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" 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, "iterative" 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 and work requirements, 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 book 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

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

In case that $a_{jj} \neq 0$, $j = 1, \ldots, n$, we obtain the solution by "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 "(arithmetic) complexity", we use the notion "arithmetic 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*" and "*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 1-st and the r-th row. Let the result be the matrix $[\tilde{A}^{(0)}, \tilde{b}^{(0)}]$. Then, for $j = 2, \ldots, n$, we multiply the 1-st row by q_{j1} and subtract the result from the j-th 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.

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

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 1-st row with the r-th row and the 1-st column with the s-th 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.

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

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

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 i-th elimination step $[A^{(i-1)}, b^{(i-1)}] \to [A^{(i)}, b^{(i)}]$ the subdiagonal elements in the i-th 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

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

Here, the subdiagonal elements $\lambda_{k+1,k}, \ldots, \lambda_{nk}$ in the k-th 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*

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

arithmetic 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 $n^2 + O(n)$ arithmetic operations.

Proof. The k-th 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 $\boxed{\cdot}$.

$$\begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 7 \\ 4 \end{bmatrix} \quad \rightarrow$$

pivoting

$\boxed{3}$	1	6	2
2	1	3	7
1	1	1	4

elimination

3	1	6	2
2/3	1/3	−1	17/3
1/3	$\boxed{2/3}$	−1	10/3

$\rightarrow$

pivoting

3	1	6	2
1/3	2/3	−1	10/3
2/3	1/3	−1	17/3

elimination

3	1	6	2
1/3	2/3	−1	10/3
2/3	1/2	−1/2	4

$\rightarrow$

$$\begin{aligned} x_3 &= -8 \\ x_2 &= \tfrac{3}{2}(\tfrac{10}{3} - x_3) = -7 \\ x_1 &= \tfrac{1}{3}(2 - x_2 - 6x_3) = 19\,. \end{aligned}$$

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	
$0.1\cdot10^{-3}$	$0.1\cdot10^{1}$	$0.1\cdot10^{1}$
0	$-0.1\cdot10^{5}$	$-0.1\cdot10^{5}$
$x_2 = 1\,,$	$x_1 = 0$	

b) with pivoting:

x_1	x_2	
$0.1\cdot10^{1}$	$0.1\cdot10^{1}$	$0.2\cdot10^{1}$
0	$0.1\cdot10^{1}$	$0.1\cdot10^{1}$
$x_2 = 1\,,$	$x_1 = 1$	

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,

$$Ax = b \quad \to \quad DAx = Db\,, \qquad d_i = \Big(\sum_{j=1}^{n} |a_{ij}|\Big)^{-1}, \tag{2.1.9}$$

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 (see Wilkinson[2] [23]).

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")*

$$\frac{\|\delta A\|_\infty}{\|A\|_\infty} \le 1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\mathrm{eps}. \tag{2.1.10}$$

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:

$$\frac{\|\delta x\|_\infty}{\|x\|_\infty} \le \frac{\mathrm{cond(A)}}{1 - \mathrm{cond(A)}\|\delta \mathrm{A}\|_\infty/\|\mathrm{A}\|_\infty}\left\{1.01 \cdot 2^{n-1}(n^3 + 2n^2)\,\mathrm{eps}\right\}. \tag{2.1.11}$$

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 cond(A), which described the conditioning of the numerical problem to be solved.

Direct LR and Cholesky decomposition

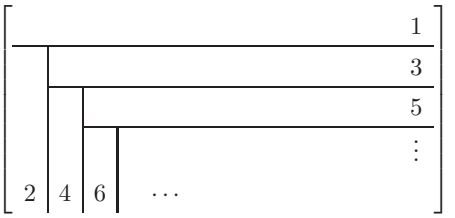
The Gaussian algorithm for the computation of the LR decomposition $A = LR$ (if it exists) can also be written in direct form, in which the elements l_{jk} of L and r_{jk} of

[2] 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).

R are computed recursively. The equation $A = LR$ yields n^2 equations for the n^2 unknown elements $r_{jk}\,,\ j \le k\,,\ l_{jk}\,,\ j > k\,(l_{jj} = 1)\,:$

$$a_{jk} = \sum_{i=1}^{\min(j,k)} l_{ji}\, r_{ik}. \tag{2.1.12}$$

Here, the ordering of the computation of l_{jk}, r_{jk} is not prescribed a priori. In the so-called "algorithm of Crout[3]" the matrix $A = LR$ is tessellated as follows:



The single steps of this algorithm are $(l_{ii} \equiv 1)\,:$

$$\begin{aligned}
&k = 1, \cdots, n\ : && a_{1k} = \sum_{i=1}^{1} l_{1i} r_{ik} \ \Rightarrow\ r_{1k} := a_{1k},\\
&j = 2, \cdots, n\ : && a_{j1} = \sum_{i=1}^{1} l_{ji} r_{i1} \ \Rightarrow\ l_{j1} := r_{11}^{-1} a_{j1},\\
&k = 2, \cdots, n\ : && a_{2k} = \sum_{i=1}^{2} l_{2i} r_{ik} \ \Rightarrow\ r_{2k} := a_{2k} - l_{21} r_{1k},\\
&\vdots
\end{aligned}$$

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 arithmetic operations and are algebraically equivalent.

[3]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.

2.1.2 Accuracy improvement by defect correction

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) "residual" (negative "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 "correction 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 correction 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 residual (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

$$\frac{\|x^{(0)} - x\|}{\|x\|} \leq \frac{\mathrm{cond(A)}}{1 - \mathrm{cond(A)}\frac{\|\mathrm{A}-\tilde{\mathrm{L}}\tilde{\mathrm{R}}\|}{\|\mathrm{A}\|}} \underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim\,\varepsilon}\,.$$

Here, the loss of exact decimals corresponds to the condition $\mathrm{cond(A)}$. Additionally round-off errors are neglected. The exact residual $\hat{d}^{(0)}$ is replaced by the expression

$d^{(0)} := \tilde{A}x^{(0)} - b$ 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

$$\begin{aligned} \|(\tilde{L}\tilde{R})^{-1}\| &\leq \|A^{-1}\|\,\|\,[I - A^{-1}(A - \tilde{L}\tilde{R})]^{-1}\| \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}(A - \tilde{L}\tilde{R})\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\,\|A - \tilde{L}\tilde{R}\|} = \frac{\|A^{-1}\|}{1 - \mathrm{cond}(A)\,\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}. \end{aligned}$$

This eventually implies

$$\frac{\|x^{(1)} - x\|}{\|x\|} \;\sim\; \mathrm{cond}(A)\,\Big[\underbrace{\frac{\|A - \tilde{L}\tilde{R}\|}{\|A\|}}_{\sim\,\varepsilon}\;\underbrace{\frac{\|x^{(0)} - x\|}{\|x\|}}_{\sim\,\mathrm{cond(A)}\varepsilon} + \underbrace{\frac{\|A - \tilde{A}\|}{\|A\|}}_{\sim\,\tilde{\varepsilon}}\;\frac{\|x^{(0)}\|}{\|x\|}\Big].$$

This correction procedure can be iterated to a "defect correction" 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 residual

$$d^{(0)} = b - Ax^{(0)} = \begin{cases} (0,0)^T & \text{3-decimal computation,} \\ (0,065, 0,035)^T & \text{6-decimal computation.} \end{cases}$$

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):

$$\begin{array}{c|ccc} & 1 & & 0 \\ A & & \ddots & \\ & 0 & & 1 \end{array} \quad \rightarrow \quad \overset{\text{forward elimination}}{\begin{array}{ccc|ccc} r_{11} & \cdots & r_{1n} & 1 & & 0 \\ & \ddots & \vdots & & \ddots & \\ & & r_{nn} & * & & 1 \end{array}}$$

$$\begin{array}{c|c} \text{backward elimination} & \\ \hline \begin{matrix} r_{11} & & 0 \\ & \ddots & \\ 0 & & r_{nn} \end{matrix} & * \end{array} \quad \rightarrow \quad \begin{array}{c|c} \text{scaling} & \\ \hline \begin{matrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{matrix} & A^{-1} \end{array}$$

Example 2.5: The pivot elements are marked by $\boxed{\cdot}$.

$$A = \begin{bmatrix} 3 & 1 & 6 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \quad : \quad \begin{array}{c} \text{forward elimination} \\ \hline \boxed{3} \;\; 1 \;\; 6 \;\vert\; 1 \;\; 0 \;\; 0 \\ 2 \;\; 1 \;\; 3 \;\vert\; 0 \;\; 1 \;\; 0 \\ 1 \;\; 1 \;\; 1 \;\vert\; 0 \;\; 0 \;\; 1 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{c} \text{row permutation} \\ \hline 3 \;\; 1 \;\; 6 \;\vert\; 1 \;\; 0 \;\; 0 \\ 0 \;\; 1/3 \;\; -1 \;\vert\; -2/3 \;\; 1 \;\; 0 \\ 0 \;\; \boxed{2/3} \;\; -1 \;\vert\; -1/3 \;\; 0 \;\; 1 \end{array} \quad \rightarrow \quad \begin{array}{c} \text{forward elimination} \\ \hline 3 \;\; 1 \;\; 6 \;\vert\; 1 \;\; 0 \;\; 0 \\ 0 \;\; 2/3 \;\; -1 \;\vert\; -1/3 \;\; 0 \;\; 1 \\ 0 \;\; 1/3 \;\; -1 \;\vert\; -2/3 \;\; 1 \;\; 0 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{c} \text{backward elimination} \\ \hline 3 \;\; 1 \;\; 6 \;\vert\; 1 \;\; 0 \;\; 0 \\ 0 \;\; 2/3 \;\; -1 \;\vert\; -1/3 \;\; 0 \;\; 1 \\ 0 \;\; 0 \;\; -1/2 \;\vert\; -1/2 \;\; 1 \;\; -1/2 \end{array} \quad \rightarrow \quad \begin{array}{c} \text{backward elimination} \\ \hline 3 \;\; 1 \;\; 0 \;\vert\; -5 \;\; 12 \;\; -6 \\ 0 \;\; 2/3 \;\; 0 \;\vert\; 2/3 \;\; -2 \;\; 2 \\ 0 \;\; 0 \;\; -1/2 \;\vert\; -1/2 \;\; 1 \;\; -1/2 \end{array} \quad \rightarrow$$

$$\rightarrow \quad \begin{array}{c} \text{scaling} \\ \hline 3 \;\; 0 \;\; 0 \;\vert\; -6 \;\; 15 \;\; -9 \\ 0 \;\; 2/3 \;\; 0 \;\vert\; 2/3 \;\; -2 \;\; 2 \\ 0 \;\; 0 \;\; -1/2 \;\vert\; -1/2 \;\; 1 \;\; -1/2 \end{array} \quad \rightarrow \quad \begin{array}{c} \\ \hline 1 \;\; 0 \;\; 0 \;\vert\; -2 \;\; 5 \;\; -3 \\ 0 \;\; 1 \;\; 0 \;\vert\; 1 \;\; -3 \;\; 3 \\ 0 \;\; 0 \;\; 1 \;\vert\; 1 \;\; -2 \;\; 1 \end{array}$$

$$\Rightarrow \quad A^{-1} = \begin{bmatrix} -2 & 5 & -3 \\ 1 & -3 & 3 \\ 1 & -2 & 1 \end{bmatrix} .$$

An alternative method for computing the inverse of a matrix is the so-called "exchange algorithm" (sometimes called "Gauß-Jordan 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 p-th 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 choose 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.*

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

$$\begin{matrix} r \\ m-r \end{matrix}\left\{\begin{array}{c|c} * & * \\ \hline * & 0 \end{array}\right] \begin{bmatrix} y_1 \\ \vdots \\ y_r \\ x_{r+1} \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_r \\ y_{r+1} \\ \vdots \\ y_m \end{bmatrix} .$$

$$\underbrace{}_{r} \quad \underbrace{}_{n-r}$$

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 k-th 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 $(4, 4)$:

$$A = \left[\begin{array}{cccc} B & -I & & \\ -I & B & -I & \\ & -I & B & -I \\ & & -I & B \end{array}\right]\Bigg\}16, \quad B = \left[\begin{array}{cccc} 4 & -1 & & \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ & & -1 & 4 \end{array}\right]\Bigg\}4, \quad I = \left[\begin{array}{cccc} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{array}\right]\Bigg\}4.$$

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

$$L = \begin{bmatrix} 1 & & & \\ \gamma_2 & \ddots & & \\ & \ddots & 1 & \\ & & \gamma_n & 1 \end{bmatrix}, \qquad R = \begin{bmatrix} \alpha_1 & \beta_1 & & \\ & \ddots & \ddots & \\ & & \alpha_{n-1} & \beta_{n-1} \\ & & & \alpha_n \end{bmatrix}$$

are simply be obtained by short recursion formulas (sometimes called "Thomas[4] algorithm"),

$$\begin{array}{rrcl} & & \alpha_1 = a_1 & , \quad \beta_1 = b_1 \,, \\ i = 2, \ldots, n-1 : & \gamma_i = c_i/\alpha_{i-1} & , \quad \alpha_i = a_i - \gamma_i \beta_{i-1} & , \quad \beta_i = b_i \,, \\ & \gamma_n = c_n/\alpha_{n-1} & , \quad \alpha_n = a_n - \gamma_n \beta_{n-1} & . \end{array}$$

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.*

[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.

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

$$\begin{aligned}
\sum_{k=2,k\neq j}^{n} |a_{jk}^{(1)}| &\leq \sum_{k=2,k\neq j}^{n} |a_{jk}| + |q_{j1}| \sum_{k=2,k\neq j}^{n} |a_{1k}| \\
&\leq \underbrace{\sum_{k=1,k\neq j}^{n} |a_{jk}| - |a_{j1}|}_{\leq\, |a_{jj}|} + \underbrace{|q_{j1}|}_{= \left|\frac{a_{j1}}{a_{11}}\right|} \underbrace{\sum_{k=2}^{n} |a_{1k}| - |q_{j1}||a_{1j}|}_{\leq\, |a_{11}|} \\
&\leq |a_{jj}| - |q_{j1} a_{1j}| \leq |a_{jj} - q_{j1} a_{1j}| = |a_{jj}^{(1)}|.
\end{aligned}$$

The matrix $A^{(1)} = G_1 A^{(0)}$ 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,

$$\begin{aligned} 0 < \sum_{j,k=1}^{n} a_{jk}x_jx_k &= \sum_{j,k=2}^{n} a_{jk}x_jx_k + 2x_1 \sum_{k=2}^{n} a_{1k}x_k + a_{11}x_1^2 \\ &\quad \underbrace{- \frac{1}{a_{11}} \sum_{j,k=2}^{n} a_{k1}a_{1j}x_kx_j + \frac{1}{a_{11}} \Big(\sum_{k=2}^{n} a_{1k}x_k \Big)^2}_{= 0 \ (a_{jk} = a_{kj})} \\ &= \sum_{j,k=2}^{n} \underbrace{\Big(a_{jk} - \frac{a_{k1}a_{1j}}{a_{11}} \Big)}_{= a_{jk}^{(1)}} x_jx_k + a_{11} \underbrace{\Big(x_1 + \frac{1}{a_{11}} \sum_{k=2}^{n} a_{1k}x_k \Big)^2}_{= 0} \end{aligned}$$

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*

$$N_{Cholesky}(n) = \tfrac{1}{6}n^3 + O(n^2) \quad a.\ op. \tag{2.2.23}$$

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 $\tilde{L}$:

$$\tilde{l}_{11}^2 = a_{11}\,, \quad \tilde{l}_{21}\,\tilde{l}_{11} = a_{21}\,, \quad \ldots\,, \quad \tilde{l}_{n1}\,\tilde{l}_{11} = a_{n1}\,,$$

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,

$$\begin{aligned} \tilde{l}_{ii} &= \sqrt{a_{ii} - \tilde{l}_{i1}^2 - \tilde{l}_{i2}^2 - \ldots - \tilde{l}_{i,i-1}^2}\,, \\ \tilde{l}_{ji} &= \tilde{l}_{ii}^{-1}\big\{a_{ji} - \tilde{l}_{j1}\tilde{l}_{i1} - \tilde{l}_{j2}\tilde{l}_{i2} - \ldots - \tilde{l}_{j,i-1}\tilde{l}_{i,i-1}\big\}, \quad j = i+1, \ldots, n, \end{aligned}$$

[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".*

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 $\text{rank}(A) = \text{rank}([A, b])$. 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, $\text{rank}(A) < n$, a particular solution $\hat{x}$ of the normal system is not unique, but of the general form $\hat{x} + y$ with any element $y \in \text{kern}(A)$. 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 $A^T A$ is symmetric and positive semi-definite, and even positive definite if A has maximal rank $\text{rank}(A) = n$. 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 $A^T A$ is rather ill-conditioned. In fact, for $m = n$, we have that

$$\text{cond}_2(\text{A}^{\text{T}}\text{A}) \sim \text{cond}_2(\text{A})^2. \tag{2.3.27}$$

Example 2.9: Using 3-decimal arithmetic, we have

$$A = \begin{bmatrix} 1.07 & 1.10 \\ 1.07 & 1.11 \\ 1.07 & 1.15 \end{bmatrix} \quad \rightarrow \quad A^T A = \begin{bmatrix} 3.43 & 3.60 \\ 3.60 & 3.76 \end{bmatrix}.$$

But $A^T A$ is *not* positive definite: $(-1,1) \cdot A^T A \cdot (-1,1)^T = -0.01$, 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 $A^T A$. 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 $\operatorname{rank}(A) = n$. Then, there exists a uniquely determined orthonormal matrix $Q \in \mathbb{K}^{m\times n}$ with the property*

$$\bar{Q}^T Q = I \quad (\mathbb{K} = \mathbb{C}), \qquad Q^T Q = I \quad (\mathbb{K} = \mathbb{R}), \tag{2.3.28}$$

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 $\operatorname{rank}(A) = n$ 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 $A = Q_1 R_1$ and

$A = Q_2 R_2$ be two such decompositions. Since R_1 and R_2 are regular and $(\det(R_i) > 0)$ it follows that

$$\begin{array}{rcl} Q := \bar{Q}_2^T Q_1 = R_2 R_1^{-1} & \text{right upper triangular,} \\ \bar{Q}^T = \bar{Q}_1^T Q_2 = R_1 R_2^{-1} & \text{right upper triangular.} \end{array}$$

Since $\bar{Q}^T Q = R_1 R_2^{-1} R_2 R_1^{-1} = I$ 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 $A^T Ax = A^T b$ transforms into

$$A^T Ax = R^T Q^T QRx = R^T Rx = R^T Q^T b,$$

and, consequently, in view of the regularity of R^T,

$$Rx = Q^T b. \tag{2.3.30}$$

This triangular system can now be solved by backward substitution in $\mathcal{O}(n^2)$ arithmetic operations. Since

$$A^T A = R^T R \tag{2.3.31}$$

with the triangular matrix R, we are given a Cholesky decomposition of $A^T A$ without explicit computation of the matrix product $A^T A$.

Example 2.10: The 3×3-matrix

$$A = \begin{bmatrix} 12 & -51 & 4 \\ 6 & 167 & -68 \\ -4 & 24 & -41 \end{bmatrix}$$

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, $\mathbb{K} = \mathbb{R}$. 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 followimg form:

[6] Alston Scott Householder (1904–1993): US-American mathematician; Director of Oak Ridge National Laboratory (1948-1969), thereafter Prof. at the Univ. of Tennessee; worked in mathematical biology, best known for his fundamental contributions to numerics, especially to numerical linear algebra.

$$A^{(i-1)} = \left[\begin{array}{ccc|c|ccc} * & & & & & & * \\ & \ddots & & & & & \vdots \\ & & * & & & & \\ \hline & & & * & \cdots & * & \\ & 0 & & & & & \\ & & & * & \cdots & * & \end{array}\right] \begin{matrix} \\ \\ \\ i \\ \\ \\ \end{matrix}$$

$$\begin{matrix} & & & & & & & & i \end{matrix}$$

In the i-th 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

$$\tilde{R} = \left[\begin{array}{ccc} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \\ \hline 0 & \cdots & 0 \end{array}\right] \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \text{n} \\ \left.\begin{matrix} \\ \end{matrix}\right\} \text{m}. \end{matrix}$$

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 $\tilde{R}$:

$$A = \underbrace{\left[\begin{array}{c|c} & \\ Q & * \\ & \end{array}\right]}_{\text{n}\quad\text{m - n}} \cdot \left[\begin{array}{c} R \\ \hline 0 \end{array}\right] \begin{matrix} \left.\begin{matrix} \\ \end{matrix}\right\} \text{n} \\ \left.\begin{matrix} \\ \end{matrix}\right\} \text{m} - \text{n} \end{matrix} = QR.$$

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 $S_1 a_1 \in \text{span}\{e_1\}$. The vector a_1 is reflected with respect to one of the axes $\text{span}\{a_1 + \|a_1\|e_1\}$ or $\text{span}\{a_1 - \|a_1\|e_1\}$) into the x_1-axis. The choice of the axis is oriented by $\text{sgn}(a_{11})$ 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

$$a_1^{(1)} = -\|a_1\|_2 e_1, \quad a_k^{(1)} = a_k - 2(a_k, v_1)v_1, \quad k = 2, \ldots, n.$$

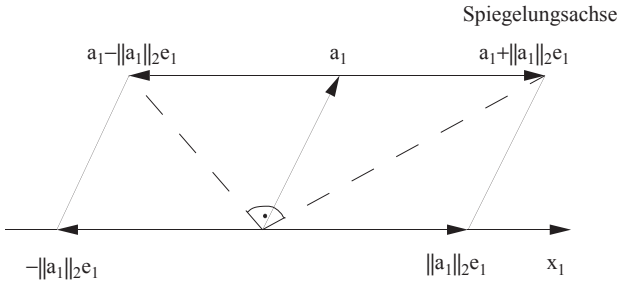


Figure 2.1: *Scheme of the Householder transformation*

Let now the transformed matrix $A^{(i-1)}$ be already computed.

***i*-th step**: For S_i we make the following ansatz:

$$S_i = \underbrace{\begin{bmatrix} I & 0 \\ \hline 0 & I - 2\tilde{v}_i\bar{\tilde{v}}_i^T \end{bmatrix}}_{\text{i - 1}} = I - 2v_i\bar{v}_i^T, \qquad v_i = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tilde{v}_i \end{bmatrix} \begin{matrix} \Big\} \text{i - 1} \\ \Big\} \text{m} \end{matrix}$$

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 1-st 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 by Deuflhard & Hohmann [33].

Let $A \in \mathbb{K}^{m\times n}$ be given. Further let $Q \in \mathbb{K}^{m\times m}$ and $Z \in \mathbb{K}^{n\times n}$ 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*

$$A = U\Sigma\bar{V}^T, \quad \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m\times n}, \quad p = \min(m, n), \tag{2.4.33}$$

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

$$\left(\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_m & \end{array}\right) \qquad or \qquad \left(\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ \hline & 0 & \end{array}\right).$$

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 $\Lambda = \mathrm{diag}(\lambda_i)$, 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. The σ_i are the so-called "singular values" of the matrix A . 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):

- $\mathrm{rank}(A) = r$,
- $\mathrm{kern}(A) = \mathrm{span}\{v^{r+1}, \ldots, v^n\}$,
- $\mathrm{range}(A) = \mathrm{span}\{u^1, \ldots, u^r\}$,
- $A = U_r \Sigma_r V_r^T \equiv \sum_{i=1}^r \sigma_i u^i v^{iT}$ (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

$$\mathrm{rank}(A, \varepsilon) = \min_{\|A-B\|_2 \leq \varepsilon} \mathrm{rank}(B).$$

The matrix is called "numerically rank-deficient" if

$$\operatorname{rank}(A,\varepsilon) < \min(m,n)\,, \qquad \varepsilon = \operatorname{eps}\|A\|_2,$$

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 $k < r = \operatorname{rank}(A)$, then in the truncated singular value decomposition,*

$$A_k = \sum_{i=1}^{k} \sigma_i u^i v^{iT},$$

there holds the estimate

$$\min_{\operatorname{rank}(B)=k} \|A-B\|_2 = \|A-A_k\|_2 = \sigma_{k+1}\,.$$

This implies for $r_\varepsilon = \operatorname{rank}(A,\varepsilon)$ the relation

$$\sigma_1 \geq \cdots \geq \sigma_{r_\varepsilon} > \varepsilon \geq \sigma_{r_\varepsilon+1} \geq \cdots \geq \sigma_p\,, \quad p = \min(m,n).$$

Proof. Since

$$U^T A_k V = \operatorname{diag}(\sigma_1,\ldots,\sigma_k,0,\ldots,0)$$

it follows that $\operatorname{rank}(A_k) = k$. Further, we obtain

$$U^T(A-A_k)V = \operatorname{diag}(0,\ldots,0,\sigma_{k+1},\ldots,\sigma_p)$$

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

$$\|A-B\|_2 \geq \sigma_{k+1}.$$

To this end, we choose an ONB $\{x^1,\ldots,x^{n-k}\}$ of $\operatorname{kern}(B)$. For dimensional reasons there obviously holds

$$\operatorname{span}\{x^1,\ldots,x^{n-k}\} \cap \operatorname{span}\{v^1,\ldots,v^{k+1}\} \neq \emptyset.$$

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 $z = \sum_{i=1}^{k+1} (v^{iT} z) v^i$ 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 x,

$$\|Ax - b\|_2 = \min!$$

necessarily solves the normal equation $A^T Ax = A^T b$. But this solution is unique only in the case of maximal $\operatorname{rank}(A) = n$, 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 $\Sigma = \operatorname{diag}(\sigma_i)$, $\sigma_i := \lambda_i^{1/2}$, $V = [v^1, \ldots, v^n]$, $u^i := \lambda_i^{-1/2} A v^i$, 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 $\operatorname{rank}(A) < n$ 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

$$\min_{X\in\mathbb{R}^{n\times m}} \|AX - I\|_F,$$

with the Frobenius norm $\|\cdot\|_F$. Since the identity in (2.4.37) holds for all b it follows that

[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.

$$\begin{aligned} \operatorname{rank}(A) = n \quad &\Rightarrow \quad A^+ = (A^T A)^{-1} A^T, \\ \operatorname{rank}(A) = n = m \quad &\Rightarrow \quad A^+ = A^{-1}. \end{aligned}$$

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 by Golub & van Loan [36].

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

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

holds

$$Aw = T^{-1}BTw = \lambda w,$$

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*

[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.

$$\bar{U}^T A U = \begin{bmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}. \tag{2.5.40}$$

If $A \in \mathbb{C}^{n\times n}$ is Hermitian, $A^T = \bar{A}$, so is also $\bar{U}^T A U$ Hermitian. Hence, Hermitian matrices $A \in \mathbb{C}^{n\times n}$ are "unitary similar" to a diagonal matrix $\bar{U}^T A U = \mathrm{diag}(\lambda_i)$, i. e., "diagonalizable".

Lemma 2.3 (Diagonalization): *For any matrix $A \in \mathbb{C}^{n\times n}$ the following statements are equivalent:*

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

$$A = A^{(0)} \to A^{(1)} = T_1^{-1} A^{(0)} T_1 \to \ldots A^{(m)} = T_m^{-1} A^{(m-1)} T_m.$$

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 any 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,

$$\|B\| \le \mathrm{cond(T)}\, \|\mathrm{A}\|, \quad \|\delta \mathrm{A}\| \le \mathrm{cond(T)}\, \|\delta \mathrm{B}\|.$$

This implies that

$$\frac{\|\delta A\|}{\|A\|} \le \mathrm{cond(T)}^2\, \frac{\|\delta \mathrm{B}\|}{\|\mathrm{B}\|}. \tag{2.5.41}$$

Hence, for large $\mathrm{cond(T)} \gg 1$ 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

$$\mathrm{cond(T)} = \mathrm{cond(T_1 \ldots T_m)} \le \mathrm{cond(T_1)} \cdot \ \ldots \ \cdot \mathrm{cond(T_m)},$$

the transformation matrices T_i are to be chosen such that $\mathrm{cond(T_i)}$ does not become too large. This is especially achieved for the following three types of transformations:

a) Rotations (Givens transformation):

$$T = \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos(\varphi) & & & -\sin(\varphi) & & & \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & \sin(\varphi) & & & \cos(\varphi) & & & \\ & & & & & & & 1 & & \\ & & & & & & & & \ddots & \\ & & & & & & & & & 1 \end{bmatrix} \implies \mathrm{cond}_2(\mathrm{T}) = 1.$$

b) Reflections (Householder transformation):

$$T = I - 2u\bar{u}^T \implies \mathrm{cond}_2(\mathrm{T}) = 1.$$

The Givens and the Householder transformations are *unitary* with spectral condition $\mathrm{cond}_2(\mathrm{T}) = 1$.

c) Elimination

$$T = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{i+1,i} & 1 & \\ & & \vdots & & \ddots \\ & & l_{n,i} & & 1 \end{bmatrix}, \quad |l_{jk}| \leq 1 \implies \mathrm{cond}_\infty(\mathrm{T}) \leq 4.$$

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,

$$A^{(1)} = T_1 A T_1 = \underbrace{\left[\begin{array}{c|ccc} a_{11} & a_{12} & \dots & a_{1n} \\ \hline //// & * & & \\ \hline 0 & & & \end{array}\right]}_{T_1 A} \underbrace{\left[\begin{array}{c|ccc} 1 & 0 & \dots & \\ \hline 0 & * & & \\ \vdots & & & \end{array}\right]}_{T_1^T} = \left[\begin{array}{c|c} a_{11} & * \\ \hline //// & \\ \hline 0 & \tilde{A}^{(1)} \end{array}\right].$$

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 $A^{(1)} = T_1 A T_1$ 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 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 form goes back to Givens[10] (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 by Stoer & Bulirsch II [50].

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[11] (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., $a_{j+1,j} \neq 0$, $j = 1, \dots, n-1$. With a function $c(\cdot)$ still to be chosen, we consider the linear system

$$\begin{aligned} (a_{11} - z)x_1 + a_{12}x_2 + \cdots + a_{1,n-1}x_{n-1} + a_{1n}x_n &= -c\,(z) \\ a_{21}x_1 + (a_{22} - z)x_2 + \cdots + a_{2,n-1}x_{n-1} + a_{2n}x_n &= 0 \\ &\vdots \\ a_{n,n-1}x_{n-1} + (a_{nn} - z)x_n &= 0. \end{aligned}$$

[10] 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).

[11] Morton Allan Hyman: Dutch mathematician; PhD Techn. Univ. Delft 1953, *Eigenvalues and eigenvectors of general matrices*, Twelfth National Meeting A.C.M., Houston, Texas, 1957.

Setting $x_n = 1$ the values $x_{n-1}, \dots, x_1$ and $c(z)$ can be successively determined. By Cramer's rule there holds

$$1 = x_n = \frac{(-1)^n c(z) a_{21} a_{32} \dots a_{n,n-1}}{\det(A - zI)}.$$

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 $b_i \neq 0$, $i = 1, \dots, n-1$:

$$A = \begin{bmatrix} a_1 & b_1 & & 0 \\ b_1 & & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ 0 & & b_{n-1} & a_n \end{bmatrix}.$$

For the computation of the characteristic polynomial $\chi_A(\cdot)$, we have the recursion formulas

$$p_0(z) = 1, \quad p_1(z) = a_1 - z, \qquad p_i(z) = (a_i - z)\, p_{i-1}(z) - b_{i-1}^2 p_{i-2}(z), \quad i = 2, \dots, n.$$

The polynomials $p_i \in P_i$ are the i-th principle minors of $\det(zI - A)$, i. e., $p_n = \chi_A$. To see this, we expand the $(i+1)$-th principle minor with respect to the $(i+1)$-th column:

$$\begin{bmatrix} a_1 - z & b_1 & & & & \\ b_1 & \ddots & \ddots & & & \\ \ddots & & & & & \\ & & & b_{i-1} & & \\ & & b_{i-1} & a_i - z & b_i & \\ & & & b_i & a_{i+1} - z & \ddots \\ & & & & \ddots & \ddots \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} = \underbrace{(a_{i+1} - z) p_i(z) - b_i^2 p_{i-1}(z)}_{=:\, p_{i+1}(z)}.$$

$$\begin{matrix} & & i-1 & i & i+1 & \end{matrix}$$

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:* $p_{n-1}(\lambda)p_n'(\lambda) < 0$.

(S3) *For each real zero ζ of p_{i-1}:* $p_i(\zeta)p_{i-2}(\zeta) < 0,\ i = 2, \dots, n;$
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[12] chain" of p.*

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.

[12] *Jacques Charles Franois Sturm (1803–1855): French-Swiss mathematician; Prof. at École Polytechnique in Paris since 1840; contributions to Mathematical Physics, differential equations, ("Sturm-Liouville problem") and Differential Geometry.*

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 $p_j(a)$, $j \in \{i-1, i, i+1\}$ for sufficiently small $h > 0$ shows 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

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n.$$

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 i-th 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 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:

$$k = 1, \ldots, n-1: \qquad a_{ii}^{(k)} \leq a_{ii}^{(k-1)}, \quad i = 1, \ldots, n, \quad \max_{1\leq i,j\leq n} |r_{ij}| \leq \max_{1\leq i,j\leq n} |a_{ij}|.$$

(Hint: Use the recursion formula and employ an induction argument.)

Exercise 2.3: The "LR decomposition" of a regular matrix $A \in \mathbb{R}^{n\times n}$ is the represen-

tation of A as a product $A = LR$ consisting of a lower-left triangular matrix L with normalized diagonal ($l_{ii} = 1, 1 \leq i \leq n$) 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 text it is 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 text 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?