$

where $\lambda_{ijk} = (i^2 + j^2 + k^2)\pi^2$ are the exact eigenvalues of the Laplace operator and $c \approx 1$.

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

(ii) How many of the eigenvalues of the Laplace operator can be approximated reliably if a uniform accuracy of $TOL = 10^{-4}$ is required?

(iii) How small has the mesh size h to be chosen if the first 1.000 eigenvalues of the Laplace operator have to be computed with absolute accuracy $TOL = 10^{-2}$? How large would the dimension n of the resulting system matrix A be in this case?

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

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

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

(ii) Formulate 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-pseudospectrum $\sigma_\varepsilon(T)$ 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

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

5 Multigrid Methods

Multigrid methods belong to the class of preconditioned defect correction methods, in which the preconditioning uses a hierarchy of problems of similar structure but decreasing dimension. They are particularly designed for the solution of the linear systems resulting from the discretization of partial differential equations by grid methods such as finite difference or finite element schemes. But special versions of this method can also be applied to other types of problems not necessarily originating from differential equations. Its main concept is based on the existence of a superposed "continuous" problem of infinite dimension, from which all the smaller problems are obtained in a nested way by some projection process (e. g., a "finite difference discretization" or a "Galerkin method"). On the largest subspace (on the finest grid) the errors and the corresponding defects are decomposed into "high-frequency" and "low-frequency" components, which are treated separately by simple fixed-point iterations for "smoothing" out the formers and by correcting the latters on "coarser" subspaces (the "preconditioning" or "coarse-space correction"). This "smoothing" and "coarse-space correcting" is applied recursively on the sequence of nested spaces leading to the full "multigrid algorithm". By an appropriate combination of all these algorithmical components one obtains an "optimal" solution algorithm, which solves a linear system of dimension n, such as the model problem considered above, in $\mathcal{O}(n)$ operations. In the following, for notational simplicity, we will decribe and analyze the multigrid method within the framework of a low-order finite element Galerkin discretization of the model problem of Section 3.4. In fact, on uniform cartesian meshes this discretization is almost equivalent to the special finite difference scheme considered in Section 3.4. For the details of such a finite element scheme and its error analysis, we refer to the literature, e. g., Rannacher [3].

5.1 Multigrid methods for linear systems

For illustration, we consider the linear system

$$A_h x_h = b_h, \tag{5.1.1}$$

representing the discretization of the model problem of Section 3.4 on a finite difference mesh $\mathbb{T}_h$ with mesh size $h \approx m^{-1}$ and dimension $n = m^2 \approx h^{-4}$. Here and below, the quantities related to a fixed subspace (corresponding to a mesh $\mathbb{T}_h$) are labeled by the subscript h.

The solution of problem (5.1.1) is approximated by the damped Richardson iteration

$$x_h^{t+1} = x_h^t + \theta_h(b_h - A_h x_h^t) = (I_h - \theta_h A_h)x_h^t + \theta_h b_h, \tag{5.1.2}$$

with a damping parameter $0 < \theta_h \leq 1$. The symmetric, positive definite matrix A_h possesses an ONS of eigenvectors $\{w_h^i,\, i = 1, ..., n_h\}$ corresponding to the ordered eigenvalue $\lambda_{\min}(A_h) = \lambda_1 \leq ... \leq \lambda_n = \lambda_{\max}(A_h) =: \Lambda_h$. The expansion of the initial error

$$e_h^0 := x_h^0 - x_h = \sum_{i=1}^{n_h} \varepsilon_i w_h^i,$$

induces the corresponding expansion of the iterated errors

$$e_h^t = (I_h - \theta_h A_h)^t e_h^0 = \sum_{i=1}^{n_h} \varepsilon_i (I_h - \theta_h A_h)^t w_h^i = \sum_{i=1}^{n_h} \varepsilon_i (1 - \theta_h \lambda_i)^t w_h^i.$$

Consequently,

$$|e_h^t|^2 = \sum_{i=1}^{n_h} \varepsilon_i^2 (1 - \theta_h \lambda_i)^{2t}. \tag{5.1.3}$$

The assumption $0 < \theta_h \leq \Lambda_h^{-1}$ is sufficient for the convergence of the Richardson iteration. Because of $|1 - \theta_h \lambda_i| \ll 1$ for larger λ_i and $|1 - \theta_h \lambda_1| \approx 1$ "high-frequency" components of the error decay fast, but "low-frequency" components only very slowly. The same holds for the residuum $r_h^t = b_h - A_h x_h^t = A_h e_h^t$, i. e., already after a few iterations there holds

$$|r_h^t|^2 \approx \sum_{i=1}^{[N/2]} \varepsilon_i^2 \lambda_i^2 (1 - \theta_h \lambda_i)^{2t}, \qquad [n/2] := \max\{m \in \mathbf{N} | \, m \leq n/2\}. \tag{5.1.4}$$

This may be interpreted as follows: The iterated defect r_h^t on the mesh $\mathbb{T}_h$ is "smooth". Hence, it can be approximated well on the coarser mesh $\mathbb{T}_{2h}$ with mesh size $2h$. The resulting defect equation for the computation of the correction to the approximation x_h^t on $\mathbb{T}_h$ would be less costly because of its smaller dimension $n_{2h} \approx n_h/4$.

This defect correction process in combination with recursive coarsening can be carried on to a coarsest mesh, on which the defect equation can finally be solved exactly. The most important components of this multigrid process are the "smoothing iteration", $x_h^\nu = S_h^\nu(x_h^0)$ and certain transfer operations between functions defined on different meshes. The smoothing operation $S_h(\cdot)$ is usually given in form of a fixed-point iteration (e. g., the Richardson iteration)

$$x_h^{\nu+1} = S_h(x_h^\nu) := (I_h - C_h^{-1} A_h) x_h^\nu + C_h^{-1} b_h,$$

with the iteration matrix $S_h := I_h - C_h^{-1} A_h$.

5.1.1 Multigrid methods in the finite element context

For the formulation of the multigrid process, we consider a sequence of nested grids $\mathbb{T}_l = \mathbb{T}_{h_l}$, $l = 0, ..., L$, of increasing fineness $h_0 > ... > h_l > ... > h_L$ (for instance obtained by successively refining a coarse starting grid) and corresponding finite element spaces $V_l := V_{h_l}$ of increasing dimension n_l, which are subspaces of the "continuous" solution space $V = H_0^1(\Omega)$ (first-order Sobolev space on Ω including zero Dirichlet boundary conditions). Here, we think of spaces of continuous, with respect to the mesh $\mathbb{T}_h$ piecewise *linear* (on triangular meshes) or piecewise (isoparametrical) *bilinear* (on quadrilateral meshes) functions. For simplicity, we assume that the finite element spaces are hierachically ordered,

$$V_0 \subset V_1 \subset ... \subset V_l \subset ... \subset V_L \,. \tag{5.1.5}$$

This structural assumption eases the analysis of the multigrid process but is not essential for its practical success.

The finite element Galerkin scheme

As usual, we write the continuous problem and its corresponding finite element Galerkin approximation in compact variational form

$$a(u,\varphi) = (f,\varphi)_{L^2} \quad \forall \varphi \in V, \tag{5.1.6}$$

and, analogously on the mesh $\mathbb{T}_h$

$$a(u_h,\varphi_h) = (f,\varphi_h)_{L^2} \quad \forall \varphi_h \in V_h. \tag{5.1.7}$$

Here, $a(u,\varphi) := (Lu,\varphi)_{L^2}$ is the "energy bilinear form" corresponding to the (elliptic) differential operator L and $(f,\varphi)_{L^2}$ the L^2-scalar product on the solution domain Ω. In the model problem discussed above this notation has the explicit form $Lu = -\Delta u$ and

$$a(u,\varphi) = \int_\Omega \nabla u(x)\nabla\varphi(x)\,dx, \quad (f,\varphi)_{L^2} = \int_\Omega f(x)\varphi(x)\,dx.$$

The finite element subspace $V_h \subset V$ has a natural so-called "nodal basis" (Lagrange basis) $\{b^1,\ldots,b^{n_h}\}$ characterized by the interpolation property $b^i(a_j) = \delta_{ij}$, $i,j = 1,\ldots,n_h$, where a_j are the nodal points of the mesh $\mathbb{T}_h$. Between the finite element function $u_h \in V_h$ and the corresponding nodal-value vector $x_h \in \mathbb{R}^{n_h}$, we have the connection $u_h(a_j) = x_{h,j}$, $j = 1,...,n_h$,

$$u_h = \sum_{j=1}^{n_h} x_{h,j} b^j = \sum_{j=1}^{n_h} u_h(a_j) b^j.$$

Using this notation the discrete problems (5.1.7) can be written in the form

$$\sum_{j=1}^{n_h} x_{h,j} a(b^j,b^i) = (f,b^i)_{L^2}, \quad i = 1,\ldots,n_h,$$

which is equivalent to the linear system

$$A_h x_h = b_h, \tag{5.1.8}$$

with the "system matrix" ("stiffness matrix") $A_h = (a_{ij})_{i,j=1}^{n_h} \in \mathbb{R}^{n_h\times n_h}$ and "load vector" $b_h = (b_j)_{j=1}^{n_h} \in \mathbb{R}^{n_h}$ defined by

$$a_{ij} := a(b^j,b^i), \quad b_j := (f,b^j)_{L^2}, \quad i,j = 1,\ldots,n_h.$$

For finite element functions $u_h = \sum_{i=1}^{n_h} x_{h,i} b^i$ and $v_h = \sum_{i=1}^{n_h} y_{h,i} b^i$ there holds

$$a(u_h,v_h) = (A_h x_h, y_h)_2.$$

The system matrix A_h is symmetric and positive definite by construction and has a condition number of size $\mathrm{cond}_2(A_h) = \mathcal{O}(h^{-2})$. Additionally, we will use the so-called "mass matrix" $M_h = (m_{ij})_{i,j=1}^{n_h}$ defined by

$$m_{ij} := (b^j,b^i)_{L^2}, \quad i,j = 1,\ldots,n_h.$$

For finite element functions $u_h = \sum_{i=1}^{n_h} x_{h,i} b^i$ and $v_h = \sum_{i=1}^{n_h} y_{h,i} b^i$ there holds

$$(u_h, v_h)_{L^2} = (M_h x_h, y_h)_2.$$

The mass matrix M_h is also symmetric and positive definite by construction and has a condition number of size $\mathrm{cond}_2(A_h) = \mathcal{O}(1)$.

For the exact "discrete" solution $u_h \in V_h$ there holds the error estimate

$$\|u - u_h\|_{L^2} \leq c\, h^2 \, \|f\|_{L^2}. \tag{5.1.9}$$

Now, we seek a solution process which produces an approximation $\tilde{u}_h \in V_h$ to u_h satisfying

$$\|u_h - \tilde{u}_h\|_{L^2} \leq c\, h^2 \, \|f\|_{L^2}. \tag{5.1.10}$$

This process is called "complexity-optimal" if the arithmetic work for achieving this accuracy is of size $\mathcal{O}(n_h)$ uniformly with respect to the mesh size h. We will see below that the multigrid method is optimal is this sense if all its components are properly chosen.

The multigrid process

Let $u_L^0 \in V_L$ be an initial guess for the exact discrete solution $u_L \in V_L$ on mesh level L (For example, $u_L^0 = 0$ or $u_L^0 = u_{L-1}$ if such a coarse-grid solution is available.). Then, u_L^0 is "smoothed" ("pre-smoothed") by applying ν steps, e. g., of the damped Richardson iteration starting from $\bar{u}_L^0 := u_L^0$. This reads in variational from as follows:

$$(\bar{u}_L^k, \varphi_L)_{L^2} = (\bar{u}_L^{k-1}, \varphi_L)_{L^2} + \theta_L \big\{ (f, \varphi_L)_{L^2} - a(\bar{u}_L^{k-1}, \varphi_L) \big\} \quad \forall \varphi_L \in V_L, \tag{5.1.11}$$

where $\theta_L = \lambda_{max}(A_h)^{-1}$. For the resulting smoothed approximation $\bar{u}_L^\nu$, we define the "defect" $d_L \in V_L$ (without actually computing it):

$$(d_L, \varphi_L)_{L^2} := (f, \varphi_L)_{L^2} - a(\bar{u}_L^\nu, \varphi_L), \quad \varphi_L \in V_L. \tag{5.1.12}$$

Since $V_{L-1} \subset V_L$, we obtain the "defect equation" on the next coarser mesh $\mathbb{T}_{L-1}$ as

$$a(q_{L-1}, \varphi_{L-1}) = (d_L, \varphi_{L-1})_{L^2} = (f, \varphi_{L-1})_{L^2} - a(\bar{u}_L^\nu, \varphi_{L-1}) \quad \forall \varphi_{L-1} \in V_{L-1}. \tag{5.1.13}$$

The correction $q_{L-1} \in V_{L-1}$ is now computed either exactly (for instance by a direct solver) or only approximately by a defect correction iteration $q_{L-1}^0 \to q_{L-1}^R$ using the sequence of coarser meshes $\mathbb{T}_{L-2}, ..., \mathbb{T}_0$. The result $q_{L-1}^R \in V_{L-1}$ is then interpreted as an element of V_L and used for correcting the preliminary approximation $\bar{u}_L^\nu$:

$$\bar{\bar{u}}_L^0 := \bar{u}_L^\nu + \omega_L q_{L-1}^R. \tag{5.1.14}$$

The correction step may involve a damping parameter $\omega_L \in (0, 1)$ in order to minimize the residual of $\bar{\bar{u}}_L$. This practically very useful trick will not be further discussed here, i. e., in the following, we will mostly set $\omega_L = 1$. The obtained corrected approximation $\bar{\bar{u}}_L$ is again smoothed ("post-smoothing") by applying another μ steps of the damped Richardson iteration

starting from $\bar{\bar{u}}_L^0 := \bar{\bar{u}}_L$:

$$(\bar{\bar{u}}_L^k, \varphi_L)_{L^2} = (\bar{\bar{u}}_L^{k-1}, \varphi_L)_{L^2} + \theta_L\big\{(f, \varphi_L)_{L^2} - a(\bar{\bar{u}}_L^{k-1}, \varphi_L)\big\} \quad \forall \varphi_L \in V_L. \tag{5.1.15}$$

The result is then accepted as the next multigrid iterate, $u_L^1 := \bar{\bar{u}}_L^\mu$, completing one step of the multigrid iteration ("multigrid cycle") on mesh level L. Each such cycle consists of $\nu + \mu$ Richardson smoothing steps (on level L), which each requires the inversion of the mass matrix M_h, and the solution of the correction equation on the next coarser mesh.

Now, we will formulate the multigrid algorithm using a more abstract, functional analytical notation, in order to better understand its structure and to ease its convergence analysis. To the system matrices $A_l = A_{h_l}$ on the sequence of meshes $\mathbb{T}_l$, $l = 0, 1, \ldots, L$, we associate operators $\mathcal{A}_l : V_l \to V_l$ by setting

$$(\mathcal{A}_l v_l, w_l)_{L^2} = a(v_l, w_l) = (A_l y_l, z_l)_2 \quad \forall v_l, w_l \in V_l, \tag{5.1.16}$$

where $v_l = (y_{l,i})_{i=1}^{n_l}$, $w_l = (z_{l,i})_{i=1}^{n_l}$. Further, let $\mathcal{S}_l(\cdot)$ denote the corresponding smoothing operations with (linear) iteration operators (Richardson operator) $\mathcal{S}_l = \mathcal{I}_l - \theta_l \mathcal{A}_l : V_l \to V_l$ where $\mathcal{A}_l$ is the "system operator" defined above and $\mathcal{I}_l$ denotes the identity operator on V_l. Finally, we introduce the operators by which the transfers of functions between consecutive subspaces are accomplished:

$$r_l^{l-1} : V_l \to V_{l-1} \ \text{("restriction")}, \quad p_{l-1}^l : V_{l-1} \to V_l \ \text{("prolongation")}.$$

In the finite element context these operators are naturally chosen as $r_l^{l-1} = \mathcal{P}_{l-1}$, the L^2 projection onto V_{l-1}, and $p_{l-1}^l = id.$, the natural embedding of $V_{l-1} \subset V_l$ into V_l.

Now, using this notation, we reformulate the multigrid process introduced above for solving the linear system on the finest mesh $\mathbb{T}_L$:

$$\mathcal{A}_L u_L = f_L := \mathcal{P}_L f. \tag{5.1.17}$$

Multigrid process: Starting from an initial vector $u_L^0 \in V_L$ iterates u_L^t are constructed by the recursive formula

$$u_L^{(t+1)} = MG(L, u_L^{(t)}, f_L). \tag{5.1.18}$$

Let the t-th multigrid iterate $u_L^{(t)}$ be determined.

Coarse grid solution: For $l = 0$, the operation $MG(0, 0, g_0)$ yields the exact solution of the system $\mathcal{A}_0 v_0 = g_0$ (obtained for instance by a direct method),

$$v_0 = MG(0, \cdot, g_0) = \mathcal{A}_0^{-1} g_0. \tag{5.1.19}$$

Recursion: Let for some $1 \le l \le L$ the system $\mathcal{A}_l v_l = g_l$ to be solved. With parameter values $\nu, \mu \ge 1$ the value

$$MG(l, v_l^0, g_l) := v_l^1 \approx v_l \tag{5.1.20}$$

is recursively definined by the following steps:

(i) Pre-smoothimg:

$$\bar{v}_l := \mathcal{S}_l^{\nu}(v_l^0); \tag{5.1.21}$$

(ii) Defect formation:

$$d_l := g_l - \mathcal{A}_l \bar{v}_l; \tag{5.1.22}$$

(iii) Restriction:

$$\tilde{d}_{l-1} := r_l^{l-1} d_l; \tag{5.1.23}$$

(iv) Defect equation: Starting from $q_{l-1}^0 := 0$ for $1 \le r \le R$ the iteration proceeds as follows:

$$q_{l-1}^r := MG(l-1, q_{l-1}^{r-1}, \tilde{d}_{l-1}); \tag{5.1.24}$$

$$\tag{5.1.25}$$

(v) Prolongation:

$$q_l := p_{l-1}^l q_{l-1}^R\,; \tag{5.1.26}$$

(vi) Correction: With a damping parameter $\omega_l \in (0, 1]$,

$$\bar{\bar{v}}_l := \bar{v}_l + \omega_l q_l; \tag{5.1.27}$$

(vii) Post-smoothing:

$$v_l^1 := \mathcal{S}_l^{\mu}(\bar{\bar{v}}_l); \tag{5.1.28}$$

In case $l = L$, one sets:

$$u_L^{t+1} := v_l^1. \tag{5.1.29}$$

We collect the afore mentioned algorithmical steps into a compact systematics of the multigrid cycle $u_L^t \to u_L^{t+1}$:

$$u_L^t \;\to\; \bar{u}_L^t = S_L^{\nu}(u_L^t) \;\to\; d_L = f_L - \mathcal{A}_L \bar{u}_L^t$$

$$\downarrow \quad \tilde{d}_{L-1} = r_L^{L-1} d_{L-1} \quad \text{(restriction)}$$

$$q_{L-1} = \tilde{\mathcal{A}}_{L-1}^{-1} \tilde{d}_{L-1} \quad (R\text{-times defect correction})$$

$$\downarrow \quad \tilde{q}_L = p_{L-1}^L q_{L-1} \quad \text{(prolongation)}$$

$$\bar{\bar{u}}_L^t = \bar{u}_L^t + \omega_L \tilde{q}_L \;\to\; u_L^{t+1} = S_L^{\mu}(\bar{\bar{u}}_L^t).$$

If the defect equation $\mathcal{A}_{L-1} q_{L-1} = \tilde{d}_{L-1}$ on the coarser mesh $\mathbb{T}_{L-1}$ is solved "exactly", one speaks of a "two-grid method". In practice, the two-grid process is continued recursively to the

"multigrid method" up to the coarsest mesh $\mathbb{T}_0$. This process can be organized in various ways depending essentially on the choice of the iteration parameter R, which determines how often the defect correction step is repeated on each mesh level. In practice, for economical reasons, only the cases $R = 1$ and $R = 2$ play a role. The schemes of the corresponding multigrid cycles, the "V-cycle" and the "W-cycle", are shown in Figure 5.1. Here, "•" represnt "smoothing" and "defect correction" on the meshes $\mathbb{T}_l$, and lines "−" stand for the transfer between consecutive mesh levels.

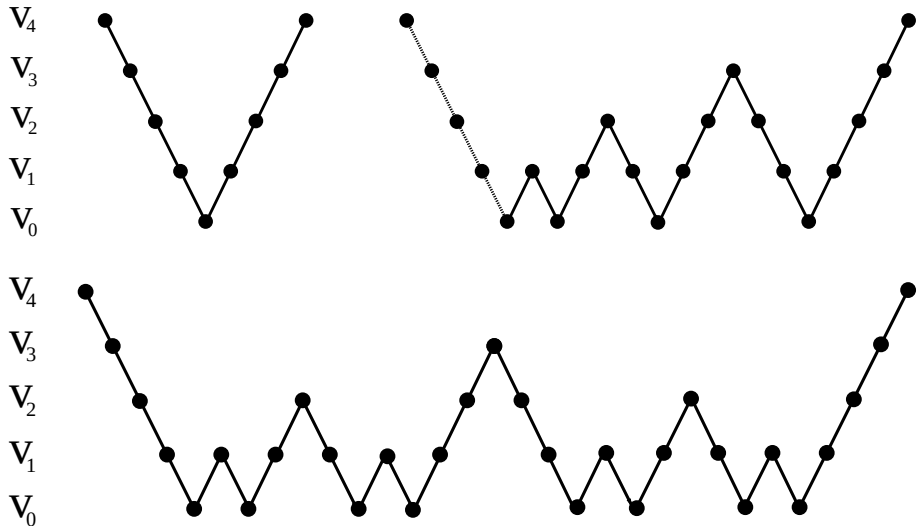


Figure 5.1: Scheme of a multigrid algorithm organized as V- (top left), F- (top right), and W-cycle (bottom line).

The V-cycle is very efficient (if it works at all), but often suffers from instabilities caused by irregularities in the problem to be solved, such as strong anisotropies in the differential operator, boundary layers, corner singularities, nonuniformities and deteriorations in the meshes (local mesh refinement and cell stretching), etc.. In contrast to that, the W-cycle is more robust but usually significantly more expensive. Multigrid methods with $R \geq 3$ are too inefficient. A good compromise between robustness and efficiency is provided by the so-called "F-cycle" shown in Figure 5.1. This process is usually started on the finest mesh $\mathbb{T}_L$ with an arbitrary initial guess u_L^0 (most often $u_L^0 = 0$). However, for economical reasons, one may start the whole multigrid process on the coarsest mesh $\mathbb{T}_0$ and then use the approximate solutions obtained on intermediate meshes as starting values for the iteration on the next finer meshes. This "nested" version of the multigrid method will be studied more systematically below.

Remark 5.1: Though the multigrid iteration in V-cycle modus may be unstable, it can be used as preconditioners for an "outer" CG (in the symmetric case) or GMRES iteration (in the nonsymmetric case). This approach combines the robustness of the Krylov space method with the efficiency of the multigrid methods and has been used very successfully for the solution of various nonstandard problems, involving singularities, indefiniteness, saddle-point structure, and multi-physics coupling.

Nested multigrid: Starting from some initial vector $u_0 := \mathcal{A}_0^{-1} f_0$ on the coarsest mesh $\mathbb{T}_0$, for $l = 1, ..., L$, successively approximations $\tilde{u}_l \approx u_l$ are computed by the following recursion:

$$\begin{aligned} u_l^0 &= p_{l-1}^l \tilde{u}_{l-1} \\ u_l^t &= MG(l, u_l^{t-1}, f_l), \quad t = 1, ..., t_l, \quad \|u_l^{t_l} - u_l\|_{L^2} \leq \hat{c}\, h_l^2\, \|f\|_{L^2}, \\ \tilde{u}_l &= u_l^{t_l}. \end{aligned}$$

Remark 5.2: There is not something like *the* multigrid algorithm. The successful realization of the multigrid concept requires a careful choice of the various components such as the "smoother" $\mathcal{S}_l$, the coarse-mesh operators $\mathcal{A}_l$, and the mesh-transfer operators r_l^{l-1}, p_{l-1}^l specially adapted to the particular structure of the problem considered. In the following, we discuss these algorithmical componenents in the context of the finite element discretization, e. g., of the model problem from above.

i) Smoothers: "Smoothers" are usually simple fixed-point iterations, which could principally also be used as "solvers", but with a very bad convergence rate. They are applied on each mesh level only a few times ($\nu, \mu \sim 1 - 4$), for damping out the high-frequency components in the errors or the residuals. In the following, we consider the damped Richardson iteration with iteration matrix

$$\mathcal{S}_l := \mathcal{I}_l - \theta_l \mathcal{A}_l\,, \quad \theta_l = \lambda_{max}(\mathcal{A}_l)^{-1}, \tag{5.1.30}$$

as smoother, which, however, only works for very simple (scalar) and non-degenerate problems.

Remark 5.3: More powerful smoothers are based on the Gauß-Seidel and the ILU iteration. These methods also work well for problems with certain pathologies. For example, in case of strong advection in the differential equation, if the mesh points are numbered in the direction of the transport, the system matrix has a dominant lower triangular part L, for which the Gauß-Seidel method is "exact". For problems with degenerate coefficients in one space direction or on strongly anisotropic meshes the system matrix has a dominant tridiagonal part, for which the ILU method is almost "exact". For indefinite saddle-point problems certain "block" iterations are used, which are specially adapted to the structure of the problem. Examples are the so-called "Vanka-type" smoothers, which are used, for example, in solving the "incompressible" Navier-Stokes equations in Fluid Mechanics.

ii) Grid transfers: In the context of a finite element discretization with nested subspaces $V_0 \subset V_1 \subset ... \subset V_l \subset ... \subset V_L$ the generic choice of the prolongation $p_{l-1}^l : V_{l-1} \to V_l$ is the cellwise embedding, and of the restriction $r_l^{l-1} : V_l \to V_{l-1}$ the L^2 projection. For other discretizations (e. g., finite difference schemes), one uses appropriate interpolation operators (e. g., bi/trilinear interpolation).

iii) Corse-grid operators: The operators $\mathcal{A}_l$ on the several spaces V_l do not need to correspond to the same discretization of the original "continuous" problem. This aspect becomes important in the use of mesh-dependent numerical diffusion ("upwinding", "streamline diffusion", etc.) for the treatment of stronger transport. Here, we only consider the ideal case that all $\mathcal{A}_l$ are defined by the same finite element discretization on the mesh family $\{\mathbb{T}_l\}_{l=0,...,L}$. In this case,

we have the following useful identity:

$$\begin{aligned}(\mathcal{A}_{l-1}v_{l-1}, w_{l-1})_{L^2} &= a(v_{l-1}, w_{l-1}) \\ &= a(p_{l-1}^l v_{l-1}, p_{l-1}^l w_{l-1}) \\ &= (\mathcal{A}_l p_{l-1}^l v_{l-1}, p_{l-1}^l w_{l-1})_{L^2} = (r_l^{l-1}\mathcal{A}_l p_{l-1}^l v_{l-1}, w_{l-1})_{L^2},\end{aligned} \tag{5.1.31}$$

for all $w_{l-1} \in V_{l-1}$, which means that

$$\mathcal{A}_{l-1} = r_l^{l-1}\mathcal{A}_l p_{l-1}^l. \tag{5.1.32}$$

iv) Coarse-grid correction: The correction step contains a damping parameter $\omega_l \in (0, 1]$. It has proved very useful in practice to choose this parameter such that the defect $\mathcal{A}_l\bar{\bar{v}}_l - \tilde{d}_{l-1}$ becomes minimal. This leads to the formula

$$\omega_l = \frac{(\mathcal{A}_l\bar{\bar{v}}_l, \tilde{d}_{l-1} - \mathcal{A}_l\bar{\bar{v}}_l)_{L^2}}{\|\mathcal{A}_l\bar{\bar{v}}_l\|_{L^2}^2}. \tag{5.1.33}$$

In the following analysis of the multigrid process, for simplicity, we will make the choice $\omega_l = 1$.

5.1.2 Convergence analysis

The classical analysis of the multigrid process is based on its interpretation as a defect-correction iteration and the concept of recursive application of the two-grid method. For simplicity, we assume that only pre-smoothing is used, i. e., $\nu > 0$, $\mu = 0$, and that in the correction step no damping is applied, i. e., $\omega_l = 1$. Then, the two-grid algorithm can be written in the form

$$\begin{aligned}u_L^{t+1} &= S_L^\nu(u_L^t) + p_{L-1}^L\mathcal{A}_{L-1}^{-1}r_L^{L-1}\big(f_L - \mathcal{A}_L S_L^\nu(u_L^t)\big) \\ &= S_L^\nu(u_L^t) + p_{L-1}^L\mathcal{A}_{L-1}^{-1}r_L^{L-1}\mathcal{A}_L\big(u_L - S_L^\nu(u_L^t)\big).\end{aligned}$$

Hence, for the iteration error $e_L^t := u_L^t - u_L$ there holds

$$e_L^{t+1} = \big(\mathcal{I}_L - p_{L-1}^L\mathcal{A}_{L-1}^{-1}r_L^{L-1}\mathcal{A}_L\big)\big(S_L^\nu(u_L^t) - u_L\big). \tag{5.1.34}$$

The smoothing operation is given in (affin)-lineare form as

$$S_L(v_L) := S_L v_L + g_L,$$

and as fixed-point iteration satisfies $S_L(u_L) = u_L$. From this, we conclude that

$$S_L^\nu(u_L^t) - u_L = S_L\big(S_L^{\nu-1}(u_L^t) - u_L\big) = ... = S_L^\nu e_L^t.$$

With the so-called "two-grid operator"

$$ZG_L(\nu) := (\mathcal{I}_L - p_{L-1}^L\mathcal{A}_{L-1}^{-1}r_L^{L-1}\mathcal{A}_L)\mathcal{S}_L^\nu$$

there consequently holds

$$e_L^{t+1} = ZG_L(\nu)e_L^t\,. \tag{5.1.35}$$

Theorem 5.1 (Two-grid convergence): *For sufficiently frequent smoothimg, $\nu > 0$, the two-grid method converges with a rate independent of the mesh level L:*

$$\|ZG_L(\nu)\|_{L^2} \leq \rho_{ZG}(\nu) = c\,\nu^{-1} < 1\,. \tag{5.1.36}$$

Proof. We write

$$ZG_L(\nu) = (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1})\mathcal{A}_L \mathcal{S}_L^\nu \tag{5.1.37}$$

and estimate as follows:

$$\|ZG_L(\nu)\|_{L^2} \leq \|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\|_{L^2} \|\mathcal{A}_L \mathcal{S}_L^\nu\|_{L^2}. \tag{5.1.38}$$

The first term on the right-hand side describes the quality of the approximation of the fine-grid solution on the next coarser mesh, while the second term represents the smoothing effect. The goal of the further analysis is now to show that the smoother $S_L(\cdot)$ possesses the so-called "smoothing property",

$$\|\mathcal{A}_L \mathcal{S}_L^\nu v_L\|_{L^2} \leq c_s \nu^{-1} h_L^{-2} \|v_L\|_{L^2}, \quad v_L \in V_L\,, \tag{5.1.39}$$

and the coarse-grid correction possesses the so-called "approximation property" ,

$$\|(\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) v_L\|_{L^2} \leq c_a h_L^2 \|v_L\|_{L^2}, \quad v_L \in V_L\,, \tag{5.1.40}$$

with positive constants c_s, c_a independent of the mesh level L. Combination of these two estimates then yields the asserted estimate (5.1.36). For sufficiently frequent smoothing, we have $\rho_{ZG} := c\nu^{-1} < 1$ and the two-grid algorithm converges with a rate uniformly with respect to the mesh level L. All constants appearing in the following are independent of L.

(i) Smoothing property: The selfadjoint operator $\mathcal{A}_L$ possesses real, positive eigenvalues $0 < \lambda_1 \leq ... \leq \lambda_i \leq ... \leq \lambda_{n_L} =: \Lambda_L$ and a corresponding L^2-ONS of eigenfunctions $\{w^1, ..., w^{n_L}\}$, such that each $v_L \in V_L$ can be written as $v_L = \sum_{i=1}^{n_L} \gamma_i w^i$, $\gamma_i = (v_L, w^i)_{L^2}$. For the Richardson iteration operator,

$$\mathcal{S}_L := \mathcal{I}_L - \theta_L \mathcal{A}_L : V_L \to V_L, \quad \theta_L = \Lambda_L^{-1}, \tag{5.1.41}$$

there holds

$$\mathcal{A}_L \mathcal{S}_L^\nu v_L = \sum_{i=1}^{n_L} \gamma_i \lambda_i \Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^\nu w^i, \tag{5.1.42}$$

and, consequently,

$$\begin{aligned}
\|\mathcal{A}_L \mathcal{S}_L^\nu v_L\|_{L^2}^2 &= \sum_{i=1}^{n_L} \gamma_i^2 \lambda_i^2 \Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^{2\nu} \\
&\leq \Lambda_L^2 \max_{1 \leq i \leq n_L} \Big\{ \Big(\frac{\lambda_i}{\Lambda_L}\Big)^2 \Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^{2\nu} \Big\} \sum_{i=1}^{n_L} \gamma_i^2 \\
&= \Lambda_L^2 \max_{1 \leq i \leq n_L} \Big\{ \Big(\frac{\lambda_i}{\Lambda_L}\Big)^2 \Big(1 - \frac{\lambda_i}{\Lambda_L}\Big)^{2\nu} \Big\} \|v_L\|_{L^2}^2.
\end{aligned}$$

By the relation (exercise)

$$\max_{0\le x\le 1}\{x^2(1-x)^{2\nu}\} \le (1+\nu)^{-2} \tag{5.1.43}$$

it follows that

$$\|\mathcal{A}_L\mathcal{S}_L^\nu v_L\|_{L^2}^2 \le \Lambda_L^2(1+\nu)^{-2}\|v_L\|_{L^2}^2. \tag{5.1.44}$$

The relation $\Lambda_L \le ch_L^{-2}$ eventually implies the asserted inequality for the Richardson iteration operator:

$$\|\mathcal{A}_L\mathcal{S}_L^\nu\|_{L^2} \le c_s\nu^{-1}h_L^{-2}, \quad \nu \ge 1. \tag{5.1.45}$$

(ii) Approximation property: We recall that in the present context of nested subspaces V_l prolongationen and restriction operators are given by

$$p_{L-1}^L = id. \text{ (identity)}, \quad r_L^{L-1} = \mathcal{P}_{L-1} \ (L^2 \text{ projection}).$$

Further, the operator $\mathcal{A}_L : V_L \to V_L$ satisfies

$$(\mathcal{A}_L v_L, \varphi_L)_{L^2} = a(v_L, \varphi_L), \quad v_L, \varphi_L \in V_L.$$

For an arbitrary but fixed $f_L \in V_L$ and functions $v_L := \mathcal{A}_L^{-1}f_L$, $v_{L-1} := \mathcal{A}_{L-1}^{-1}r_L^{L-1}f_L$ there holds:

$$\begin{aligned} a(v_L, \varphi_L) &= (f_L, \varphi_L)_{L^2} \quad \forall \varphi_L \in V_L, \\ a(v_{L-1}, \varphi_{L-1}) &= (f_L, \varphi_{L-1})_{L^2} \quad \forall \varphi_{L-1} \in V_{L-1}. \end{aligned}$$

To the function $v_L \in V_L$, we associate a function $v \in V \cap H^2(\Omega)$ as solution of the "continuous" boundary value problem

$$Lv = f_L \ \text{ in } \Omega, \quad v = 0 \ \text{ on } \partial\Omega, \tag{5.1.46}$$

or in "weak" formulation

$$a(v, \varphi) = (f_L, \varphi)_{L^2} \quad \forall \varphi \in V. \tag{5.1.47}$$

For this auxiliary problem, we have the following a priori estimate

$$\|\nabla^2 v\|_{L^2} \le c\|f_L\|_{L^2}. \tag{5.1.48}$$

There holds

$$\begin{aligned} a(v_L, \varphi_L) &= (f_L, \varphi_L)_{L^2} = a(v, \varphi_L), \quad \varphi_L \in V_L, \\ a(v_{L-1}, \varphi_{L-1}) &= (f_L, \varphi_{L-1})_{L^2} = a(v, \varphi_{L-1}), \quad \varphi_{L-1} \in V_{L-1}, \end{aligned}$$

i. e., v_L and v_{L-1} are the Ritz projections of v into V_L and V_{L-1}, respectively. For these the following L^2-error estimates hold true:

$$\|v_L - v\|_{L^2} \le ch_L^2\|\nabla^2 v\|_{L^2}, \quad \|v_{L-1} - v\|_{L^2} \le ch_{L-1}^2\|\nabla^2 v\|_{L^2}. \tag{5.1.49}$$

This together with the a priori estimate (5.1.48) and observing $h_{L-1} \leq 4h_L$ implies that

$$\|v_L - v_{L-1}\|_{L^2} \leq ch_L^2 \|\nabla^2 v\|_{L^2} \leq ch_L^2 \|f_L\|_{L^2}, \tag{5.1.50}$$

and, consequently,

$$\|\mathcal{A}_L^{-1} f_L - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1} f_L\|_{L^2} \leq ch_L^2 \|f_L\|_{L^2}. \tag{5.1.51}$$

From this, we obtain the desired estimate

$$\|\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}\|_{L^2} \leq ch_L^2, \tag{5.1.52}$$

which completes the proof. Q.E.D.

The foregoing result for the two-grid algorithm will now be used for inferring the convergence of the full multigrid method.

Theorem 5.2 (Multigrid convergence): *Suppose that the two-grid algorithm converges with rate $\rho_{ZG}(\nu) \to 0$ for $\nu \to \infty$, uniformly with respect to the mesh level L. Then, for sufficiently frequent smoothing the multigrid method with $R \geq 2$ (W-cycle) converges with rate $\rho_{MG} < 1$ independent of the mesh level L,*

$$\|u_L - MG(L, u_L^t, f_L)\|_{L^2} \leq \rho_{MG} \|u_L - u_L^t\|_{L^2}. \tag{5.1.53}$$

Proof. The proof is given by induction with respect to the mesh level L. We consider only the relevant case $R = 2$ (W-cycle) and, for simplicity, will not try to optimize the constants occurring in the course of the argument. Let ν be chosen sufficiently large such that the convergence rate of the two-grid algorithm is $\rho_{ZG} \leq 1/8$. We want to show that then the convergence rate of the full multigrid algorithm is $\rho_{MG} \leq 1/4$, uniformly with respect to the mesh level L. For $L = 1$ this is obviously fulfilled. Suppose now that also $\rho_{MG} \leq 1/4$ for mesh level $L-1$. Then, on mesh level L, starting from the iterate u_L^t, with the approximative solution q_{L-1}^2 (after 2-fold application of the coarse-mesh correction) and the exact solution $\hat{q}_{L-1}$ of the defect equation on mesh level $L-1$, there holds

$$u_L^{t+1} = MG(L, u_L^t, f_L) = ZG(L, u_L^t, f_L) + p_{L-1}^L (q_{L-1}^2 - \hat{q}_{L-1}). \tag{5.1.54}$$

According to the induction assumption (observing that the starting value of the multigrid iteration on mesh level $L-1$ is zero and that $\hat{\rho}_{L-1} = \mathcal{A}_{L-1}^{-1} r_L^{L-1} d_L$) it follows that

$$\|\hat{q}_{L-1} - q_{L-1}^2\|_{L^2} \leq \rho_{MG}^2 \|\hat{q}_{L-1}\|_{L^2} = \rho_{MG}^2 \|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu (u_L - u_L^t)\|_{L^2}. \tag{5.1.55}$$

Combination of the last two relations implies for the iteration error $e_L^t := u_L^t - u_L$ that

$$\|e_L^{t+1}\|_{L^2} \leq \left(\rho_{ZG} + \rho_{MG}^2 \|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu\|_{L^2}\right) \|e_L^t\|_{L^2}. \tag{5.1.56}$$

The norm on the right-hand side has been estimated already in connection with the convergence analysis of the two-grid algorithm. Recalling the two-grid operator

$$ZG_L = (\mathcal{A}_L^{-1} - p_{L-1}^L \mathcal{A}_{L-1}^{-1} r_L^{L-1}) \mathcal{A}_L \mathcal{S}_L^\nu = S_L^\nu - p_L^{L-1} \mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu,$$

there holds

$$\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu = S_L^\nu - ZG_L,$$

und, consequently,

$$\|\mathcal{A}_{L-1}^{-1} r_L^{L-1} \mathcal{A}_L \mathcal{S}_L^\nu\|_{L^2} \leq \|S_L^\nu\|_{L^2} + \|ZG_L\|_{L^2} \leq 1 + \rho_{ZG} \leq 2. \tag{5.1.57}$$

This eventually implies

$$\|e_L^{t+1}\|_{L^2} \leq \left(\rho_{ZG} + 2\rho_{MG}^2\right) \|e_L^t\|_{L^2}. \tag{5.1.58}$$

By the assumption on ρ_{ZG} and the induction assumption, we conclude

$$\|e_L^{t+1}\|_{L^2} \leq \left(\tfrac{1}{8} + 2\tfrac{1}{16}\right) \|e_L^t\|_{L^2} \leq \tfrac{1}{4}\|e_L^t\|_{L^2}, \tag{5.1.59}$$

which completes the proof. Q.E.D.

Remark 5.4: For well-conditioned problems (symmetrical, positive definite operator, regular coefficients, quasi-uniform meshes, etc.) one achieves multigrid convergence rates in the range $\rho_{MG} = 0,05 - 0,5$. The above analysis only applies to the W-cycle since in part (ii), we need that $R \geq 2$. The V-cycle cannot be treated on the basis of the two-grid analysis. In the literature there are more general approaches, which allow to prove convergence of multigrid methods under much weaker conditions.

Next, we analyze the computational complexity of the full multigrid algorithm. For this, we introduce the following notation:

$$\begin{aligned}
OP(T) &= \text{number of a. op. for performing the operation } T,\\
R &= \text{number of defect-correction steps on the different mesh levels},\\
n_l &= \dim V_l \approx h_l^{-d} \ (d = \text{space dimension}),\\
\kappa &= \max_{1\leq l\leq L} n_{l-1}/n_l < 1,\\
C_0 &= OP(\mathcal{A}_0^{-1})/n_0,\\
C_s &= \max_{1\leq l\leq L}\{OP(\mathcal{S}_l)/n_l\}, \quad C_d = \max_{1\leq l\leq L}\{OP(d_l)/n_l\},\\
C_r &= \max_{1\leq l\leq L}\{OP(r_l)/n_l\}, \quad C_p = \max_{1\leq l\leq L}\{OP(p_l)/n_l\}.
\end{aligned}$$

In practice mostly $\kappa \approx 2^{-d}$, $C_s \approx C_d \approx C_r \approx C_p \approx \#\{a_{nm} \neq 0\}$ and $C_0 n_0 \ll n_L$.

Theorem 5.3 (Multigrid complexity): *Under the condition $q := R\kappa < 1$, for the full multigrid cycle MG_L there holds*

$$OP(MG_L) \leq C_L n_L, \tag{5.1.60}$$

where

$$C_L = \frac{(\nu+\mu)C_s + C_d + C_r + C_p}{1-q} + C_0 q^L.$$

The multigrid algorithm for approximating the n_L-dimensional discrete solution $u_L \in V_L$ on the finest mesh $\mathbb{T}_L$ within discretization accuracy $\mathcal{O}(h_L^2)$ requires $\mathcal{O}(n_L \ln(n_L))$ a. op., and therefore has (almost) optimal complexity.

Proof. We set $C_l := OP(MG_l)/n_l$. One multigrid cycle contains the R-fold application of the same algorithm on the next coarser mesh. Observing $n_{l-1} \leq \kappa n_l$ and setting $\hat{C} := (\nu+\mu)C_s + C_d + C_r + C_p$ it follows that

$$C_L n_L = OP(MG_L) \leq \hat{C} n_L + R \cdot OP(MG_{L-1}) = \hat{C} n_L + R \cdot C_{L-1} n_{L-1} \leq \hat{C} n_L + q C_{L-1} n_L,$$

and consequently $C_L \leq \hat{C} + q C_{L-1}$. Recursive use of this relation yields

$$\begin{aligned} C_L &\leq \hat{C} + q\big(\hat{C} + qC_{L-2}\big) = \hat{C}(1+q) + q^2 C_{L-1} \\ &\vdots \\ &\leq \hat{C}(1 + q + q^2 + ... + q^{L-1}) + q^L C_0 \leq \frac{\hat{C}}{1-q} + q^L C_0. \end{aligned}$$

This implies the asserted estimate (5.1.60). The total complexity of the algotrithm then results from the relations

$$\rho_{MG}^t \approx h_L^2 \approx n_L^{-2/d}, \qquad t \approx -\frac{\ln(n_L)}{\ln(\rho_{MG})}.$$

The proof is complete. Q.E.D.

It should be emphasized that in the proof of (5.1.60) the assumption

$$q := R\kappa = R \max_{1 \leq l \leq L} n_{l-1}/n_l < 1 \tag{5.1.61}$$

is essential. This means for the W-cycle ($R = 2$) that by the transition from mesh $\mathbb{T}_{l-1}$ to the next finer mesh $\mathbb{T}_l$ the number of mesh points (dimension of spaces) sufficiently increases, comparibly to the situation of uniform mesh refinement,

$$n_l \approx 4 n_{l-1}. \tag{5.1.62}$$

Remark 5.5: In an adaptively driven mesh refinement process with only local mesh refinement the condition (5.1.61) is usually not satisfied. Mostly only $n_l \approx 2n_{l-1}$ can be expected. In such a case the multigrid process needs to be modified in order to preserve good efficiency. This may be accomplished by applying the cost-intensive smoothing only to those mesh points, which have been newly created by the transition from mesh $\mathbb{T}_{l-1}$ to mesh $\mathbb{T}_l$. The implementation of a multigrid algorithm on locally refined meshes requires much care in order to achieve optimal complexity of the overall algorithm.

The nested multigrid algorithm turns out to be complexity optimal, as it requires only $O(n_L)$ a. op. for producing a sufficiently accurate approximation to the discrete solution $u_L \in V_L$.

Theorem 5.4 (Nested multigrid): *The nested multigrid algorithm is complexity-optimal, i. e., it delivers an approximation to the discrete solution $u_L \in V_L$ on the finest mesh $\mathbb{T}_L$ with discretization accuracy $\mathcal{O}(h_L^2)$ with complexity $\mathcal{O}(n_L)$ a. op. independent of the mesh level L.*

Proof. The accuracy requirement for the multigrid algorithm on mesh level $\mathbb{T}_L$ is

$$\|e_L^t\|_{L^2} \leq \hat{c} h_L^2 \|f\|_{L^2}. \tag{5.1.63}$$

(i) First, we want to show that, under the assumptions of Theorem 5.2, the result (5.1.63) is achievable by the nested multigrid algorithm on each mesh level L by using a fixed (sufficiently large) number t_* of multigrid cycles. Let $e_L^t := u_L^t - u_L$ be again the iteration error on mesh level L. By assumption $e_0^t = 0$, $t \geq 1$. In case $u_L^0 := u_{L-1}^t$ there holds

$$\begin{aligned}\|e_L^t\|_{L^2} &\leq \rho_{MG}^t \|e_L^0\|_{L^2} = \rho_{MG}^t \|u_{L-1}^t - u_L\|_{L^2} \\ &\leq \rho_{MG}^t \big(\|u_{L-1}^t - u_{L-1}\|_{L^2} + \|u_{L-1} - u\|_{L^2} + \|u - u_L\|_{L^2}\big) \\ &\leq \rho_{MG}^t \big(\|e_{L-1}^t\|_{L^2} + ch_L^2 \|f\|_{L^2}\big).\end{aligned}$$

Recursive use of this relation for $L \geq l \geq 1$ then yields (observing $h_l \leq \kappa^{l-L} h_L$)

$$\begin{aligned}\|e_L^t\|_{L^2} &\leq \rho_{MG}^t \big(\rho_{MG}^t \big(\|e_{L-2}^t\|_{L^2} + ch_{L-1}^2 \|f\|_{L^2}\big) + ch_L^2 \|f\|_{L^2}\big) \\ &\vdots \\ &\leq \rho_{MG}^{Lt} \|e_0^t\|_{L^2} + \big(c\rho_{MG}^t h_L^2 + c\rho_{MG}^{2t} h_{L-1}^2 + \ldots + c\rho_{MG}^{Lt} h_1^2\big) \|f\|_{L^2} \\ &= ch_L^2 \kappa^2 \left(\rho_{MG}^t \kappa^{-2\cdot 1} + \rho_{MG}^{2t} \kappa^{-2\cdot 2} + \ldots + \rho_{MG}^{Lt} \kappa^{-2L}\right) \|f\|_{L^2} \\ &\leq ch_L^2 \kappa^2 \|f\|_{L^2} \, \frac{\kappa^{-2} \rho_{MG}^t}{1 - \kappa^{-2} \rho_{MG}^t},\end{aligned}$$

provided that $\kappa^{-2} \rho_{MG}^t < 1$. Obviously there exists a t_*, such that (5.1.63) is satisfied for $t \geq t_*$ uniformly with respect to L.

(ii) We can now carry out the complexity analysis. Theorem 5.3 states that one cycle of the simple multigrid algorithm $MG(l, \cdot, \cdot)$ on the l-th mesh level requires $W_l \leq c_* n_l$ a. op. (uniformly with respect to l). Let now $\hat{W}_l$ be the number of a. op. of the nested multigrid algorithm on mesh level l. Then, by construction there holds

$$\hat{W}_L \leq \hat{W}_{L-1} + t_* W_L \leq \hat{W}_{L-1} + t_* c_* n_L.$$

Iterating this relation, we obtain with $\kappa := \max_{1 \leq l \leq L} n_{l-1}/n_l < 1$ that

$$\begin{aligned}\hat{W}_L &\leq \hat{W}_{L-1} + t_* c_* n_L \leq \hat{W}_{L-2} + t_* c_* n_{L-1} + t_* c_* n_L \\ &\vdots \\ &\leq t_* c_* \{n_L + \ldots + n_0\} \leq ct_* c_* n_L \{1 + \ldots + \kappa^L\} \leq \frac{ct_* c_*}{1 - \kappa} n_L,\end{aligned}$$

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

5.2 Multigrid methods for eigenvalue problems (a short review)

The application of the "multigrid concept" to the solution of high-dimensional eigenvalue problems can follow different pathes. First, there is the possibility of using it directly for the eigenvalue problem based on its reformulation as a nonlinear system of equations, which allows for the formation of "residuals". Second, the multigrid concept may be used as components of other iterative methods, such as the Krylov space methods, for accelerating certain computation-intensive substeps. In the following, we will only briefly describe these different approaches.

5.2.1 Direct multigrid approach

The algebraic eigenvalue problem

$$Az = \lambda z, \quad \lambda \in \mathbb{C},\ z \in \mathbb{C}^n,\ \ \|z\|_2 = 1, \tag{5.2.64}$$

is equivalent to the following nonlinear system of equations

$$\left\{ \begin{array}{c} Az - \lambda z \\ \|z\|_2^2 - 1 \end{array} \right\} = 0. \tag{5.2.65}$$

To this system, we may apply a nonlinear version of the multigrid method described in Section 5.1 again yielding an algorithm of optimal complexity, at least in principly (for details see, e. g., Brand et al. [24] and Hackbusch [34]). However, this approach suffers from stability problems in case of irregularities of the underlying continuous problem, such as anisotropies in the operator, the domain or the computational mesh, which may spoil the convergence of the method or render it inefficient. One cause may be the lack of approximation property in case that the continuous eigenvalue problem is not well approximated on coarser meshes, which is essential for the convergence of the multigrid method. The possibility of such a pathological situation is illustrated by the following example, which suggests to use the multigrid concept not directly but rather for accelerating the cost-intensive components of other more robust methods such as the Krylov space methods (or the Jacobi-Davidson method) described above.

Example 5.1: We consider the following non-symmetric convection-diffusion eigenvalue problem on the unit square $\Omega = (0,1)^2 \in \mathbb{R}^2$:

$$-\nu\Delta u + b \cdot \nabla u = \lambda u, \ \text{ in } \Omega, \quad u = 0, \ \text{ on } \partial\Omega, \tag{5.2.66}$$

with coefficients $\nu > 0$ and $c = (c_1, c_2) \in \mathbb{R}^2$. The (real) eigenvalues are explicitly given by

$$\lambda = \frac{b_1^2 + b_2^2}{4\nu} + \nu\pi^2(n_1^2 + n_2^2), \quad n_1, n_2 \in \mathbb{N},$$

with corresponding (non-normalized) eigenfunctions

$$w(x_1, x_2) = \exp\Big(\frac{b_1 x_1 + b_2 x_2}{2\nu}\Big) \sin(n_1\pi x_1) \sin(n_2\pi x_2).$$

The corresponding adjoint eigenvalue problem has the eigenfunctions

$$w^*(x_1, x_2) = \exp\Big(-\frac{b_1 x_1 + b_2 x_2}{2\nu}\Big) \sin(n_1\pi x_1) \sin(n_2\pi x_2).$$

This shows firstly that the underlying differential operator in (5.2.66) is non-normal and secondly that the eigenfunctions develop strong boundary layers for small parameter values ν (transport-dominant case). In particular, the eigenvalues depend very strongly on ν. For the "direct" application of the multigrid method to this problem, this means that the "coarse-grid problems", due to insufficient mesh resolution, have completely different eigenvalues than the "fine-grid" problem, leading to insufficient approximation for computing meaningful corrections. This renders the multigrid iteration, being based on successive smoothing and coarse-grid correction, inefficient and may even completely spoil convergence.

5.2.2 Accelerated Arnoldi and Lanczos method

The most computation-intensive part of the Arnoldi and Lanczos methods in the case of the approximation of the smallest (by modulus) eigenvalues of a high-dimensional matrix $A \in \mathbb{K}^{n\times n}$ is the generation of the Krylov space

$$K_m = \operatorname{span}\{q, A^{-1}q, \ldots, (A^{-1})^{m-1}q\},$$

which requires the solution of a small number $m \ll n$ but high-dimensional linear systems with A as coefficient matrix. Even though the Krylov space does not need to be explicitly constructed in the course of the modified Gram-Schmidt algorithm for the generation of an orthonormal basis $\{q^1, \ldots, q^m\}$ of K_m, this process requires the same amount of computation. This computational "acceleration" by use of multigrid techniques is exploited in Section 4.3.2 on the computation of pseudospectra. We want to illustrate this for the simpler situation of the "inverse iteration" for computing the smallest eigenvalue of a symmetric and positive definite matrix $A \in \mathbb{R}^{n\times n}$.

Recall Example 4.1 in Section 4.1.1, the eigenvalue problem of the Laplace operator on the unit square:

$$\begin{gathered} -\frac{\partial^2 w}{\partial x^2}(x,y) - \frac{\partial^2 w}{\partial y^2}(x,y) = \mu w(x,y) \quad \text{for } (x,y) \in \Omega, \\ w(x,y) = 0 \quad \text{for } (x,y) \in \partial\,\Omega. \end{gathered} \tag{5.2.67}$$

The discretization of this eigenvalue problem by the 5-point difference scheme on a uniform cartesian mesh or the related finite element method with piecewise linear trial functions leads to the matrix eigenvalue problem

$$Az = \lambda z, \quad \lambda = h^2\mu, \tag{5.2.68}$$

with the same block-tridiagonal matrix A as occurring in the corresponding discretization of the boundary value problem discussed in Section 3.4. We are interested in the smallest eigenvalue $\lambda_1 = \lambda_{\min}$ of A, which by $h^{-2}\lambda_{\min} \approx \mu_{\min}$ yields an approximation to the smallest eigenvalue of problem (5.2.67). For λ_1 and the next eigenvalue $\lambda_2 > \lambda_1$ there holds

$$\lambda_1 = 2\pi^2 h^2 + O(h^4), \quad \lambda_2 = 5\pi^2 h^2 + O(h^4).$$

For computing λ_1, we may use the inverse iteration with shift $\lambda = 0$. This requires in each step the solution of a problem like

$$A\tilde{z}^t = z^{t-1}, \quad z^t := \|\tilde{z}^t\|_2^{-1}\tilde{z}^t. \tag{5.2.69}$$

For the corresponding eigenvalue approximation

$$\frac{1}{\lambda_1^t} := \frac{(A^{-1}z^t, z^t)_2}{\|z^t\|_2^2} = (\tilde{z}^{t+1}, z^t)_2, \tag{5.2.70}$$

there holds the error estimate (see exercise in Section 4.1.1)

$$\Big|\frac{1}{\lambda_1^t} - \frac{1}{\lambda_1}\Big| \le \Big|\frac{1}{\lambda_1}\Big| \frac{\|z^0\|_2^2}{|\alpha_1|^2}\Big(\frac{\lambda_2}{\lambda_1}\Big)^{2t}, \tag{5.2.71}$$

where α_1 is the coefficient in the expansion of z^0 with respect to the eigenvector w^1. From this relation, we infer that

$$|\lambda_1 - \lambda_1^t| \leq \lambda_1^t \frac{\|z^0\|_2^2}{|\alpha_1|^2} \Big(\frac{\lambda_2}{\lambda_1}\Big)^{2t}. \tag{5.2.72}$$

Observing that $\lambda_1^t \approx \lambda_1 \approx h^2$ and $h^2\|z^0\|_2^2 = h^2 \sum_{i=1}^n |z_i^0|^2 \approx \|v^0\|_{L^2}^2$, where $v^0 \in H_0^1(\Omega)$ is the continuous eigenfunction corresponding to the eigenvector z^0, we obtain

$$|\lambda_1 - \lambda_1^t| \leq c\Big(\frac{\lambda_2}{\lambda_1}\Big)^{2t} \leq c\,0.4^{2t}. \tag{5.2.73}$$

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 $\mu_1 = h^{-2}\lambda_1$ achieving a prescribed accuracy in the approximation of μ_1 requires the scaling of the tolerance in computing λ_1 by a factor h^2, which introduces a logarithmic h-dependence in the work count of the algorithm,

$$t(\varepsilon) \approx \frac{\log(\varepsilon h^2)}{\log(2/5)} \approx \log(n). \tag{5.2.74}$$

Now, using a multigrid solver of optimal complexity $\mathcal{O}(n)$ in each iteration step (4.1.20) the total complexity of computing the smallest eigenvalue λ_1 becomes $\mathcal{O}(n\log(n))$.

Remark 5.6: For the systematic use of multigrid acceleration within the Jacobi-Davidson method for nonsymmetric eigenvalue problems, we refer to Heuveline & Bertsch [38]. This combination of a robust iteration and multigrid acceleration seems presently to be the most efficient approach to solving large-scale symmetric or unsymmetric eigenvalue problems.

5.3 Exercises

Exercise 5.1: Consider the discretization of the Poisson problem

$$-\Delta u = f, \ \text{ in } \Omega, \quad u = 0, \ \text{ on } \partial\Omega,$$

on the unit square $\Omega = (0,1)^2 \subset \mathbb{R}^2$ by the finite element Galerkin method using linear shape and test functions on a uniform cartesian triangulation $\mathbb{T}_h = \{K\}$ with cells K (rectangular triangles) of width $h > 0$. The lowest-order finite element space on the mesh $\mathbb{T}_h$ is given by

$$V_h = \{v_h \in C(\bar{\Omega}) \,|\, v_h|_K \in P_1(K), \ K \in \mathbb{T}_h, \ v_h|_{\partial\Omega} = 0\}.$$

Its dimension is $\dim V_h = n_h$, which coincides with the number of interior nodal points a_i, $i = 1, \ldots, n_h$, of the mesh $\mathbb{T}_h$. Let $\{\varphi_h^1, \ldots, \varphi_h^{n_h}\}$ denote the usual "nodal basis" (so-called "Lagrange basis") of the finite element subspace V_h defined by the interpolation condition $\varphi_h^i(a_j) = \delta_{ij}$. Make a sketch of this situation, especially of the mesh $\mathbb{T}_h$ and a nodal basis function φ_h^i.

Then, the finite element Galerkin approximation in the space V_h as described in class results in the following linear system for the nodal value vector $x_h = (x_h^1, \ldots, x_h^{n_h})$:

$$A_h x_h = b_h,$$

with the matrix $A_h = (a_{ij})_{i,j=1}^{n_h}$ and right-hand side vector $b_h = (b_i)_{i=1}^{n_h}$ given by $a_{ij} = (\nabla\varphi_h^j, \nabla\varphi_h^i)_{L^2}$ and $b_i = (f, \varphi_h^i)_{L^2}$. Evaluate these elements a_{ij} and b_i using the trapezoidal rule for triangles

$$\int_K w(x)\,dx \approx \frac{|K|}{3} \sum_{i=1}^{3} w(a_i),$$

where a_i, $i = 1, 2, 3$, are the three vertices of the triangle K and $|K|$ its area. This quadrature rule is exact for linear polynomials. The result is a matrix and right-hand side vector which are exactly the same as resulting from the finite difference discretization of the Poisson problem on thze mesh $\mathbb{T}_h$ described in class.

Exercise 5.2: Analyse the proof for the convergence of the two-grid algorithm given in class for its possible extension to the case the restriction $r_l^{l-1} : V_l \to V_{l-1}$ is defined by local bilinear interpolation rather than by global L^2-projection onto the coarser mesh $\mathbb{T}_{l-1}$. What is the resulting difficulty? Do you have an idea how to get around it?

Exercise 5.3: The FE-discretization of the convection-diffusion problem

$$-\Delta u + \partial_1 u = f \quad \text{in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega,$$

leads to asymmetric system matrices A_h. In this case the analysis of the multigrid algorithm requires some modifications. Try to extend the proof given in class for the convergence of the two-grid algorithm for this case if again the (damped) Richardson iteration is chosen as smoother,

$$x_h^{t+1} = x_h^t - \theta_t(A_h x_h^t - b_h), \quad t = 0, 1, 2, \ldots.$$

What is the resulting difficulty and how can one get around it?

Exercise 5.4: Consider the discretization of the Poisson problem

$$-\Delta u = f, \ \text{ in } \Omega, \quad u = 0, \ \text{ on } \partial\Omega,$$

on the unit square $\Omega = (0,1)^2 \subset \mathbb{R}^2$ by the finite element Galerkin method using linear shape and test functions. Let $(\mathbb{T}_l)_{l\geq 0}$ be a sequence of equidistant cartesian meshes of width $h_l = 2^{-l}$. The discrete equations on mesh level l are solved by a multigrid method with (damped) Richardson smooting and the natural embedding as prolongation and the L^2-projection as restriction. The number of pre- and postsmoothing steps is $\nu = 2$ and $\mu = 0$, respectively. How many arithmetic operations are approximately required for a V-cycle and a W-cycle depending on the dimension $n_l = \dim V_l$?

Bibliography

[1] R. Rannacher: *Numerische Mathematik 0 (Einf. in die Numerische Mathematik)*, Lecture Notes, Heidelberg University, `http://numerik.uni-hd.de/~lehre/notes/`

[2] R. Rannacher: *Numerische Mathematik 1 (Numerik gewöhnlicher Differentialgleichungen)*, Lecture Notes, Heidelberg University, `http://numerik.uni-hd.de/~lehre/notes/`

[3] R. Rannacher: *Numerische Mathematik 2 (Numerik partieller Differentialgleichungen)*, Lecture Notes, Heidelberg University, `http://numerik.uni-hd.de/~lehre/notes/`

[4] R. Rannacher: *Numerische Mathematik 3 (Numerische Methoden der Kontinuumsmechanik)*, Lecture Notes, Heidelberg University, `http://numerik.uni-hd.de/~lehre/notes/`

(I) References on Functional Analysis, Linear Algebra and Matrix Analysis

[5] N. Dunford and J. T. Schwartz: *Linear Operators I, II, III*, Interscience Publishers and Wiley, 1957, 1963, 1971.

[6] H. J. Landau: *On Szegö's eigenvalue distribution theory and non-Hermitian kernels*, J. Analyse Math. 28, 335–357 (1975).

[7] R. A. Horn and C. R. Johnson: *Matrix Analysis*, Cambridge University Press, 1985-1999, 2007.

[8] P. M. Halmos: *Finite Dimensional Vector Spaces*, Springer, 1974.

[9] T. Kato: *Perturbation Theory for Linear Operators*, Springer, 2nd ed., 1980.

[10] H.-J. Kowalsky: *Lineare Algebra*, De Gruyter, 1967.

[11] H.-O. Kreiss: *Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximaieren*, BIT, 153–181 (1962).

[12] P. Lancester and M. Tismenetsky: *The Theory of Matrices with Applications*, Academic Press, 1985.

[13] D. W. Lewis: *Matrix Theory*, World Scientific, 1991.

[14] J. M. Ortega: *Matrix Theory, A Second Course*, Springer, 1987.

[15] B. N. Parlett: *The Symmetric Eigenvalue Problem*, Prentice-Hall, 1980.

[16] B. Rajendra: *Matrix Analysis*, Springer, 1997.

[17] L. N. Trefethen: *Pseudospectra of linear operators*, SIAM Rev. 39, 383–406 (1997).

[18] L. N. Trefethen: *Computation of pseudospectra*, Acta Numerica 8, 247–295, 1999.

[19] L. N. Trefethen and M. Embree: *Spectra and Pseudospectra*, Princeton University Press Europe, 2005.

[20] J. H. Wilkinson: *Rounding Errors in Algebraic Processes*, Prentice-Hall, 1963.

[21] J. H. Wilkinson: *The Algebraic Eigenvalue Problem*, Clarendon Press, 1965.

(II) References on Numerical Linear Algebra

[22] G. Allaire and S. M. Kaber: *Numerical Linear Algebra*, Springer, 2007.

[23] A. Björck and C. C. Paige: *Loss and recapture of orthogonality in the modified GramSchmidt algorithm*, SIAM J. Matrix Anal. Appl. 13, 176–190 (1992).

[24] A. Brandt, S. McCormick, and J. Ruge: *Multigrid method for differential eigenvalue problems*, J. Sci. Stat. Comput. 4, 244–260 (1983).

[25] Ph. G. Ciarlet: *Introduction to Numerical Linear Algebra and Optimization*, Cambridge University Press, 1989.

[26] M. Crouzeix, B. Philippe, and M. Sadkane: *The Davidson method*, SIAM J. Sci. Comput. 15, 62–76 (1994).

[27] E. R. Davidson: *The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices*, J. Comput. Phys. 17, 87–94 (1975).

[28] B. N. Datta: *Numerical Linear Algebra and Applications*, Springer, 2008.

[29] J. W. Demmel: *Applied Numerical Linear Algebra*, SIAM, 1997.

[30] P. Deuflhard and A. Hohmann: *Numerische Mathematik I*, De Gruyter, 2002 (3rd edition).

[31] D. K. Faddeev and W. N. Faddeeva: *Numerische Methoden der linearen Algebra*, Deutscher Verlag der Wissenschaften, 1964.

[32] D. Gerecht, R. Rannacher and W. Wollner: *Computational aspects of pseudospectra in hydrodynamic stability analysis*, J. Math. Fluid Mech. 14, 661–692 (2012).

[33] G. H. Golub and C. F. van Loan: *Matrix Computations*, Johns Hopkins University Press, 1984.

[34] W. Hackbusch: *Multi-Grid Methods and Applications*, Springer, 1985.

[35] W. Hackbusch: *Iterative Lösung großer schwachbesetzter Gleichungssysteme*, Teubner, 1991.

[36] G. Hämmerlin and K.-H. Hoffmann: *Numerische Mathematik*; Springer, 1989.

[37] W. W. Hager: *Applied Numerical Linear Algebra*, Prentice Hall, 1988.

[38] V. Heuveline and C. Bertsch: *On multigrid methods for the eigenvalue computation of nonselfadjoint elliptic operators*, East-West J. Numer. Math. 8, 257–342 (2000).

[39] J. G. Heywood, R. Rannacher, and S. Turek: *Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations*, Int. J. Numer. Meth. Fluids 22, 325–352 (1996).

[40] D. Meidner, R. Rannacher, and J. Vihharev: *Goal-oriented error control of the iterative solution of finite element equations*, J. Numer. Math. 17, 143-172 (2009).

[41] B. N. Parlett: *Convergence of the QR algorithm*, Numer. Math. 7, 187–193 (1965); corr. in 10, 163-164 (1965).

[42] R. Rannacher, A. Westenberger, and W. Wollner: *Adaptive finite element approximation of eigenvalue problems: balancing discretization and iteration error*, J. Numer. Math. 18, 303–327 (2010).

[43] Y. Saad: *Numerical Methods for Large Eigenvalue Problems*, Manchester University Press, 1992.

[44] H. R. Schwarz, H. Rutishauser and E. Stiefel: *Numerik symmetrischer Matrizen*, Teubner, 1968.

[45] G. L. G. Sleijpen and H. A. Van der Vorst: *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM Review42, 267–293 (2000).

[46] C. E. Soliverez and E. Gagliano: *Orthonormalization on the plane: a geometric approach*, Mex. J. Phys. 31, 743-758 (1985).

[47] J. Stoer and R. Bulirsch: *Numerische Mathematik 1/2*, Springer, 2007 (10th editions).

[48] G. Strang: *Linear Algebra and its Applications*, Academic Press, 1980.

[49] G. W. Stewart: *Introduction to Matrix Computations*, Academic Press, 1973.

[50] J. Todd: *Basic Numerical Mathematics, Vol. 2: Numerical Algebra*, Academic Press, 1977.

[51] L. N. Trefethen and D. Bau, III: *Numerical Linear Algebra*, SIAM, 1997.

[52] R. S. Varga: *Matrix Iterative Analysis*, Springer, 2000 (2nd edition).

[53] T.-L. Wang and W. B. Gragg: *Convergence of the shifted QR algorithm for unitary Hessenberg matrices*, Math. Comput. 71, 1473–1496 (2001).

[54] D. M. Young: *Iterative Solution of Large Linear Systems*, Academic Press, 1971.

(III) References on the Origin of Problems and Applications

[55] O. Axelsson and V. A. Barker: *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press, 1984.

[56] D. Braess: *Finite Elemente*, Springer 2003 (3rd edition).

[57] H. Goering, H.-G. Roos, and L. Tobiska: *Finite-Elemente-Methode*, Akademie-Verlag, 1993 (3rd edition).

[58] C. Großmann, H.-G. Roos: *Numerik partieller Differentialgleichungen*, Teubner, 1992

[59] W. Hackbusch: *Theorie und Numerik elliptischer Differentialgleichungen*, Teubner, 1986.

[60] A. R. Mitchell and D. F. Griffiths: *The Finite Difference Method in Partial Differential Equations*, Wiley, 1980

[61] A. Quarteroni and A. Valli: *Numerical Approximation of Partial Differential Equations*, Springer, 1994.

[62] M. Schäfer and S. Turek: *Benchmark computations of laminar flow around a cylinder*, in Flow Simulation with High-Performance Computer II, Notes on Numerical Fluid Mechanics, vol. 48 (Hirschel, E. H., ed.), pp. 547–566, Vieweg, 1996.

[63] H. R. Schwarz: *Numerische Mathematik*, B. G. Teubner, 1986

[64] G. Strang and G. J. Fix: *An Analysis of the Finite Element Method*, Prentice-Hall, 1973.

[65] A. Tveito and R. Winther: *Introduction to Partial Differential Equations: A Computational Approach*, Springer, 1998.

Index

A-orthogonal, 116, 121
A-scalar product, 116
additivity, 16
adjoint transpose, 22
adjuncts, 22
algorithm
    classical Gram-Schmidt, 20
    Crout, 60
    exchange, 65
    Gauß-Jordan, 63, 65
    Givens, 87
    Gram-Schmidt, 73
    Householder, 74, 86
    modified Gram-Schmidt, 21, 155
    Thomas, 68
angle, 24
ansatz space, 127
approximation property, 182
arithmetic operation, 53
Arnoldi (1917-1995), 152
Arnoldi basis, 155
Arnoldi relation, 156

backward substitution, 53, 57
Banach space, 14
band matrix, 7
band type, 67
band width, 67
basis
    orthogonal, 19
    orthonormal, 20
best approximation, 19
bilinear form, 16
Burgers equation, 165

Cauchy sequence, 14
central difference quotient, 130
characteristic polynomial, 27, 90
Chebyshev (1821-1894), 4
Chebyshev equalization, 4
Chebyshev polynomial, 111, 125
Cholesky (1975-1918), 71, 100
Cholesky decomposition, 71, 100
coarse-grid correction, 181
column pivoting, 55
condition number, 43
conditioning, 43
contraction constant, 97
coordinate relaxation, 117

defect, 5, 25, 27, 61, 72
    correction, 93
defect equation, 61, 176
definiteness, 13
Descartes (1596-1650), 1
descent direction, 117
determinant, 22
deviation
    maximal, 4
    mean, 4
difference approximation, 5
difference equation, 6
discretization, 5
dyadic product, 75

eigenspace, 3, 27
eigenvalue, 3, 27
    deficient, 27
eigenvalue equation, 27
eigenvalue problem
    full, 27
    partial, 27
eigenvector, 3, 27
energy form, 175
equalization parabola, 5
exponential stability, 36

fill-in, 7
final iteration, 61, 63
fixed-point iteration, 93
fixed-point problem, 93
forward substitution, 53, 57
Frobenius (1849-1917), 31

Galerkin (1871-1945), 121
Galerkin equation, 121
Galerkin orthogonality, 98
Gauß (1777-1855), 54
Gaussian elimination, 54, 86
Gaussian equalization, 4
generalized eigenvector, 28
Gerschgorin circle, 46
Gershgorin (1901-1933), 46
Givens (1910-1993), 87
Givens transformation, 85
Gram (1850-1916), 20
grid transfer, 180

Hölder (1859-1937), 17
half-band width, 7

Hessenberg (1904-1959), 46, 83
Hessenberg normal form, 86
Hestenes (1906-1991), 121
homogeneity, 13
Householder (1904-1993), 75
Householder transformation, 75, 86

identity
  parallelogram, 19
  Parseval, 20
identity matrix, 22
inequality
  Cauchy-Schwarz, 16
  Hölder, 18
  Minkowski, 14, 18
  Young, 17
inverse iteration, 143
iteration matrix, 93

Jacobi (1804-1851), 99
Jordan (1838-1922), 29, 65
Jordan normal form, 29, 84

Kantorovich (1912-1986), 118
kernel, 23
Kronecker symbol, 22
Krylov (1879-1955), 122
Krylov matrix, 154
Krylov space, 122

Lanczos (1893-1974), 152
Lanczos relation, 158
Laplace (1749-1827), 2
line search, 117
linear mapping, 21
linear system, 3
  overdetermined, 3
  quadratic, 3
  underdetermined, 3
load vector, 175
LR decomposition, 57, 69

machine accuracy, 60
mass matrix, 175
matrix, 22
  band, 67, 131
  consistently ordered, 107, 131
  diagonalizable, 30, 84, 144
  diagonally dominant, 69
  Frobenius, 54
  hermitian, 23
  Hessenberg, 46, 83
  ill-conditioned, 45
  inverse, 22
  irreducible, 102, 131
  Jacobi, 131
  lower triangular, 67
  normal, 23
  orthogonal, 24
  orthonormal, 24, 73
  permutation, 54
  positive definite, 23, 70, 116
  positive semi-definite, 23
  rank-deficient, 80
  reducible, 102
  regular, 22
  similar, 29, 83
  sparse, 69
  strictly diagonally dominant, 70
  symmetric, 23, 116
  triangular, 53
  tridiagonal, 46, 68
  tridiagonal matrix, 83
  unitarily diagonalizable, 30
  unitary, 24
  upper triangular, 67
matrix norm
  compatible, 31
  natural, 31
mesh-point numbering
  checkerboard, 8
  diagonal, 8
  row-wise, 7
method
  ADI, 100
  Arnoldi, 152, 154
  bisection, 90
  CG, 121, 132, 191
  Cholesky, 71
  descent, 116
  direct, 4
  finite element, 98
  Gauß-Seidel, 9, 99, 101, 105, 131
  gradient, 118, 132
  Hyman, 87
  ILU, 100
  iterative, 4

Jacobi, 9, 99, 101
Jacobi-Davidson, 146
Krylov space, 127
Lanczos, 152
Least-Error Squares, 25
LR, 146
multigrid, 173
PCG, 128
power, 141
projection, 122
QR, 147
reduction, 84
Richardson, 93, 173
SOR, 99, 104, 131
SSOR, 115
two-grid, 178
minimal solution, 26, 81
Minkowski (1864-1909), 18
Mises, von (1883-1953), 141
multigrid cycle, 177
multiplicity
algebraic, 27
geometric, 27

Navier-Stokes equation, 9
neighborhood, 14
nested multigrid, 180
Neumann 1832-1925, 34
Neumann series, 34
nodal basis, 175
norm
l_1, 13, 48
l_∞, 13, 48
l_p, 14
euclidian, 13
Frobenius, 31, 83
maximal row-sum, 32
maximum, 13
spectral, 31, 52
submultiplicative, 31
normal equation, 26, 72
normal form, 30
normed space, 13
null space, 23
null vector, 1
numerical rank, 80

operator
5-point, 6
7-point, 6
coarse-grid, 180
divergence, 2
gradient, 2
Laplace, 2, 5, 130, 145, 189
nabla, 2
two-grid, 181
orthogonal
complement, 19
projection, 19
system, 19
overrelaxation, 104

Parseval (1755-1836), 20
Penrose (1931-), 82
perturbation equation, 36
pivot element, 55, 64
pivot search, 55
pivoting, 58
point, 1
accumulation, 16
isolated, 16
poor man's pseudospectra, 160
post-smoothing, 176
pre-smoothing, 176, 178
preconditioner, 93
preconditioning, 126, 128
diagonal, 128
ICCG, 129
SSOR, 129
product space, 16
prolongation, 177
pseudoinverse, 82
pseudospectrum, 38

QR decomposition, 73

range, 23
Rayleigh (1842-1919), 28
Rayleigh quotient, 28, 142
reflection, 75, 86
relaxation parameter, 99
relaxation step, 104
residual, 97
resolvent, 27
restriction, 177
Richardson (1881-1953), 93
Ritz eigenvalue, 153
rotation, 24, 85

row sum criterion, 101, 103
Rutishauser (1918-1970), 146

scalar product, 16
  euclidian, 16
  semi, 16
Schmidt (1876-1959), 20
Schur (1875-1941), 84
Schur normal form, 84
Seidel, von (1821-1896), 9
sequence
  bounded, 14
  convergent, 14
sesquilinear form, 16, 49
set
  closed, 14
  compact, 16
  open, 14
  resolvent, 27
  sequentially compact, 16
similarity transformation, 29
singular value decomposition, 78
smoother, 180
smoothing operation, 180
smoothing property, 182
Sobolev space, 174
solution operator, 37
spectral condition, 44
spectral radius, 94, 96
spectrum, 27
step length, 117
Stiefel (1909-1978), 121
stiffness matrix, 175
stopping criterion, 97
Sturm (1803-1855), 89
Sturm chain, 89
Sturm-Liouville problem, 164
system matrix, 175

test space, 127
theorem
  Banach, 94
  Cauchy, 14
  Gerschgorin, 46
  Kantorovich, 119
  multigrid complexity, 185
  multigrid convergence, 184
  nested multigrid, 186
  norm equivalence, 15
  Pythagoras, 19
  two-grid convergence, 181
total pivoting, 55, 59
trace, 22
triangle inequality, 13

underrelaxation, 104

V-cycle, 179
Vandermonde (1735-1796), 5
Vandermondian determinant, 5
vector, 1
  norm, 13
  space, 1

W-cycle, 179
Wielandt (1910-2001), 143
Wilkinson (1919-1986), 86

Young (1863-1942), 17