1.1 Direct Methods for Linear Algebraic Systems

Problem: Given an \((n, n)\)-matrix \(A \in \mathbb{R}^{n \times n}\), \(A = (a_{ij})_{i,j=1}^n\), and a vector \(b \in \mathbb{R}^n\), \(b = (b_1, \ldots, b_n)^T\). Find a vector \(x \in \mathbb{R}^n\), \(x = (x_1, \ldots, x_n)^T\), such that

\[
(*) \quad A \cdot x = b,
\]

or explicitly

\[
\begin{pmatrix}
  a_{11} & a_{12} & \cdots & a_{1n} \\
  a_{21} & a_{22} & \cdots & a_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  a_{n1} & a_{n2} & \cdots & a_{nn}
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n
\end{pmatrix}
= 
\begin{pmatrix}
  b_1 \\
  b_2 \\
  \vdots \\
  b_n
\end{pmatrix}.
\]

Definition 1.1 Linear Algebraic System
Equation \((*)\) is called a linear algebraic system of order \(n\). \(A\) is said to be the coefficient matrix and \(b\) the right-hand side of the linear system. The vector \(x\) is referred to as the vector of unknowns.
Theorem 1.2 Existence and Uniqueness
Let $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. If $\det A \neq 0$, then there exists a unique $x \in \mathbb{R}^n$ such that $Ax = b$.

Theorem 1.3 Cramer’s Rule (only of theoretical interest)
Let $A \in \mathbb{R}^{n \times n}$ such that $\det A \neq 0$ and $b \in \mathbb{R}^n$.
Denote by $A_{i,b}$, $1 \leq i \leq n$, the matrix obtained from $A$ by replacing the $i$-th column by the vector $b$. Then, the solution $x \in \mathbb{R}^n$ of $Ax = b$ is given by

$$x_i = \frac{\det A_{i,b}}{\det A}, \quad 1 \leq i \leq n.$$

Definition 1.4 Computational Work, Complexity, Efficiency
The computational work for the realization of an algorithm is the number of basic arithmetical operations $+, -, \times, /$.
The complexity of a problem is the minimal required computational work for the computation of the solution.
An algorithm is the more efficient, the closer the computational work is to the complexity of the problem.
Remark 1.5  Computational work for Cramer’s rule
The computation of the determinant \( \det A \) of a regular matrix \( A \in \mathbb{R}^{n \times n} \) by means of
\[
\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij}
\]
requires approximately \( 2^n \) essential operations, where \( A_{ij} \in \mathbb{R}^{(n-1) \times (n-1)} \) denotes the matrix obtained from \( A \) by eliminating the \( i \)-th row and the \( j \)-th column.
The computational work grows exponentially with the dimension of the problem and is significantly
greater as for the algorithms to be presented next.
Therefore, never solve a linear algebraic system by Cramer’s rule.

Remark 1.6  Computation of the inverse
Denoting by \( A^{-1} \in \mathbb{R}^{n \times n} \) the inverse of the regular matrix \( A \in \mathbb{R}^{n \times n} \), the solution \( x \in \mathbb{R}^n \) of \( Ax = b \) is formally given by
\[
x = A^{-1} b
\]
The strategy to invert \( A \) numerically, i.e., to compute \( A^{-1} \) and then form the matrix-vector product of \( A^{-1} \) and \( b \), can not be recommended (e.g., the computation of \( A^{-1} \) can be ’numerically unstable’, whereas the computation of \( x \) such that \( Ax = b \) turns out to be ’numerically stable’).
Remark 1.7 Special cases:

(i) Diagonal matrix (decoupled system)

\[ A = (a_{ij} \delta_{ij})_{i,j=1}^{n} \quad [ \text{notation: } A = \text{diag}(a_{ii}) ] \]

For regular $A$, the solution of $Ax = b$ is given by $x_i = b_i/a_{ii}$, $1 \leq i \leq n$.

The computational work comprises $n$ divisions. This is also a lower bound for the complexity of the problem.

(ii) Staggered linear systems

Definition 1.7 Lower, upper triangular matrix / Staggered systems

A matrix $R = (r_{ij})_{i,j=1}^{n}$ resp. $L = (\ell_{ij})_{i,j=1}^{n}$ is called an upper resp. a lower triangular matrix, if

\[ r_{ij} = 0 \quad i > j \quad \text{resp.} \quad \ell_{ij} = 0 \quad i < j. \]

A linear system of the form $Rx = b$ or $Lx = b$ is called a staggered linear system.
Lemma 1.1.8  Solution of staggered linear systems

Consider the staggered linear system

\[
\begin{pmatrix}
  r_{11} & r_{12} & \cdots & r_{1,n-1} & r_{1n} \\
  0 & r_{22} & \cdots & r_{2,n-1} & r_{2n} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  \vdots & \vdots & \cdots & 0 & r_{n-1,n-1} \\
  0 & 0 & \cdots & 0 & r_{nn}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
x_3 \\
\vdots \\
x_{n-1} \\
x_n
\end{pmatrix}
=
\begin{pmatrix}
b_1 \\
b_2 \\
b_3 \\
\vdots \\
b_{n-1} \\
b_n
\end{pmatrix}
\]

Since \( \det R = \prod_{i=1}^{n} r_{ii} \), we have \( \det R \neq 0 \) if and only if \( r_{ii} \neq 0, \; 1 \leq i \leq n \). The solution can be obtained recursively, starting with the \( n \)-th row:

\[
x_n = \frac{b_n}{r_{nn}}, \\
x_{n-1} = \frac{b_{n-1} - r_{n-1,n} x_n}{r_{n-1,n-1}}, \\
x_{n-i} = \frac{b_i - \sum_{j=0}^{i-1} r_{n-i,n-j} x_{n-j}}{r_{n-i,n-i}}, \; 0 \leq i \leq n-1.
\]
Lemma 1.9  Computational work for the solution of staggered systems
Computational work for the solution of $x_{n-i}$:

\[ i \text{ multiplications and addition }, \quad 1 \text{ division} \]

Total:
\[ n + 2 \sum_{i=0}^{n-1} i = n + 2 \frac{n(n-1)}{2} = n^2. \]

Remark 1.10: Similar considerations apply to a staggered linear system
with a lower triangular matrix $L \in \mathbb{R}^{n \times n}$ (starting with the first row)

\[
\begin{pmatrix}
\ell_{11} & 0 & 0 & \cdots & 0 \\
\ell_{21} & \ell_{22} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & \ell_{nn}
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{pmatrix}
= 
\begin{pmatrix}
b_1 \\
b_2 \\
\vdots \\
b_n
\end{pmatrix}.
\]

Definition 1.11  Forward/backward substitution
The solution of a staggered linear system is called backward substitution in case of an upper triangular matrix and forward substitution for a lower triangular matrix.
Idea behind Gaussian elimination:
Transformation of a linear algebraic system to a staggered linear system per triangular matrix by successive elimination of the variables $x_i$, $1 \leq i \leq n - 1$ from row $i + 1$ to row $n$:

$$
\begin{pmatrix}
  a_{11} & a_{12} & \cdots & a_{1n} \\
  a_{21} & a_{22} & \cdots & a_{2n} \\
  \quad & \quad & \ddots & \cdots \\
  a_{n1} & a_{n2} & \cdots & a_{nn}
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2 \\
  \quad \\
  x_n
\end{pmatrix}
= 
\begin{pmatrix}
  b_1 \\
  b_2 \\
  \quad \\
  b_n
\end{pmatrix}
\quad \Rightarrow
\begin{pmatrix}
  a_{11}^n & a_{12}^n & \cdots & a_{1n}^n \\
  0 & a_{22}^n & \cdots & a_{2n}^n \\
  \quad & \quad & \ddots & \cdots \\
  0 & 0 & \cdots & a_{nn}^n
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2 \\
  \quad \\
  x_n
\end{pmatrix}
= 
\begin{pmatrix}
  b_1^n \\
  b_2^n \\
  \quad \\
  b_n^n
\end{pmatrix}.
$$

Realization:
Subtraction of a multiple of row $i$ from row $i + 1$ to row $n$. 
Step 1 of Gaussian elimination:

Elimination of $x_1$ from row 2 to row n:

If $a_{11} \neq 0$, subtract the $a_{i1}/a_{11}$-fold of row 1 from rows $2 \leq i \leq n$:

$$
\left( a_{i1} - \frac{a_{i1}}{a_{11}} a_{11} \right) x_1 + \left( a_{i2} - \frac{a_{i1}}{a_{11}} a_{12} \right) x_2 + \ldots + \left( a_{in} - \frac{a_{i1}}{a_{11}} a_{1n} \right) x_n = b_i - \frac{a_{i1}}{a_{11}} b_1,
$$

$2 \leq i \leq n$.

We obtain:

$$
\begin{pmatrix}
  a_{11} & a_{12} & \cdots & a_{1n} \\
  0 & a_{22} & \cdots & a_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & a_{n2} & \cdots & a_{nn}
\end{pmatrix}
\begin{pmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n
\end{pmatrix}
= \begin{pmatrix}
  b_1 \\
  b_2 \\
  \vdots \\
  b_n
\end{pmatrix} = \begin{pmatrix}
  b_1^{(2)} \\
  b_2^{(2)} \\
  \vdots \\
  b_n^{(2)}
\end{pmatrix}.
$$
**k-th step of Gaussian elimination:**

Elimination of \( x_k \) from row \( k+1 \) to row \( n \): If \( a_{kk}^{(k)} \neq 0 \), compute

\[
\ell_{ik} := a_{ik}^{(k)} / a_{kk}^{(k)}, \ k + 1 \leq i \leq n, \\
a_{ij}^{(k+1)} := a_{ij}^{(k)} - \ell_{ik} a_{kj}^{(k)}, \ k + 1 \leq i, j \leq n, \\
b_i^{(k+1)} := b_i^{(k)} - \ell_{ik} b_k^{(k)}, \ k + 1 \leq i \leq n.
\]

**Definition 1.12** Pivot elements, Pivot rows, Pivot columns

The elements \( a_{kk}^{(k)}, 1 \leq k \leq n - 1 \), are called Pivot elements. The corresponding rows/columns are called Pivot rows/Pivot columns.

**Lemma 1.13** Computational work of Gaussian elimination

Computational work of the k-th elimination step:

\[
n - k \text{ divisions}, \ (n - k)^2 \text{ multipl.} \ & \text{ subtract.}, \ n - k \text{ multipl.} \ & \text{ subtract.}.
\]

Total: \[2 \sum_{k=1}^{n-1} k^2 + 3 \sum_{k=1}^{n-1} k = 2 \left[ \frac{1}{6} (n - 1)n(2n - 1) \right] + 3 \left[ \frac{1}{2} (n - 1)n \right] = \frac{2}{3} n^3.\]

The computational work of Gaussian elimination amounts to \( O(n^3) \) essential operations.
Equivalence of Gaussian elimination and LR-decomposition:
Each step of the elimination represents a linear operation on the rows of the intermediate matrices:

The transformation of $A^{(k)}$ to $A^{(k+1)}$ (observe $A^{(1)} := A$) corresponds to the multiplication with a matrix $L_k \in \mathbb{R}^{n \times n}$ from the left:

$$
\begin{pmatrix}
1 & & & & \\
& \ddots & & & \\
& & 1 & & \\
& & & \ell_{k+1,k} & 1 \\
& & & & \ell_{nk}
\end{pmatrix}
\begin{pmatrix}
\begin{pmatrix} a_{11}^{(k)} \\
\vdots \\
a_{1n}^{(k)}
\end{pmatrix} \\
\begin{pmatrix} a_{k1}^{(k)} & \cdots & a_{kn}^{(k)} \\
\vdots & \ddots & \vdots \\
a_{k1}^{(k)} & \cdots & a_{kn}^{(k)}
\end{pmatrix}
\end{pmatrix}
= 
\begin{pmatrix}
\begin{pmatrix} a_{11}^{(k)} \\
\vdots \\
a_{1n}^{(k)}
\end{pmatrix} \\
\begin{pmatrix} a_{k1}^{(k+1)} & a_{k1,k+1}^{(k+1)} & a_{kn}^{(k+1)} \\
\vdots & \ddots & \vdots \\
a_{k1}^{(k+1)} & \cdots & a_{kn}^{(k+1)}
\end{pmatrix}
\end{pmatrix}
$$

Matrices of the form $L_k$ are called Frobenius matrices.
Lemma 1.14  Properties of the elimination matrices (Frobenius matrices)

The inverse $L_k^{-1}$ of the elimination matrix (Frobenius matrix) $L_k$ is given by

$$
L_k^{-1} = \begin{pmatrix}
1 & & & & \\
& 1 & & & \\
& & 1 & & \\
& & & \ell_{k+1,k} & 1 \\
& & & & \ell_{nk}
\end{pmatrix}.
$$

Theorem 1.15  LR-decomposition of matrices

Let $A \in \mathbb{R}^{n \times n}$ be regular, and assume that the Gaussian elimination can be performed with the elimination matrices $L_k, 1 \leq k \leq n-1$, and the resulting upper triangular matrix $R$. Then, the matrix $L := \prod_{k=1}^{n-1} L_k^{-1}$ is a lower triangular matrix with $\ell_{ii} = 1$, $1 \leq i \leq n$, and there holds:

$$(*) \quad A = LR.$$
Proof: We have

\[ A^{(n)} = R = L_{n-1} A^{(n-1)} = \ldots = \prod_{k=1}^{n-1} L_{n-k} A. \]

**Definition 1.16 LR-decomposition, unipotent matrices**

The decomposition \((*)\) of \(A\) into the product of a lower and an upper triangular matrix is called LR-decomposition of \(A\). Triangular matrices, whose diagonal elements are equal to 1, are called **unipotent**.

**Remark 1.17 Storage for Gaussian elimination**

During the generation of \(A^{(k)}\), the significant elements \(\ell_{ij}\) of the matrix \(L\) are stored in the remaining (lower triangular) part:

\[
\begin{pmatrix}
    a_{11}^{(k)} & a_{12}^{(k)} & \cdots & a_{1,k-1}^{(k)} & a_{ik}^{(k)} & \cdots & a_{1n}^{(k)} & b_{1}^{(k)} \\
    \ell_{21}^{(k)} & a_{22}^{(k)} & \cdots & a_{2,k-1}^{(k)} & a_{2k}^{(k)} & \cdots & a_{2n}^{(k)} & b_{2}^{(k)} \\
    & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
    \ell_{k1}^{(k)} & \ell_{k2}^{(k)} & \cdots & \ell_{k,k-1}^{(k)} & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & b_{k}^{(k)} \\
    & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
    \ell_{n1} & \ell_{n2} & \cdots & \ell_{n,k-1}^{(k)} & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & b_{n}^{(k)}
\end{pmatrix}
\]
Remark 1.18  Problems during Gaussian elimination

(i) Zero Pivot elements

The matrix

\[
A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \det A = -1, \quad a_{11} = 0
\]

is regular, but the first elimination can not be performed.

Exchanging the rows results in:

\[
\tilde{A} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = LR \text{ with } L = R = I.
\]

(ii) "Small" Pivot elements

Consider the linear system

\[
\begin{pmatrix} 10^{-4} & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}
\]

with the exact solution (up to four valid digits):

\[
x_1 = 1.000, \quad x_2 = 0.9999.
\]
Gaussian elimination with 3 valid digits results in:

\[ \ell_{21} = \frac{a_{21}}{a_{11}} = 1.00 \cdot 10^4 \]

and hence

\[
\begin{pmatrix}
1.00 \cdot 10^{-4} & 1.00 \\
0.00 & -1.00 \cdot 10^4
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix} =
\begin{pmatrix}
1.00 \\
-1.00 \cdot 10^4
\end{pmatrix}.
\]

We obtain:

\[ x_2 = 1.00 \text{ (correct)} , \quad x_1 = 0.00 \text{ (wrong)} . \]

Exchanging the rows in the coefficient matrix gives:

\[
\begin{pmatrix}
1 & 1 \\
10^{-4} & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix} =
\begin{pmatrix}
1 \\
2
\end{pmatrix}.
\]

Choosing \( \ell_{21} = a_{21}/a_{11} = 10^{-4} \) it follows that:

\[
\begin{pmatrix}
1 & 1 \\
0 & 1
\end{pmatrix}
\begin{pmatrix}
x_1 \\
x_2
\end{pmatrix} =
\begin{pmatrix}
1 \\
2
\end{pmatrix} \quad \text{and thus:}
\]

\[ x_2 = 1.00 \text{ (correct)} , \quad x_1 = 1.00 \text{ (correct)}. \]
Remark 1.19  Column pivoting with exchange of rows
In the k-th step of the elimination, the pivot row is chosen as that row which has the largest element (in absolute value) in the corresponding pivot column:

(i) In the k-th step compute \( p \in \{k, \ldots, n\} \), such that

\[
|a_{pk}^{(k)}| \geq |a_{jk}^{(k)}|, \quad k \leq j \leq n.
\]

The element \( a_{pk}^{(k)} \) is called the pivot element, and the \( p \)-th row is referred to as pivot row.

(ii) Exchange rows \( p \) and \( k \):

\[
A^{(k)} \rightarrow \tilde{A}^{(k)}, \quad \tilde{a}_{ij}^{(k)} = \begin{cases} 
 a_{kj}^{(k)}, & i = p, \\
 a_{pj}^{(k)}, & i = k, \\
 a_{ij}^{(k)}, & \text{otherwise}
\end{cases}
\]

(iii) Perform the k-th step applied to \( \tilde{A}^{(k)} \).

Remark 1.20  Computational work of partial pivoting, total pivoting
Instead of column pivoting with row exchanges, one can also do row pivoting with column exchanges. Both techniques are called partial pivoting and require \( O(n^2) \) additional operations. Complete pivoting with row and column exchanges requires \( O(n^3) \) operations and will thus not be performed.
Algebraic representation of row and column exchanges

**Definition 1.21 Permutation matrix**

Let $\mathbf{e}_i := (\delta_{i1}, \ldots, \delta_{in})^T$, $1 \leq i \leq n$, be the $i$-th unit vectors $\mathbb{R}^n$ and let $\pi$ be a permutation of $I_n := \{1, \ldots, n\}$. The matrix

$$
P_\pi := \begin{bmatrix}
\mathbf{e}_{\pi(1)} & \mathbf{e}_{\pi(2)} & \cdots & \mathbf{e}_{\pi(n)}
\end{bmatrix}
$$

is called a *permutation matrix*. $P_\pi$ is an orthogonal matrix, i.e., $P_\pi^T = P_\pi^{-1}$, and there holds $\det P_\pi = \text{sgn} \, \pi \in \{-1, +1\}$.

**Remark 1.22 Permutations in the domain of definition and in the range**

(i) row exchange (permutation in the range) \( A \mapsto P_\pi A \).

(ii) column exchange (permutation in the domain of definition) \( A \mapsto A P_\pi \).

**Theorem 1.23 Existence and uniqueness of the LR-decomposition**

Let $A \in \mathbb{R}^{n \times n}$ be regular. Then, there exists a permutation matrix $P$, such that $PA = LR$ where $L$ is a unipotent lower triangular matrix with $|\ell_{ij}| \leq 1$, $1 \leq j < i \leq n$, and $R$ is an upper triangular matrix. The matrices $L$ and $R$ are uniquely determined.
Remark 1.24 Scaling
Pivot strategies can be modified by the multiplication of rows/columns by scalar quantities.

Definition 1.25 Scaling of rows and columns, equilibration
Let $A \in \mathbb{R}^{n \times n}$ and $D_z := \text{diag}(\sigma_1, ..., \sigma_n)$ as well as $D_s := \text{diag}(\tau_1, ..., \tau_n)$.
Row scaling (column scaling) of $A$ is a left/right multiplication of $A$ by $D_z$ ($D_s$).

(i) Equilibration of rows
Let $\| \cdot \|$ be a vector norm and let $A_{(i)}$, $1 \leq i \leq n$, be those rows of $A \in \mathbb{R}^{n \times n}$ with $A_{(i)} \neq 0$, $1 \leq i \leq n$.
If we choose
\[
D_z := \text{diag}(\sigma_1, ..., \sigma_n), \quad \sigma_i := \|A_{(i)}\|^{-1}, \quad 1 \leq i \leq n
\]
then all rows of $\tilde{A} := D_z A$ have the same norm $\|\tilde{A}_{(i)}\| = 1$.

(ii) Equilibration of columns:
Let $\| \cdot \|$ be a vector norm and let $A_{(j)}$, $1 \leq j \leq n$, be those columns of $A \in \mathbb{R}^{n \times n}$ with $A_{(j)} \neq 0$, $1 \leq j \leq n$.
If we choose
\[
D_s := \text{diag}(\tau_1, ..., \tau_n), \quad \tau_j := \|A_{(j)}\|^{-1}, \quad 1 \leq j \leq n,
\]
then all columns of $\tilde{A} := AD_s$ have the same norm $\|\tilde{A}_{(j)}\| = 1$. 
Definition 1.26  Residual (defect)
Let $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and $x^{(0)} \in \mathbb{R}^n$. The vector
\[ r(x^{(0)}) := b - A x^{(0)} \]
is called the residual (defect) w.r.t. $x^{(0)}$.

Remark 1.27 Post processing
Let $x^* \in \mathbb{R}^n$ be the solution of the linear system $Ax = b$ and let $x^{(0)} \in \mathbb{R}^n$ be some approximation.
The error $e^{(0)} := x^* - x^{(0)}$ satisfies the defect equation
\[ A e^{(0)} = A x^* - A x^{(0)} = b - a x^{(0)} = r(x^{(0)}) . \]
The computation of an approximation $\tilde{e}^{(0)}$ of the defect equation and the update $x^{(1)}$ according to
\[ x^{(1)} = x^{(0)} + \tilde{e}^{(0)} \]
is a post processing iterative technique which is usually carried out in double precision.
Cholesky method for symmetric, positive definite matrices

Definition 1.28  Symmetric, positive definite matrices
A matrix \( A \in \mathbb{R}^{n \times n} \) is called symmetric, if

\[
A = A^T, \text{i.e., } a_{ij} = a_{ji}, 1 \leq i, j \leq n,
\]

and it is called positive definite, if

\[
(A x, x) > 0, \quad x \neq 0,
\]

where \((\cdot, \cdot)\) is the Euclidian inner product in \( \mathbb{R}^n \).

For symmetric, positive definite matrices we will use the abbreviation \( \text{spd} \) matrices.

Theorem 1.29  Properties of spd matrices
Let \( A \in \mathbb{R}^{n \times n} \) be spd. Then, there holds gilt:

(i) \( A \) is invertible,  
(ii) \( a_{ii} > 0 \), \( 1 \leq i \leq n \),  
(iii) \( \max_{1 \leq i, j \leq n} |a_{ij}| = \max_{1 \leq i \leq n} a_{ii} \),  
(iv) When performing Gaussian elimination without pivoting, all \( (n - i, n - i), 1 \leq i \leq n - 1 \) submatrices of the intermediate matrices are again spd.
Proof of (i): We have

\[ A x = 0 \implies (A x, x) = 0 \implies x = 0. \]

Proof of (ii): Choosing \( x = e_i \), \( 1 \leq i \leq n \), it follows that

\[ (A e_i, e_i) = a_{ii} > 0. \]

Proof of (iii): Since \( A \) is spd, there exists \( A^{1/2} \). Setting \( \| \cdot \| := (\cdot, \cdot)^{1/2} \), we get:

\[ |a_{ij}| = |(A e_i, e_j)| = |(A^{1/2} e_i, A^{1/2} e_j)| \leq \|A^{1/2} e_i\| \|A^{1/2} e_j\| = (A e_i, e_i)^{1/2} (A e_j, e_j)^{1/2} = a_{ii}^{1/2} a_{jj}^{1/2}. \]

Proof of (iv): After the first elimination step, we arrive at:

\[ A^{(2)} = L_1 A = \begin{pmatrix} a_{11} & z^T \\ - & - \\ - & - \\ 0 & B^{(2)} \end{pmatrix}, \quad L_1 A L_1^T = \begin{pmatrix} a_{11} & 0 \\ - & - \\ - & - \\ 0 & B^{(2)} \end{pmatrix}. \]

Since, \( L_1 A L_1^T \) is spd, so is \( B^{(2)}. \)
Remark 1.30:
The properties (iii),(iv) in Theorem 1.29 imply that the LR-decomposition of spd matrices does not require pivoting.

Theorem 1.31  Rational Cholesky decomposition
Let $A \in \mathbb{R}^{n \times n}$ be an spd matrix. Then, there exist a unipotent lower triangular matrix $L$ and a positive diagonal matrix $D$ such that

$$A = L D L^T.$$  

Proof: The assertion can be proved by a constructive continuation of the proof of Theorem 1.29 (iv) with $L = \prod_{k=1}^{n-1} L_k^{-1}$ and $D = \text{diag}(a_{11}^{(n)}, \ldots, a_{nn}^{(n)})$.

Corollary 1.32  Cholesky decomposition
Let $A \in \mathbb{R}^{n \times n}$ be an spd matrix. Setting $D^{1/2} := \text{diag}(\sqrt{a_{11}^{(n)}}, \ldots, \sqrt{a_{nn}^{(n)}})$ and $\bar{L} := LD^{1/2}$, we obtain:

$$A = \bar{L} \bar{L}^T.$$
Remark 1.34 Realization of the Cholesky decomposition

In explicit form, the equations $A = LL^T$ are as follows:

$$
\begin{pmatrix}
  a_{11} & \cdots & a_{1n} \\
  \vdots & \ddots & \vdots \\
  a_{n1} & \cdots & a_{nn}
\end{pmatrix} =
\begin{pmatrix}
  \ell_{11} & 0 & \cdots & 0 \\
  \vdots & \ddots & \vdots & \vdots \\
  \ell_{n1} & \cdots & \ell_{nn}
\end{pmatrix}
\begin{pmatrix}
  \ell_{11} & \cdots & \ell_{n1} \\
  0 & \ddots & \vdots \\
  0 & \cdots & \ell_{nn}
\end{pmatrix}
$$

The elementwise evaluation results in:

- For $i = k$: $a_{kk} = \ell_{k1}^2 + \cdots + \ell_{k,k-1}^2 + \ell_{kk}^2$
- For $i > k$: $a_{ik} = \ell_{i1} \ell_{k1} + \cdots + \ell_{i,k-1} \ell_{k,k-1} + \ell_{ik} \ell_{kk}$

Algorithm: Cholesky decomposition

```plaintext
for k = 1 to n do
    \ell_{kk} = (a_{kk} - \sum_{j=1}^{k-1} \ell_{kj}^2)^{1/2}
for i = k + 1 to n do
    \ell_{ik} = (a_{ik} - \sum_{j=1}^{k-1} \ell_{ij} \ell_{kj})/\ell_{kk}
```
Remark 1.35  Computational work of the Cholesky decomposition

The number of essential operations is given by:

\[ \frac{1}{3} n^3 \] multiplikations and subtractions as well as n square roots

Therefore, the computational work is roughly 50 % of that for Gaussian elimination.
1.2 Iterative methods for linear algebraic systems

1.2.1 Linear iterations

Definition 1.36 Linear iteration

Let $M, N \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n, x^{(0)} \in \mathbb{R}^n$.

Then, the recursion

\[
(*) \quad x^{(k+1)} = M x^{(k)} + N b, \quad k \in \mathbb{N}_0
\]

is called a linear iteration. The matrix $M$ is called the iteration matrix, and $x^{(0)}$ is referred to as the start vector.

Definition 1.37 Fixed point, consistency

(i) A vector $x^* \in \mathbb{R}^n$ is called a fixed point of the linear iteration, if

\[
x^* = M x^* + N b.
\]

(ii) The linear iteration $(*)$ is said to be consistent with the linear system $Ax = b$, if any solution of the linear system is a fixed point of $(*)$. 
Remark 1.38: Consistency of linear iterations
Assume $x^*$ to be a solution of $Ax = b$. Then, if the linear iteration is consistent with the linear system, there holds

$$x^* = Mx^* + Nb = Mx^* + N(Ax^*) = (M + NA)x^*.$$

Lemma 1.39 Characterization of consistency
The linear iteration (*) is consistent with the linear system $Ax = b$ if and only if,

$$M = I - NA.$$

Definition 1.40 Second normal form of linear iterations
Assume consistency and a regular matrix $A \in \mathbb{R}^{n \times n}$. Then, there holds

$$N = (I - M)A^{-1},$$

and it follows that

$$x^{(k+1)} = Mx^{(k)} + Nb = (I - NA)x^{(k)} + Nb = x^{(k)} - N(Ax^{(k)} - b) = x^{(k)} + Nr(x^{(k)})$$

with $r(x^{(k)}) := b - Ax^{(k)}$ as the residual with respect to the iterate $x^{(k)}$. This iterative scheme is called the second normal form of the linear iteration (*).
Definition 1.41: Convergence of linear iterations

A linear iteration is called convergent, if \( \lim_{k \to \infty} x^{(k)} \) exists independently of the choice of the start vector \( x^{(0)} \in \mathbb{R}^n \).

Lemma 1.42 Representation of the iterates and iteration error

(i) For a linear iteration, the iterates \( x^{(k)}, k \in \mathbb{N} \), admit the representation

\[
x^{(k)} = M^k x^{(0)} + \sum_{j=0}^{k-1} M^j N b.
\]

(ii) In case of consistent iterations, the iteration error \( e^{(k)} := x^{(k)} - x^* \) is given by

\[
e^{(k)} = M^k e^{(0)}, \quad k \in \mathbb{N}_0.
\]

Proof: The first assertion is proved by induction on \( k \in \mathbb{N} \): Assuming that the assertion holds true for \( k - 1 \), we obtain:

\[
x^{(k)} = M x^{(k-1)} + N b = M ( M^{k-1} x^{(0)} + \sum_{j=0}^{k-2} M^j N b ) + N b =
\]

\[
= M^k x^{(0)} + \sum_{j=0}^{k-2} M^{j+1} N b + N b = M^k x^{(0)} + \sum_{j=0}^{k-1} M^j N b.
\]


As far as the representation of the error is concerned, we have

\[ e^{(k)} = x^{(k)} - x^* = M x^{(k-1)} + N b - x^* = M x^{(k-1)} + N A x^* - x^* = \]

\[ = M x^{(k-1)} + (I - M)x^* - x^* = M \left( x^{(k-1)} - x^* \right) = M e^{(k-1)}. \]

**Theorem 1.43  Characterization of convergence**

A linear iteration with the iteration matrix \( M \) is convergent, if and only if

\[ \rho(M) < 1. \]

Here, \( \rho(M) \) denotes the spectral radius of the iteration matrix, i.e., \( \rho(M) := \max |\lambda| \), where \( \lambda \) is an eigenvalue of \( M \).

**Proof:** If the linear iteration is convergent, without restriction of generality we may assume \( b = 0 \). Then, there holds

\[ x^{(k)} = M^k x^{(0)} \rightarrow x^* \quad (m \rightarrow \infty). \]

For \( x^{(0)} = 0 \) we obtain \( x^* = 0 \). Assuming \( \rho(M) \geq 1 \) in contradiction to the assertion and choosing \( x^{(0)} \neq 0 \) as an eigenvector associated with the eigenvalue \( \lambda \) with \( |\lambda| = \rho(M) \), it follows that \( x^{(k)} = \lambda^k x^{(0)} \xrightarrow{?} x^* = 0 \) which contradicts the convergence of the iteration.
For the proof of the converse statement, assume $\rho(M) < 1$. For an arbitrarily chosen start vector $x^{(0)} \in \mathbb{R}^n$ we have:

$$M^k x^{(0)} \to 0, \quad \sum_{j=0}^{k-1} M^j \to (I - M)^{-1} \quad (k \to \infty).$$

Hence,

$$x^{(k)} = M^k x^{(0)} + \sum_{j=0}^{k-1} M^j N b \to (I - M)^{-1} N b \quad (k \to \infty).$$

**Corollary 1.44  Consistency and convergence**
Assume that we have a consistent and convergent linear iteration. Then, $A$ and $N$ are regular, and there holds

$$x^{(k)} \to x^* = A^{-1} b \quad (k \to \infty).$$

**Proof:** Consistency yields $NA = I - M$, whereas convergence implies $\rho(M) < 1$. Consequently, $I - M$ is regular, and we obtain $(I - M)^{-1} N = A^{-1}$ and

$$x^{(k)} \to (I - M)^{-1} N b = A^{-1} b = x^* \quad (k \to \infty).$$
1.2.2 Classical linear iterations
In the sequel, we always assume consistency.

Definition 1.45 Richardson iteration
Choose $N_R := \theta I$, $\theta \in \mathbb{R}$. Then, $M_R = M_R(\theta) = I - \theta A$, and we obtain the iteration
\[x^{(k+1)} = x^{(k)} - \theta (Ax^{(k)} - b),\]
component wise: \[x_i^{(k+1)} = x_i^{(k)} - \theta \left( \sum_{j=1}^{n} a_{ij} x_j^{(k)} - b_i \right), \quad 1 \leq i \leq n.

Lemma 1.46 Spectral radius of the Richardson iteration matrix
Assume $\sigma(A) \subseteq \mathbb{R}$ and
\[\lambda_{\min}(A) := \min_{\lambda \in \sigma(A)} \lambda, \quad \lambda_{\max}(A) := \max_{\lambda \in \sigma(A)} \lambda.
\]
Then $\sigma(M_R(\theta)) \subseteq \mathbb{R}$, $\theta \in \mathbb{R}$, and it follows that
\[\rho(M_R(\theta)) = \max \left\{ |1 - \theta \lambda_{\min}(A)|, |1 - \theta \lambda_{\max}(A)| \right\}.
\]
Proof: We have $\lambda(M_R(\theta)) = 1 - \theta \lambda(A)$ with $\lambda(A) \in \sigma(A)$. 
Theorem 1.47  Convergence of the Richardson iteration
Assume $\sigma(A) \subset (0, \infty)$, and $\theta \in \mathbb{R}$. Then, the Richardson iteration converges, if and only if

$$0 < \theta < 2/\lambda_{\text{max}}(A).$$

The optimal choice of $\theta$ is

$$\theta_{\text{opt}} = 2/(\lambda_{\text{min}}(A) + \lambda_{\text{max}}(A)) \quad \text{with} \quad \rho(M_R(\theta_{\text{opt}})) = \frac{\lambda_{\text{max}}(A) - \lambda_{\text{min}}(A)}{\lambda_{\text{min}}(A) + \lambda_{\text{max}}(A)}.$$
Definition 1.48  Triangular decomposition of the coefficient matrix

The decomposition of \( A \in \mathbb{R}^{n \times n} \) according to

\[
\begin{pmatrix}
  a_{11} & \cdots & a_{1n} \\
  \vdots & \ddots & \vdots \\
  a_{n1} & \cdots & a_{nn}
\end{pmatrix}
= \begin{pmatrix}
  a_{11} & 0 & \cdots & 0 \\
  0 & \ddots & \vdots & \vdots \\
  0 & \cdots & 0 & a_{nn}
\end{pmatrix}
- \begin{pmatrix}
  0 & \cdots & \cdots & 0 \\
  \cdots & \ddots & \ddots & \vdots \\
  \cdots & \cdots & \ddots & \cdots \\
  -a_{n1} & \cdots & -a_{n,n-1} & 0
\end{pmatrix}
- \begin{pmatrix}
  0 & -a_{12} & \cdots & -a_{1n} \\
  \cdots & \ddots & \ddots & \vdots \\
  \cdots & \cdots & \ddots & \cdots \\
  \cdots & \cdots & \cdots & -a_{n-1,n}
\end{pmatrix}
\]

is called triangular decomposition of \( A \).

Definition 1.49  Jacobi iteration

Assume \( a_{ii} \neq 0 \), \( 1 \leq i \leq n \), and choose \( N_J = D^{-1} \). Then, \( M_J = I - D^{-1}A \), and we obtain

\[
x^{(k+1)} = x^{(k)} - D^{-1}(Ax^{(k)} - b) = x^{(k)} - D^{-1}((D - E - F)x^{(k)} - b) = D^{-1}(E + F)x^{(k)} + D^{-1}b 
\]

component wise: \( x_i^{(k+1)} = \left( b_i - \sum_{j \neq i} a_{ij}x_j^{(k)} \right)/a_{ii} \), \( 1 \leq i \leq n \).

The iteration is called Jacobi iteration.
Definition 1.50  Gauss-Seidel iteration

Assume $a_{ii} \neq 0$, $1 \leq i \leq n$. If we choose $N_{GS} = (D - E)^{-1}$, then $M_{GS} = I - (D - E)^{-1}A$, and we obtain

$$x^{(k+1)} = x^{(k)} - (D - E)^{-1}(Ax^{(k)} - b) = x^{(k)} - (D - E)^{-1}((D - E - F)x^{(k)} - b) = (D - E)^{-1}Fx^{(k)} + (D - E)^{-1}b,$$

component wise: $x^{(k+1)}_i = (b_i - \sum_{j=1}^{i-1} a_{ij}x^{(k+1)}_j - \sum_{j=i+1}^{n} a_{ij}x^{(k)}_j)/a_{ii}$, $1 \leq i \leq n$.

This iteration is called Gauss-Seidel iteration.

Remark 1.51  Convergence in case of spd matrices (preliminaries)

On the set of spd matrices we introduce a partial ordering according to

$$A > B \iff A - B \text{ is spd}.$$

Moreover, we define the energy norm of an spd matrix $A$ by:

$$\|x\|_A := (A x, x)^{1/2} = (A^{1/2} x, A^{1/2} x)^{1/2} = \|A^{1/2} x\|.$$
As far as the associated matrix norm is concerned, it follows that:

$$
\|B\|_A = \sup_{x \neq 0} \frac{\|Bx\|_A}{\|x\|_A} = \sup_{x \neq 0} \frac{\|A^{1/2}Bx\|}{\|A^{1/2}x\|} = \sup_{y = A^{1/2}x} \frac{\|A^{1/2}BA^{-1/2}y\|}{\|y\|} = \|A^{1/2}BA^{-1/2}\|.
$$

**Theorem 1.52  Convergence in case of spd matrices**

Assume \( A \in \mathbb{R}^{n \times n} \) to be spd and \( N \) regular. Assume further that for \( W := N^{-1} \)

$$
W + W^T > A.
$$

Then, for \( M = I - NA \) there holds

$$
\rho(M) \leq \|M\|_A < 1.
$$

**Proof:** Since \( \rho(M) \leq \|M\|_A \), it suffices to show \( \|M\|_A < 1 \). Observing \( M = I - NA = I - W^{-1}A \), we have

$$
\|M\|_A = \|A^{1/2}MA^{1/2}\| = \|I - A^{1/2}W^{-1}A^{1/2}\| =: \bar{M}.
$$
Moreover, we have:

\[
\hat{M}^T \hat{M} = (I - A^{1/2} W^{-1} A^{1/2})^T (I - A^{1/2} W^{-1} A^{1/2}) = \\
= I - A^{1/2} (W^{-T} + W^{-1}) A^{1/2} + A^{1/2} W^{-T} A W^{-1} A^{1/2} = \\
= I - A^{1/2} W^{-T} (W + W^T) W^{-1} A^{1/2} + A^{1/2} W^{-T} A W^{-1} A^{1/2} < \\
< I - A^{1/2} W^{-T} A W^{-1} A^{1/2} + A^{1/2} W^{-T} A W^{-1} A^{1/2} = I ,
\]

whence

\[
\| M \|_A = \| \hat{M} \| = \rho(\hat{M}^T \hat{M})^{1/2} < \rho(I)^{1/2} = 1.
\]

**Theorem 1.53** Convergence of the Jacobi and the Gauss-Seidel iteration for matrices

Assume \( A \in \mathbb{R}^{n \times n} \) to be spd. Then, there holds:

(i) If \( 2D - A > 0 \), the Jacobi iteration converges with

\[
\rho(M_J) \leq \| M_J \|_A < 1.
\]

(ii) The Gauss-Seidel iteration converges with

\[
\rho(M_{GS}) \leq \| M_{GS} \|_A < 1.
\]
Proof: The proof is based on the application of Theorem 1.52:

Jacobi: \[ N_J = D - 1 \implies W_J = D \implies W_J + W_J^T 2D > A, \]

Gauss-Seidel: \[ N_{GS} = (D - E)^{-1} \implies W_{GS} = D - E \implies W_{GS} + W_{GS}^T = D - E + D - E^T = D - E - F + D = A + D > A. \]

Theorem 1.54 Theorem of Stein and Rosenberg
Assume the matrix \( E + F \) to be nonnegative. Then, for the Jacobi and the Gauss-Seidel iteration exactly one of the following relations holds true

1. \( \rho(M_{GS}) = \rho(M_J) = 0 \),
2. \( 0 < \rho(M_{GS}) < \rho(M_J) < 1 \),
3. \( \rho(M_{GS}) = \rho(M_J) = 1 \),
4. \( \rho(M_{GS}) > \rho(M_J) > 1 \).
Remark 1.55  Acceleration of convergence by Successive Over Relaxation (SOR)
The convergence of the Gauss-Seidel iteration can be improved by a relaxation of the iterates

\begin{align*}
(*) \quad x_i^{(k+1)} &= \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) / a_{ii}, \\
(**) \quad x_i^{(k+1)} &= \omega x_i^{(k+1/2)} + (1 - \omega) x_i^{(k)}, \quad 0 < \omega < 2.
\end{align*}

Inserting (*) into (**) yields:

\begin{align*}
x_i^{(k+1)} &= x_i^{(k)} + \omega \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) / a_{ii}, \quad 1 \leq i \leq n, \\
D x^{(k+1)} &= D x^{(k)} + \omega \left( b - (D - F) x^{(k)} + E x^{(k+1)} \right), \\
(D - \omega E) x^{(k+1)} &= \left[ (1 - \omega) D + \omega F \right] x^{(k)} + \omega b, \\
x^{(k+1)} &= \left( D - \omega E \right)^{-1} \left[ (1 - \omega) D + \omega F \right] x^{(k)} + \omega \left( D - \omega E \right)^{-1} b.
\end{align*}

= M_{SOR} (\omega) \\
= N_{SOR} (\omega)
Definition 1.56 The SOR iteration (Successive Over Relaxation)

The iteration

\[ x^{(k+1)} = (D - \omega E)^{-1} \left[ (1 - \omega) D + \omega F \right] x^{(k)} + \omega (D - \omega E)^{-1} b. \]

is referred to as a relaxation. For \( 0 < \omega < 1 \), it is called an under relaxation, in case \( \omega = 1 \) we obtain the Gauss-Seidel iteration, and for \( 1 < \omega < 2 \) it is called over relaxation.
Theorem 1.57  Convergence of the SOR iteration for spd matrices

Assume $A \in \mathbb{R}^{n \times n}$ to be spd. Then, the SOR iteration converges for all $0 < \omega < 2$ with

$$\rho(M_{\text{SOR}}(\omega)) \leq \|M_{\text{SOR}}(\omega)\|_A < 1.$$  

Proof: We apply Theorem 1.52 with $W_{\text{SOR}}(\omega) := \omega^{-1}D - E$:

$$W_{\text{SOR}}(\omega) + W_{\text{SOR}}^T(\omega) = 2\omega^{-1}D - E - F = A + (2\omega^{-1} - 1)D > A > 0.$$
1.2.3 The method of conjugate gradients (cg-iteration)

Theorem 1.57 Equivalence with the minimization of convex functionals
Assume \( A \in \mathbb{R}^{n \times n} \) to be spd and \( b \in \mathbb{R}^n \). Then, \( x^* \in \mathbb{R}^n \) is a solution of the linear system \( Ax = b \), if and only if \( x^* \) solves the minimization problem:

\[
(*) \quad J(x^*) = \min_{x \in \mathbb{R}^n} J(x) \quad , \quad J(x) := \frac{1}{2} \left( A \, x, x \right) - \left( b, x \right).
\]

Proof: We have:

\[
J'(x) = \text{grad} \, J(x) = A \, x - b.
\]
Since \( J''(x) = A > 0 \), the condition \( J'(x) = 0 \) is necessary and sufficient for the existence of a minimum.

Definition 1.58 Search direction
Assume that \( x \in \mathbb{R}^n \) is some approximation of the solution \( x^* \) of (*) and \( p \in \mathbb{R}^n \setminus \{0\} \) a search direction for the computation of the minimum of (*). Then, starting from \( x \), we perform a (one dimensional) minimization of \( J \) in the direction of \( p \):

\[
J(x + \lambda_{\text{opt}} \, p) = \min_{\lambda \in \mathbb{R}} J(x + \lambda \, p).
\]
Lemma 1.59  Step length selection

The solution of the one dimensional minimization problem

\[ J(x + \lambda_{\text{opt}} p) = \min_{\lambda \in \mathbb{R}} J(x + \lambda p). \]

is given by:

\[ \lambda_{\text{opt}} = \frac{(r, p)}{(Ap, p)}, \quad r := b - Ax. \]

\( \lambda_{\text{opt}} \) is called the optimal step length in the search direction \( p \).

Proof: As the necessary and sufficient optimality condition we obtain:

\[ (J'(x + \lambda_{\text{opt}} p), p) = (Ax + \lambda_{\text{opt}} Ap - b, p) = 0. \]

Remark 1.60  Steepest descent

A suitable search direction is given by the direction of the steepest descent:

\[ p = - \text{grad } J(x) = b - Ax = r. \]
Definition 1.61  Method of steepest descent
If the search direction $p$ is chosen as the residual $r = b - Ax$, we obtain the method of steepest descent:

Initialization: Choose a start vector $x^{(0)} \in \mathbb{R}^n$ and compute the residual $r^{(0)} := b - Ax^{(0)}$.

Iteration loop: For $k = 0, 1, 2, \ldots$ do:

\[
x^{(k+1)} := x^{(k)} + \lambda_{opt} r^{(k)}, \quad \lambda_{opt} := \left( \frac{r^{(k)} \cdot r^{(k)}}{A r^{(k)} \cdot r^{(k)}} \right),
\]

\[
r^{(k+1)} := r^{(k)} - \lambda_{opt} A r^{(k)}.
\]

Graphical illustration of the method of steepest descent:
Consider $A = \text{diag}(\lambda_1, \lambda_2)$, $0 < \lambda_1 \leq \lambda_2$ and $b = 0$.

Equipotential lines $N_c := \{ x \in \mathbb{R}^2 | J(x) = \frac{1}{2} (\lambda_1 x_1^2 + \lambda_2 x_2^2) = c \}$.

The iterates $x^{(k)}$ are located on the ellipses $N_k$ with $J(x^{(k)}) = c$.

The iterates $x^{(k)}$ are located on the ellipses $N_k$ with $J(x^{(k)}) = c$.

The vector $(x^{(k+1)} - x^{(k)})$ is tangential to $N_{k+1}$. 
Definition 1.62  Optimality with respect to search directions
A vector $x \in \mathbb{R}^n$ is called optimal with respect to the direction $p \in \mathbb{R}^n \setminus \{0\}$, if there holds:
$$J(x) \leq J(x + \lambda p), \quad \lambda \in \mathbb{R}. \quad \text{for all } \lambda \in \mathbb{R}.$$  
$x$ is called optimal with respect to the subspace $U \subset \mathbb{R}^n$, if:  
$$J(x) \leq J(x + y), \quad y \in U.$$

Lemma 1.63  Characterization of optimality with respect to a search direction
The optimality of $x \in \mathbb{R}^n$ with respect to $p \in \mathbb{R}^n \setminus \{0\}$ is equivalent to

$$(p, r) = 0, \quad \text{i.e.,} \quad p \perp r = b - Ax = -\nabla J(x).$$

Proof: We have:

$$J(x) = \min_{\lambda \in \mathbb{R}} J(x + \lambda p) \iff (\nabla J(x + \lambda_{\text{opt}} p, p) = (\nabla J(x), p) = - (b - Ax, p) = 0.$$
Lemma 1.64  Optimality for the method of steepest descent

For the method of steepest descent, the iterate $x^{(k+1)}$ is optimal with respect to the search direction $r^{(k)}$, i.e.,

$$r^{(k+1)} \perp r^{(k)} = \lambda_{opt}^{-1} (x^{(k+1)} - x^{(k)}).$$

**Proof:** For $r^{(k+1)} = r^{(k)} - \lambda_{opt} A r^{(k)}$ there holds:

$$(r^{(k+1)} , r^{(k)}) = (r^{(k)} , r^{(k)}) - \lambda_{opt} (Ar^{(k)} , r^{(k)}).$$

Remark 1.65  Disadvantage of the method of steepest descent

For the method of steepest descent, the optimality of $x^{(k+1)}$ with respect to $r^{(k)}$ is lost, when computing the new iterate $x^{(k+2)}$.

**Problem:** What kind of property must a search direction have to preserve optimality?

If $x$ is optimal with respect to $p$ (i.e., $r \perp p$), and $x' = x + q$, how to choose $q$ such that $x'$ is optimal with respect to $p$?

$$(r', p) = (r - Aq , p) = (r , p) - (Aq , p) = 0 \iff Aq \perp p.$$
Definition 1.66 Conjugate (A-orthogonal) directions
Two vectors $p, q \in \mathbb{R}^n \setminus \{0\}$ satisfying
$$Aq \perp p$$
are called conjugate resp. A-orthogonal.

Definition 1.67 Method of conjugate directions

**Initialization:** Choose a start vector $x^{(0)} \in \mathbb{R}^n$ and compute $r^{(0)} = b - Ax^{(0)}$.

**Iteration loop:** For $k = 0, 1, 2, \ldots$ do:

1. Choose the direction $p^{(k)}$, conjugate to $p^{(\ell)}$, $\ell < k$ ,
2. $x^{(k+1)} := x^{(k)} + \lambda_{\text{opt}} p^{(k)}$, $\lambda_{\text{opt}} := \frac{(r^{(k)}, p^{(k)})}{(Ap^{(k)}, p^{(k)})}$ ,
3. $r^{(k+1)} := r^{(k)} - \lambda_{\text{opt}} Ap^{(k)}$ .

**Remark 1.68** The method is not practicable, if there is no suitable choice of $p^{(k)}$. 
Remark 1.69  Method of conjugate gradients (first version)

Idea: Use the residuals $r^{(k)}$ for determining the search direction $p^{(k)}$:

If $p^{(0)}, p^{(1)}, ..., p^{(k-1)}$ are already known, orthogonalize $r^{(k)}$ w.r.t. $(\cdot, \cdot)_A$ according to

$$
p^{(k)} = r^{(k)} - \sum_{\ell=0}^{k-1} \frac{(Ar^{(k)}, p^{(\ell)})}{(Ap^{(\ell)}, p^{(\ell)})} p^{(\ell)}.
$$

Lemma 1.70  Properties of conjugate directions

1. $(Ap^{(k)}, p^{(\ell)}) = 0$, $0 \leq \ell \leq k - 1$,
2. $r^{(k)}$ and $p^{(k)}$ can be zero only together, i.e., either $x^{(k)} = x^*$ or $p^{(k)} \neq 0$.
3. $U_{(k)} := \text{span} \{p^{(0)}, ..., p^{(k)}\} = \text{span} \{p^{(0)}, ..., p^{(k-1)}, r^{(k)}\}$,
4. $r^{(k)} \perp U_{\ell}$, $\ell < k$. 
Definition 1.71  Method of conjugate gradients (cg - iteration)

Initialization: Choose a start vector $x^{(0)}$ and compute $r^{(0)} := b - Ax^{(0)}$. Set $p^{(0)} := r^{(0)}$.

Iteration loop: For $k = 0, 1, 2, \ldots$ do:

1. $x^{(k+1)} := x^{(k)} + \alpha p^{(k)}$, \hspace{1cm} $\alpha := \frac{(r^{(k)}, p^{(k)})}{(Ap^{(k)}, p^{(k)})}$

2. $r^{(k+1)} := r^{(k)} - \alpha Ap^{(k)}$,

3. $p^{(k+1)} := r^{(k+1)} - \beta p^{(k)}$, \hspace{1cm} $\beta := \frac{(Ar^{(k+1)}, p^{(k)})}{(Ap^{(k)}, p^{(k)})}$.

Theorem 1.72  Properties of the cg - iteration I

In case of exact computations, the iteration terminates with the exact solution $x^* = A^{-1}b$ after at most $n$ steps.

Proof: If $r^{(m)} = 0$ for $m < n$, then $x^{(m)} = x^*$. Otherwise, we have $r^{(m)}$, $p^{(m)} \neq 0$, and the iteration continues. If $m = n - 1$, we have $\|r^n\| = \text{span} \{p^{(0)}, \ldots, p^{(n-1)}\}$, i.e., there is no continuation.
Lemma 1.73 Properties of the cg - iteration II
For the subspaces \( U_k \), \( k \in \mathbb{N}_0 \) there holds:
\[
U_k := \text{span} \{ p^{(0)}, \ldots, p^{(k)} \} = \text{span} \{ r^{(0)}, \ldots, r^{(k)} \} = \text{span} \{ r^{(0)}, A r^{(0)}, \ldots, A^k r^{(0)} \}.
\]

Proof: Induction on \( k \in \mathbb{N}_0 \): For \( k = 0 \) we have \( p^{(0)} = r^{(0)} \). We assume that the assertion holds true for \( k - 1 \). Then, in view of Lemma 1.70 (3), we obtain:
\[
U_k = \text{span} \{ p^{(0)}, \ldots, p^{(k)} \} = \text{span} \{ p^{(0)}, \ldots, p^{(k-1)}, r^{(k)} \} = \text{span} \{ r^{(0)}, \ldots, r^{(k)} \}.
\]

Moreover, taking the induction’s assumption into account:
\[
\begin{align*}
r^{(k-1)} & \in \text{span} \{ r^{(0)}, \ldots, r^{(k-1)} \} = \text{span} \{ r^{(0)}, \ldots, A^{k-1} r^{(0)} \} \\
A p^{(k-1)} & \in A U_{k-1} = A \text{span} \{ r^{(0)}, \ldots, A^{k-1} r^{(0)} \} = \text{span} \{ A r^{(0)}, \ldots, A^k r^{(0)} \} \\
& \Rightarrow r^{(k)} = r^{(k-1)} - \alpha A p^{(k-1)}
\end{align*}
\]
\[
\begin{align*}
r^{(k)} & \in \text{span} \{ r^{(0)}, \ldots, A^k r^{(0)} \} \\
& \Rightarrow U_k = \text{span} \{ r^{(0)}, \ldots, r^{(k)} \} \subset \text{span} \{ r^{(0)}, \ldots, A^k r^{(0)} \}.
\end{align*}
\]

Since \( \dim U_k = k + 1 \), the subspaces must coincide.

Definition 1.74 Krylov subspace of dimension \( k \)
The subspace \( \text{span} \{ r^{(0)}, \ldots, A^{k-1} r^{(0)} \} \) is called the Krylov subspace of dimension \( k \).
Lemma 1.75 Properties of the cg - iteration III

The iterates \( x^{(k)} \), \( k \in \mathbb{N} \), in the cg - iteration are solutions of the minimization problems:

\[
\begin{align*}
(1) & \quad J(x^{(k)}) = \min_{\lambda_0, \ldots, \lambda_{k-1}} J(x^{(0)}) + \sum_{\ell=1}^{k-1} \lambda_{\ell} p^{(\ell)}) , \\
(2) & \quad J(x^{(k)}) = \min_{\mu_0, \ldots, \mu_{k-1}} J(x^{(0)}) + \sum_{\ell=1}^{k-1} \mu_{\ell} r^{(\ell)}) , \\
(3) & \quad J(x^{(k)}) = \min_{p_{k-1} \in P_{k-1}} J(x^{(0)}) + p_{k-1}(A) r^{(0)}) ,
\end{align*}
\]

where \( P_{k-1} \) denotes the linear space of polynomials of degree \( \leq k - 1 \) on \( \mathbb{R} \).

Proof: According to Lemma 1.73, all minimization problems are of the form

\[
J(x^{(k)}) = \min_{y \in U_{k-1}} J(x^{(0)} + y) .
\]
Convergence analysis of the cg - iteration

Theorem 1.76  Representation of the iteration error

Let \((x^{(k)})_{k\in\mathbb{N}_0}\) be the sequence of cg-iterates. Then, for all \(0 \leq j \leq k\) there exist polynomials \(p_j \in P_j\) with \(p_j(1) = 1\) such that

\[ e^{(k)} := x^{(k)} - x^* = p_k(I - A)e^{(0)}. \]

Proof: In view of Lemma 1.73:

\[ (+) \quad x^{(k)} - x^{(0)} \in \text{span} \{p^{(0)}, \ldots, p^{(k-1)}\} = \text{span} \{r^{(0)}, \ldots, A^{k-1}r^{(0)}\}. \]

Moreover, we have:

\[ A(x^{(k)} - x^{(0)}) = b - Ax^{(0)} - (b - Ax^{(k)}) = r^{(0)} - r^{(k)}. \]

Taking \((+)_k\) into account, it follows that:

\[ r^{(k)} - r^{(0)} \in \text{span} \{Ar^{(0)}, \ldots, A^k r^{(0)}\} \Rightarrow r^{(k)} = r^{(0)} + \sum_{j=1}^{k} \alpha_j A^j r^{(0)} \Rightarrow \]

\[ r^{(k)} = r_k(A)r^{(0)}, \text{ where } r_k(\xi) := \sum_{j=0}^{k} \alpha_j \xi^j \text{ with } \alpha_0 = 1, \text{i.e., } r_k(0) = \alpha_0 = 1. \]
We define:

\[ p_k(\xi) := r_k(1 - \xi) \implies p_k \in P_k , \quad p_k(1) = 1 \quad \text{and} \quad (\ast) \quad r^{(k)} = p_k(I - A)r^{(0)}. \]

On the other hand, we have

\[ e^{(k)} = x^{(k)} - x^* = A^{-1}(Ax^{(k)} - Ax^*) = A^{-1}(Ax^{(k)} - b) = -A^{-1}r^{(k)}. \]

Observing (\ast), we obtain

\[ e^{(k)} = -A^{-1}r^{(k)} = -A^{-1}p_k(I - A)r^{(0)} = -A^{-1}p_k(I - A)(Ax^* - Ax^{(0)}) = \\
= A^{-1}p_k(I - A)Ae^{(0)} = p_k(I - A)e^{(0)}. \]

**Corollary 1.77**  Error estimate for the cg-iteration I

For each polynomial \( p_k \in P_k \) with \( p_k(1) = 1 \), the iteration errors \( e^{(k)} = x^{(k)} - x^* \) satisfy the estimate

\[ \|e^{(k)}\|_A \leq \max \{ |p_k(1 - \lambda)| \mid \lambda \in \sigma(A) \} \|e^{(0)}\|_A. \]
Definition 1.78 Chebyshev polynomials

The Chebyshev polynomials $T_k$ of degree $k \in \mathbb{N}_0$ are given by

$$T_k(x) := \begin{cases} 
\cos \left( k \arccos x \right) , & x \in [-1,1] \\
\cosh \left( k \text{ Arcosh } x \right) , & x \in \mathbb{R} \setminus [-1,1] 
\end{cases} , \quad k \in \mathbb{N}_0 .$$

They satisfy the three-term recursion

$$T_0(x) = 1 , \quad T_1(x) = x , \quad T_{k+1}(x) = 2x T_k(x) - T_{k-1}(x) , \quad k \in \mathbb{N} .$$

For $x \in \mathbb{C}$, we have the representation

$$T_k(x) = \frac{1}{2} \left[ \left( x + (x^2 - 1)^{1/2} \right)^k + \left( x + (x^2 - 1)^{1/2} \right)^{-k} \right] .$$
Theorem 1.79  Minimax - problem

Let \([a, b] \subset \mathbb{R}\) such that \(-\infty < a < b < 1\) and \(\tilde{P}_k := \{p_k \in P_k \mid p_k(1) = 1\}\). Then, the minimax - problem

\[
\min_{p_k \in \tilde{P}_k} \max_{\lambda \in [a, b]} |p_k(\lambda)|
\]

is uniquely solvable with the solution given by

\[
p_k(\xi) = T_k\left(\frac{2\xi - a - b}{b - a}\right)/C_k,
\]

\[
\frac{1}{C_k} := \frac{1}{T_k\left(\frac{2 - a - b}{b - a}\right)} = \frac{2c^k}{1 + c^{2k}},
\]

\[
c := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \quad \kappa := \frac{1 - a}{1 - b}.
\]
Proof: We have $C_k := T_k\left(\frac{2-a-b}{b-a}\right) \neq 0$, since $\frac{2-a-b}{b-a} \notin [-1, +1]$ and $\cosh x > 0$. Moreover,

$$
\xi \in [a, b] \implies \frac{2\xi - a - b}{b - a} \in [-1, +1] \implies \max_{\xi \in [a, b]} |T_k\left(\frac{2\xi - a - b}{b - a}\right)| = 1.
$$

In order to prove that $p_k$ is the unique solution of the minimax - problem, let

$$
q_k \in \tilde{P}_k, \quad \max_{\xi \in [a, b]} |q_k(\xi)| \leq \frac{1}{C_k}, \quad z(\xi) := p_k(\xi) - q_k(\xi), \quad \xi \in [a, b].
$$

In view of $T_k(\cos(\nu \pi / k)) = (-1)^\nu$, $\nu = -k, -k + 1, \ldots, 0$, there holds:

$$
p_k(\xi_\nu) = (-1)^\nu / C_k, \quad \xi_\nu := \frac{1}{2} [a + b + (b - a) \cos(\nu \pi / k)] \implies z(\xi_\nu) = \begin{cases} 
> 0, & \nu = 2\mu, \\
\leq 0, & \nu = 2\mu + 1.
\end{cases}
$$

Since $z$ is continuous, in each interval $[\xi_{\nu-1}, \xi_\nu], \ 1 - k \leq \nu \leq 0$, $z$ has a zero (eventually a double zero in $\xi_\nu$), and hence at least $k$ zeroes in $[a, b]$. In view of $p_k(1) = q_k(1) = 1$, $z$ admits a $(k + 1)$-st zero. In view of $z \in P_k$, it follows that $z = 0$, i.e., $q_k = p_k$.

The representation of $1/C_k$ results from the definition of $T_k$ for $|x| > 1$. 
Theorem 1.80  Convergence rates for the cg - iteration
Let \( A \in \mathbb{R}^{n \times n} \) be spd with \( \lambda := \lambda_{\min}(A) \) and \( \Lambda := \lambda_{\max}(A) \). Moreover, let

\[
\kappa := \kappa(A) := \frac{\Lambda}{\lambda}
\]

be the spectral condition of \( A \). Then, there holds:

\[
\|e^{(k)}\|_A \leq \frac{2c^k}{1 + c^{2k}} \|e^{(0)}\|_A , \quad c := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} .
\]

Proof: According to Corollary 1.77, we have:

\[
\|e^{(k)}\|_A \leq \max \{ |p_k(1 - \mu)| \mid \mu \in \sigma(A) \} \|e^{(0)}\|_A , \quad \lambda \leq \mu \leq \Lambda \quad \Rightarrow \quad 1 - \Lambda =: a \leq 1 - \mu \leq b := 1 - \lambda .
\]

Applying Theorem 1.79 with

\[
\kappa = \frac{1 - a}{1 - b} = \frac{\Lambda}{\lambda} ,
\]

gives the assertion.
Remark 1.81 Growth of the spectral condition
Consider the tridiagonal matrix $A \in \mathbb{R}^{n \times n}$ with

$$A := h^{-2} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 \end{pmatrix}, \quad h := 1/(n + 1).$$

$A$ has the eigenvalues:

$$\lambda_i(A) = 4 h^{-2} \sin^2 \frac{i \pi}{2(n + 1)} , \quad 1 \leq i \leq n \quad \Rightarrow$$

$$\lambda_{\max}(A) = \lambda_n(A) = 4 h^{-2} \cos^2 \frac{\pi}{2(n + 1)} , \quad \lambda_{\min}(A) = \lambda_1(A) = 4 h^{-2} \sin^2 \frac{\pi}{2(n + 1)}.$$

It follows that:

$$\kappa(A) = \frac{b \lambda_{\max}(A)}{\lambda_{\min}(A)} = \frac{1 - \sin^2 \frac{\pi}{2(n+1)}}{\sin^2 \frac{\pi}{2(n+1)}} = O(h^{-2}) = O(n^2).$$
Remark 1.82 About preconditioning

Let $A$, $W \in \mathbb{R}^{n \times n}$ be spd and assume that $0 < \gamma \leq \Gamma$ satisfy

\[
\lambda_{\min}(A) \ll \gamma \quad \text{and} \quad \Gamma \ll \lambda_{\max}(A), \quad \text{such that}
\]

\[
\frac{\Gamma}{\gamma} \ll \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \kappa(A) \quad \Rightarrow \quad \gamma I \leq W^{-1/2} A W^{-1/2} \leq \Gamma I.
\]

Idea: Apply the cg - iteration to the transformed linear system:

\[
\underbrace{W^{-1/2} A W^{-1/2}}_{=:\hat{A}} \underbrace{W^{1/2}x}_{=:\hat{x}} = \underbrace{W^{-1/2}b}_{=:\hat{b}}.
\]

Initialization: \( \hat{x}^{0} := W^{1/2}x^{0} \), \( \hat{r}^{0} := \hat{b} - \hat{A}\hat{x}^{0} \), \( \hat{p}^{0} := \hat{r}^{0} \).

Iteration loop: For $k = 0, 1, 2, \ldots$, do:

1. \( \hat{x}^{(k+1)} := \hat{x}^{(k)} + \hat{\alpha} \hat{p}^{(k)} \), \( \hat{\alpha} := \frac{(\hat{r}^{(k)}, \hat{p}^{(k)})}{(\hat{A}\hat{p}^{(k)}, \hat{p}^{(k)})} \),

2. \( \hat{r}^{(k+1)} := \hat{r}^{(k)} - \hat{\alpha} \hat{A}\hat{p}^{(k)} \),

3. \( \hat{p}^{(k+1)} := \hat{r}^{(k+1)} - \hat{\beta} \hat{p}^{(k)} \), \( \hat{\beta} := \frac{(\hat{A}\hat{p}^{(k)}, \hat{r}^{(k+1)})}{(\hat{A}\hat{p}^{(k)}, \hat{p}^{(k)})} \).
Definition 1.83 Preconditioned cg - iteration (pcg - iteration)

With \( \hat{x}^{(k)} = W^{1/2}x^{(k)} \) and \( \hat{p}^{(k)} = W^{1/2}p^{(k)} \) there holds:

\[
W^{-1/2}r^{(k)} = W^{-1/2}(b - Ax^{(k)}) = \hat{b} - \hat{A}\hat{x}^{(k)} = \hat{r}^{(k)} \implies \\
(\hat{r}^{(k)} , \hat{p}^{(k)}) = (W^{-1/2}r^{(k)}, W^{1/2}p^{(k)}) = (r^{(k)} , p^{(k)}) , \\
(\hat{A}\hat{p}^{(k)}, \hat{p}^{(k)}) = (W^{-1/2}Ap^{(k)}, W^{1/2}p^{(k)}) = (A\hat{p}^{(k)}, \hat{p}^{(k)}) , \\
(\hat{A}\hat{p}^{(k)}, \hat{r}^{(k+1)}) = (W^{-1/2}Ap^{(k)}, W^{-1/2}r^{(k+1)}) = (Ap^{(k)}, W^{-1}r^{(k+1)}) .
\]

Therefore, the cg - iteration can be performed w.r.t. the original variables requiring only one additional matrix inversion of \( W \) (\( W \) is called the **preconditioner**):

**Initialization:** Choose a start vector \( x^{(0)} \). Compute: \( r^{(0)} := b - Ax^{(0)} \), \( p^{(0)} := W^{-1}r^{(0)} \), \( \rho_0 := (p^{(0)}, r^{(0)}) \).

**Iteration loop:** For \( k = 0, 1, 2, \ldots \) do:

1. \( a^{(k)} := Ap^{(k)} \) , \( x^{(k+1)} := x^{(k)} + \alpha p^{(k)} \) , \( \alpha := \frac{\rho_k}{a^{(k)} , p^{(k)}} \),
2. \( r^{(k+1)} := r^{(k)} - \alpha a^{(k)} \) , \( q^{(k+1)} := W^{-1}r^{(k+1)} \),
3. \( p^{(k+1)} := q^{(k+1)} + \frac{\rho_{k+1}}{\rho_k} p^{(k)} \) , \( \rho_{k+1} := (q^{(k+1)}, r^{(k+1)}) \).
Theorem 1.84  Convergence rates for the pcg - iteration
Let $A, W \in \mathbb{R}^{n \times n}$ be spd with $\gamma W \leq A \leq \Gamma W$, $0 < \gamma \leq \Gamma$. Then, the iterates $(x^{(k)})_{k=0}^{N_0}$ of the pcg - iteration (with $W$ as the preconditioner) satisfy:

$$\|e^{(k)}\|_A \leq \frac{2c^k}{1 + c^2} \|e^{(0)}\|_A$$

$$c := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \quad \kappa := \frac{\Gamma}{\gamma}.$$

Proof: The proof follows readily by applying Theorem 1.80 to the matrix $W^{-1/2}AW^{-1/2}$.

Remark 1.85  Choice of preconditioners

1. $W := \theta I$  RCG  :  Richardson accelerated cg - iteration,
2. $W := D$  JCG  :  Jacobi accelerated cg - iteration,
3. $W := (D - E)D^{-1}(D - F)$  SGSCG  :  symmetric Gauss Seidel accelerated cg - iteration,
4. $W := \omega^{-1}(D - \omega E)D^{-1}\omega^{-1}(D - \omega F)$  SSORCG  :  symmetric SOR accelerated cg - iteration
1.3 Linear least squares problems
1.3.1 The Gauss method of least squares

Remark 1.85 Problem
Assume a functional dependence between a measurable quantity $b$ and the variables $t$

$$b(t) = \varphi(t; x_1, \ldots, x_n),$$

where the function $\varphi$ depends on $n$ parameters $x_j$, $1 \leq j \leq n$.

For the determination of the parameters, perform $m \geq n$
measurements in $t_i$, $1 \leq i \leq m$, providing data $b_i$, $1 \leq i \leq m$.
Compute the deviations

$$\delta_i := b_i - \varphi(t_i; x_1, \ldots, x_n), \quad 1 \leq i \leq m.$$

Compute $x_1, \ldots, x_n$ in such a way that

$$\sum_{i=1}^{m} \Delta_i^2$$

is minimized. This technique is called the Gauss method of least squares.
Definition 1.86  Linear least squares problem
If the function $\varphi$ is linear in the parameters $x_j$, $1 \leq j \leq n$, i.e.,

$$\varphi(t; x_1, \ldots, x_n) := a_1(t) x_1 + \ldots + a_n(t) x_n(t),$$

then the problem is called a linear least squares problem.
Setting $A \in \mathbb{R}^{m \times n}$ with $a_{ij} := a_j(t_i)$, $1 \leq i \leq m$, $1 \leq j \leq n$, and $b := (b_1, \ldots, b_m)^T$, we have

$$\min_{x_1, \ldots, x_n} \sum_{i=1}^{m} (b_i - \varphi(t_i; x_1, \ldots, x_n))^2 \implies (*) \ \min_{x \in \mathbb{R}^n} \|b - Ax\|_2,$$

where $\| \cdot \|_2$ denotes the Euclidian norm on $\mathbb{R}^n$.

Remark 1.87  Similar problems
If we choose in (*) the $\ell_1$-norm $\| \cdot \|_1$ instead of the Euclidian norm, the problem is called a linear optimization problem. If the maximum - norm $\| \cdot \|_\infty$ is chosen, the problem is called a Chebyshev least squares problem.
1.3.2 Normal equations

Definition 1.88 Orthogonal projection
Let $U \subset \mathbb{R}^m$ be a subspace and $v \in \mathbb{R}^m$. Then, $Pv \in U$ is called the orthogonal projection of $v$ onto $U$, if
\[(+) \quad (Pv - v, u) = 0, \quad u \in U.
\]
The mapping $P : \mathbb{R}^m \to U$, which satisfies $(+)$, is called the orthogonal projection of $\mathbb{R}^m$ onto $U$.

Lemma 1.89 Characterization of the orthogonal projection
Given $v \in \mathbb{R}^m$, the vector $Pv \in U$ is the orthogonal projection of $v$ onto the subspace $U \subset \mathbb{R}^m$, if and only if
\[(++) \quad \|v - Pv\|_2 = \min_{u \in U} \|v - u\|_2.
\]

Proof: For $u \in U$ we have:
\[
\|v - u\|_2^2 = (v - u, v - u) = (v - Pv + Pv - u, v - Pv + Pv - u) = \\
= \|v - Pv\|_2^2 + 2 (v - Pv, Pv - u) + \|Pv - u\|_2^2.
\]
If $Pv$ is the orthogonal projection of $v$ onto $U$, then $(v - Pv, Pv - u) = 0$ and hence,
\[
\|v - u\|_2 \geq \|v - Pv\|_2, \quad u \in U.
\]
On the other hand, suppose that $(++)$ holds true. Then, we obtain:

$$ (v - P v , P v - u ) \geq 0 \ , \ u \in U . $$

Choosing $u = P v \perp z , z \in U$, it follows that

$$ (v - P v , \perp z ) 0 , z \in U \quad \Rightarrow \quad (v - P v , z ) = 0 , z \in U . $$

**Lemma 1.90** Characterization of the solution of the linear least squares problem $x \in \mathbb{R}^n$ is the solution of the linear least squares problem $(\ast)$, if and only if $Ax \in \mathbb{R}^m$ is the orthogonal projection of $b \in \mathbb{R}^m$ onto $\text{R}(A)$, i.e.,

$$ (b - Ax , Ay ) = 0 \ , \ y \in \mathbb{R}^n . $$

**Theorem 1.91** Normal equations $x \in \mathbb{R}^n$ is the solution of the linear least squares problem $(\ast)$, if and only if $x$ satisfies the normal equations

$$ A^T A x = A^T b . $$
1.3.3 Orthogonalization methods
Definition 1.93 QR - decomposition of matrices
Let \( A \subset \mathbb{R}^{m \times n} \), \( m \geq n \), and \( Q \in \mathbb{R}^{m \times m} \) be an orthogonal matrix such that

\[
(**) \quad Q^T A = \begin{pmatrix} R \\ 0 \end{pmatrix} \iff A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}
\]

with an upper triangular matrix \( R \in \mathbb{R}^{n \times n} \). Then, \((*)\) is called a QR - decomposition of \( A \).

Theorem 1.94 Solution of the linear least squares problem by QR - decomposition
Let \( A \subset \mathbb{R}^{m \times n} \), \( m \geq n \), \( \text{Rang} \ A = n \), \( b \in \mathbb{R}^m \), and let \( Q \in \mathbb{R}^{m \times m} \) be an orthogonal matrix such that

\[
Q^T A = \begin{pmatrix} R \\ 0 \end{pmatrix}, \quad Q^T b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},
\]

where \( b_1 \in \mathbb{R}^n \), \( b_2 \in \mathbb{R}^{m-n} \) and \( R \in \mathbb{R}^{n \times n} \) is a regular upper triangular matrix. Then, the solution of the linear least squares problem \((*)\) is given by

\[
x = R^{-1} b_1.
\]
Proof: We have:

\[ \|b - Ax\|_2^2 = \|Q^T(b - Ax)\|_2^2 = \left\| \begin{pmatrix} b_1 - Rx \\ b_2 \end{pmatrix} \right\|_2^2 = \|b_1 - Rx\|_2^2 + \|b_2\|_2^2. \]

Obviously, \( \|b - Ax\|_2 \) attains its minimum for \( x = R^{-1}b_1 \).

Remark 1.95 Construction of orthogonal transformations by reflections

Reflection of \( a \in \mathbb{R}^2 \) on the line \( \ell \) perpendicular to \( v \in \mathbb{R}^2 \):

\[
\sin \left( \frac{\pi}{2} - \alpha \right) = \frac{\kappa}{\|a\|_2} = \cos \alpha = \frac{(v, a)}{\|v\|_2 \|a\|_2} \quad \Rightarrow
\]

\[
\kappa = \frac{(v, a)}{\|v\|_2}.
\]

We thus obtain:

\[
a \mapsto \|a\|_2 e_1 = Qa := a - 2 \kappa \frac{v}{\|v\|_2} = a - 2 \frac{(v, a)}{(v, v)} v = (I - 2 \frac{vv^T}{v^Tv}) a.
\]
Definition 1.96  Householder reflections

The matrix

\[ Q := I - 2 \frac{vv^T}{v^Tv} \]

is called Householder reflection. It has the following properties

1. Q is symmetric: \( Q^T = Q \),
2. Q is orthogonal: \( QQ^T = Q^TQ = I \),
3. Q is idempotent: \( Q^2 = I \).
Remark 1.97 Implementation of the Householder reflections

For \( v := a - \alpha e_1 \), \( \alpha := \|a\|_2 \) we get:

\[
v^T v = a^T a - 2 \alpha a^T e_1 + \alpha^2 = 2 \alpha^2 - 2 \alpha a_1 = 2 \alpha (\alpha - a_1) .
\]

For \( x \in \mathbb{R}^m \), it follows that:

\[
Qx = x - \frac{v^T x}{\alpha (\alpha - a_1)} v .
\]

If \( \alpha = (\sum_{i=1}^{m} a_i^2)^{1/2} \sim a_1 \), cancellation of essential digits may occur. In this case, we map \( a \) onto \(-\alpha e_1\).

Therefore, it is advisable always to choose

\[
\alpha = - \text{sgn}(a_1) \|a\|_2
\]
Remark 1.98  Implementation of the QR-decomposition of a matrix

The QR-decomposition of a matrix $A \in \mathbb{R}^{m \times n}$, $m \geq n$, is realized as follows:

1. Step: Choose $v_1 := a_1 - \alpha_1 e_1$

2. Step: Choose $\bar{Q}_2 := I - 2 \frac{(v_2 v_2^T)/(v_2^T v_2)}{\|v_2\|^2}$

where $v_2 := t_1^{(2)} - \alpha_2 e_1$, $\alpha_2 := -\text{sgn}(t_1^{(2)}) \|t_1^{(2)}\|_2$.

After $p = \min(m - 1, n)$ steps, we obtain the upper triangular matrix

$$
\begin{pmatrix}
R \\
0
\end{pmatrix} = Q_p \ldots Q_1 A.
$$
**Remark 1.99  Storage for the QR-decomposition of a matrix**

During the implementation of the Householder method, the diagonal elements $r_{ii}$, $1 \leq i \leq n$, are stored in a separate vector such that the Householder vectors $v_i$, $1 \leq i \leq p$, can be stored in the lower part of $A$ ("parquetting"): 

![Diagram showing the storage of Householder vectors in the lower part of $A$](image)