„FŐNIX ME" – Megújuló Egyetem felsőoktatási intézményi fejlesztések a felsőfokú oktatás minőségének és hozzáférhetőségének együttes javítása érdekében
EFOP 3.4.3-16-2016-00015

Numerical Methods and Optimization

Készítette:
Körei Attila

A tananyagfejlesztést az EFOP-3.4.3-16-2016-00015 számú, "Megújuló egyetem- felsőoktatási intézményi fejlesztések a felsőoktatás minőségének és hozzáférhetőségének együttes javítása érdekében" elnevezésű projekt 5. Nemzetköziesítés részprojektjében támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap és Magyarország költségvetése társfinanszírozásában valósul meg.

Contents

1 Preliminaries: Vectors and Matrices

Definition 1.1 (Real Vector) An n-dimensional real vector is an ordered set of n real numbers $(x_1, x_2, \ldots, x_n)$ and is usually written in the column vector form

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

We will also write $x = [x_i]_{i=1}^n$. The numbers $x_1, x_2, \ldots, x_n$ are called the components or coordinates of x.

Definition 1.2 (Real Coordinate Space) The set of all n-dimensional vectors is called the n-dimensional real coordinate space and is denoted by $\mathbb{R}^n$.

Definition 1.3 (Special Vectors) The zero vector $0 \in \mathbb{R}^n$ has only zero components: $0 = [0]_{i=1}^n$. The unit vector $e^{(i)} \in \mathbb{R}^n$ has its ith component 1, and the others are zero:

$$e_j^{(i)} = \begin{cases} 1, & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Definition 1.4 (Operations with Vectors) Given vectors $x = [x_i]_{i=1}^n$ and $y = [y_i]_{i=1}^n \in \mathbb{R}^n$, their sum is defined by

$$x + y = [x_i + y_i]_{i=1}^n.$$

For a vector $x \in \mathbb{R}^n$ and a scalar $\alpha \in \mathbb{R}$, the scalar multiplication αx is defined as

$$\alpha x = [\alpha x_i]_{i=1}^n.$$

Definition 1.5 (Linear Combination) Given vectors $v^{(1)}, v^{(2)}, \ldots, v^{(k)} \in \mathbb{R}^n$ and scalars $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}$, a sum in the form

$$\alpha_1 v^{(1)} + \alpha_2 v^{(2)} + \cdots + \alpha_k v^{(k)}$$

is called a linear combination of $v^{(1)}, v^{(2)}, \ldots, v^{(k)}$.

Definition 1.6 (Linear Indepence) Given vectors $v^{(1)}, v^{(2)}, \ldots, v^{(k)} \in \mathbb{R}^n$ are called linearly independent if

$$\alpha_1 v^{(1)} + \alpha_2 v^{(2)} + \cdots + \alpha_k v^{(k)} = 0$$

implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$. Otherwise, the vectors are linearly dependent.

Definition 1.7 (Basis) Consider n vectors $v^{(1)}, v^{(2)}, \dots, v^{(n)}$ of the real coordinate space $\mathbb{R}^n$. We say, that vectors $v^{(1)}, v^{(2)}, \dots, v^{(n)}$ form a basis of $\mathbb{R}^n$ if the vectors are linearly independent.

Remark 1.1 *If vectors $v^{(1)}, v^{(2)}, \dots, v^{(n)}$ form a basis of $\mathbb{R}^n$, then any vector x can be expressed by a linear combination of the basis vectors as*

$$x = x_1 v^{(1)} + x_2 v^{(2)} + \cdots + x_n v^{(n)}.$$

The numbers $x_1, x_2, \dots, x_n$ are the coordinates of x with respect to the basis $v^{(1)}, v^{(2)} \dots, v^{(n)}$.

Definition 1.8 (Matrix) A real matrix is a rectangular array of real numbers composed of rows and columns. The element in row i and column j is denoted by a_{ij}. The set of real matrices of order $m \times n$ is denoted by $\mathbb{R}^{m\times n}$. If $A \in \mathbb{R}^{m\times n}$, its elements are

$$A = [a_{ij}]_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

Note that vectors are special matrices, $\mathbb{R}^{m\times 1}$ is the set of all m-dimensional column vectors.

Definition 1.9 (Special Matrices)

- A matrix $0 \in \mathbb{R}^{m\times n}$ is called a zero or null matrix, if all of its elements are zero.
- A square matrix of order n is a matrix with n rows and n columns.
- A square matrix $A = [a_{ij}]_{n\times n}$ is called upper triangular if $a_{ij} = 0$ for all $i > j$.
- A square matrix $A = [a_{ij}]_{n\times n}$ is called lower triangular if $a_{ij} = 0$ for all $i < j$.
- Matrix A is said to be diagonal if $a_{ij} = 0$ when $i \neq j$.
- The identity matrix of order n is a square matrix $I_n = [\delta_{ij}]_{n\times n}$, where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

  In other words, the identity matrix is a square matrix which has ones on the main diagonal and zeros elsewhere.

Definition 1.10 (Matrix Operations) Given two matrices, $A = [a_{ij}]_{m\times n}$ and $B = [b_{ij}]_{m\times n} \in \mathbb{R}^{m\times n}$, their sum is defined as

$$A + B = [a_{ij} + b_{ij}]_{m\times n} \in \mathbb{R}^{m\times n}.$$

If $\alpha \in \mathbb{R}$ is a scalar, the scalar multiplication αA is defined as

$$\alpha A = [\alpha a_{ij}]_{m\times n}.$$

Given $A \in \mathbb{R}^{m\times n}$, and $B \in \mathbb{R}^{n\times k}$ the matrix product AB is defined as $AB = C = [c_{ij}]_{m\times k}$, where

$$c_{ij} = \sum_{l=1}^{n} a_{il} b_{lj}.$$

Matrix multiplication AB is not defined if the number of columns A is not equal to the number of rows in B.

Definition 1.11 (Matrix Transpose) Given a matrix $A \in \mathbb{R}^{m\times n}$, the transpose of A is an $n\times m$ matrix whose rows are the columns of A. The transpose matrix of A is denoted by A^T.

Definition 1.12 (Symmetric Matrix) A square matrix $A \in \mathbb{R}^{n\times n}$ is said to be symmetric if $A^T = A$.

Theorem 1.1 (Rules for Transposes) *Given matrices A and B we have*

(i) $(A^T)^T = A$;

(ii) $(A + B)^T = A^T + B^T$ *(if the sum $A + B$ is defined);*

(iii) $(AB)^T = B^T A^T$ *(if the product AB is defined).*

Definition 1.13 (Matrix Determinant) For a square matrix $A = [a_{ij}]_{n\times n}$ the determinant of A is a real number, which is denoted by $\det(A)$ and is defined as follows.

- If $n = 1$ and $A = a_{11}$, then $\det(A) = a_{11}$.
- If $n \geq 2$, then

  $$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \cdot \det(A_{ij}),$$

  where A_{ij} is the matrix obtained from A removing its ith row and jth column. This formula is called the ith row expansion where index i can be chosen arbitrarily.

Definition 1.14 (Matrix Trace) The trace $\mathrm{t}r(A)$ of a matrix $A = [a_{ij}]_{n\times n}$ is the sum of its diagonal elements, i.e.,

$$\mathrm{t}r(A) = a_{11} + a_{22} + \cdots + \alpha_{nn} = \sum_{j=1}^{n} a_{jj}.$$

Definition 1.15 (Matrix Inverse) A matrix $A = [a_{ij}]_{n\times n}$ is said to be nonsingular if there exists some matrix A^{-1} such that $AA^{-1} = A^{-1}A = I_n$. The matrix A^{-1} is called the multiplicative inverse of A. If A^{-1} does not exist, then A is said to be singular.

Theorem 1.2 (Properties of the Inverse) *Let A and B be nonsingular matrices of the same order. Then*

(i) $(A^{-1})^{-1} = A$;

(ii) $(A^{-1})^T = (A^T)^{-1}$;

(iii) $(AB)^{-1} = B^{-1}A^{-1}$.

Definition 1.16 (Norm) A function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, denoted by

$$f(x) = ||x||$$

is called a matrix norm (a vector norm in case $n = 1$), if for any $x, y \in \mathbb{R}^{m\times n}$ and any scalar $\lambda \in \mathbb{R}$ the following properties hold:

(i) $||x|| \geq 0$ with equality if and only if $x = 0$;

(ii) $||\lambda x|| = |\lambda| \cdot ||x||$;

(iii) $||x + y|| \leq ||x|| + ||y||$.

Let $x = [x_i]_{i=1}^n$. The most commonly used vector norms are the following:

- 1-norm: $||x||_1 = \sum_{i=1}^{n} |x_i|$;
- 2-norm: $||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$;
- ∞-norm: $||x||_\infty = \max_{1\leq i\leq n} |x_i|$.

Let $A = [a_{ij}]_{m\times n} \in \mathbb{R}^{m\times n}$. The most commonly used matrix norms are the following:

- 1-norm: $||A||_1 = \max_{1 \le j \le n} \sum_{i=1}^{m} |a_{ij}|$;
- 2-norm: $||A||_2 =$ the square root of the largest eigenvalue of $A^T A$ (see later in Chapter 3);
- ∞-norm: $||A||_\infty = \max_{1 \le i \le m} \sum_{j=1}^{n} |a_{ij}|$;
- Frobenius norm: $||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$.

2 Solving a System of Linear Equations

The matrix form of an $m \times n$ system of linear algebraic equations is $Ax = b$, where

$$\begin{aligned} A &\in \mathbb{R}^{m \times n} \text{ the matrix of the coefficients,} \\ x &\in \mathbb{R}^{n} \text{ the vector of the unknowns and} \\ b &\in \mathbb{R}^{m} \text{ the vector of the constant terms.} \end{aligned}$$

So, the system consists of m equations and n unknowns, where the ith equation has the form

$$a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i, \quad \text{and} \quad i = 1, 2, \ldots, m.$$

If all of the constant terms are zero ($b_i = 0,\ i = 1, 2, \ldots, m$), then the system is called homogeneous (otherwise inhomogeneous).

Definition 2.1 (Solution of a System) A system of linear equations is called consistent, if there exists at least one vector x^*, for which $Ax^* = b$. The vector x^* is called a solution of the system. A system of linear equations, which has no solution is called inconsistent.

There are three different possible outcomes when solving a system of linear equations:

- no solution (inconsistent system),
- unique solution (consistent system),
- infinitely many solutions (consistent system).

If A is a square matrix ($A \in \mathbb{R}^{n\times n}$), then the following statements are equivalent:

a) The system $Ax = b$ has a unique solution.

b) The matrix A is nonsingular (i.e.: it has an inverse matrix, A^{-1}).

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

Any of the statements is satisfied, the unique solution x can be obtained in the following way:

$$x = A^{-1}b.$$

In case of homogeneous systems ($b = 0$), the unique solution is $x = 0$, when A is nonsingular. If A is singular, the homogeneous system has infinitely many solutions.

2.1 Direct Methods for Solving Systems of Linear Equations

2.1.1 Gaussian Elimination

The method consists of two phases. In the first phase we transform the system into triangular form, then, in the second phase we compute x by back substitution.

Example 2.1 Solve the following system of linear equations:

$$\begin{aligned} 2x_1 - 5x_2 - 3x_3 &= 4 \\ x_1 + 2x_2 - 2x_3 &= -14 \\ 5x_1 + x_2 \quad + x_3 &= 12 \end{aligned}$$

First we leave the unknowns x_1, x_2, x_3, and write the coefficients and the constant terms into the form:

$$\begin{bmatrix} \boxed{2} & -5 & -3 & | & 4 \\ 1 & 2 & -2 & | & -14 \\ 5 & 1 & 1 & | & 12 \end{bmatrix}$$

We use the pivoting technique to transform the system into triangular form. The first pivotal element is $a_{11} = 2$, which is framed. Rules of pivoting:

- the pivotal row (the row of the pivotal element) remains the same

- the pivotal column contains zeros under the pivotal element
- any other element is obtained by the rectangle rule:

$$\begin{array}{ccc} \boxed{\text{p}} & - - - - - - - - - & a \\ | & & | \\ b & - - - - - - - - - - & x \end{array}$$

instead of x we have to write $x - \dfrac{ab}{p}$.

After the first pivoting step we have

$$\begin{bmatrix} 2 & -5 & -3 & | & 4 \\ 0 & \boxed{4.5} & -0.5 & | & -16 \\ 0 & 13.5 & 8.5 & | & 2 \end{bmatrix}$$

The next pivotal element is $a_{22} = 4.5$. The first two rows are unchanged, we have 0 under the pivotal element, and the remained two numbers are computed by the rectangle rule:

$$\begin{bmatrix} 2 & -5 & -3 & | & 4 \\ 0 & 4.5 & -0.5 & | & -16 \\ 0 & 0 & 10 & | & 50 \end{bmatrix}$$

We obtained a triangular system, which can be solved by back substitution. The last row means $10x_3 = 50$, from which $x_3 = 5$. The second row of the triangular system is $4.5x_2 - 0.5x_3 = -16$. Substituting $x_3 = 5$ yields $4.5x_2 - 2.5 = -16$, from which $x_2 = -3$. The first equation is $2x_1 - 5x_2 - 3x_3 = 4$. Using the known values of x_2 and x_3, we get $2x_1 + 15 - 15 = 4$, from which $x_1 = 2$.

So the unique solution of the system can be expressed by the vector

$$x = \begin{bmatrix} 2 \\ -3 \\ 5 \end{bmatrix}.$$

When using the rectangle rule, we have to divide by the pivotal element, that is why it can not be zero. In addition a pivotal element which are very small in absolute value can cause undesirable round-off errors. To illustrate this phenomenom consider the following example.

Example 2.2 Solve the system

$$\begin{aligned} 10^{-5}x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

using four-digit arithmetic.

The first phase of the Gaussian method:

$$\begin{bmatrix} 10^{-5} & 1 & | & 1 \\ 1 & 1 & | & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 10^{-5} & 1 & | & 1 \\ 0 & -99999 & | & -99998 \end{bmatrix}$$

The second phase is the back substitution:

$$x_2 = \frac{-999998}{-99999} = 1 \ \ \text{(rounding to four digits)}.$$

Replacing $x_2 = 1$ in the first equation, we have $x_1 = 0$. The obtained solution is of bad quality, because $x_1 + x_2 = 1$, instead of 2, as prescribed in the second equation of the system.

Solve the problem again after switching the rows:

$$\begin{bmatrix} 1 & 1 & | & 2 \\ 10^{-5} & 1 & | & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 1 & | & 2 \\ 0 & 0.99999 & | & 0.99998 \end{bmatrix}$$

The back substitution gives: $x_2 = \frac{0.99998}{0.99999} = 1$. Replacing $x_2 = 1$ in the first equation, we have $x_1 = 1$, which is a much more better solution than the previous one.

Partial Pivoting Strategy

To avoid the problem demonstrated in the previous example, we have to choose the actual pivotal element a_{kk} such that

$$| \, a_{kk} \, | = \max_{i \geq k} | \, a_{ik} \, | \, .$$

This partial pivoting strategy can be achieved by switching the rows, if needed.

Consider the system given in Example 4.1 and solve it using partial pivoting strategy.

$$\begin{aligned} 2x_1 - 5x_2 - 3x_3 &= 4 \\ x_1 + 2x_2 - 2x_3 &= -14 \\ 5x_1 + x_2 \quad + x_3 &= 12 \end{aligned}$$

First we have to switch row 1 and row 3:

$$\left[\begin{array}{ccc|c} \boxed{5} & 1 & 1 & 12 \\ 1 & 2 & -2 & -14 \\ 2 & -5 & -3 & 4 \end{array}\right] \to \left[\begin{array}{ccc|c} 5 & 1 & 1 & 12 \\ 0 & 1.8 & -2.2 & -16.4 \\ 0 & -5.4 & -3.4 & -0.8 \end{array}\right]$$

We have to switch row 2 and row 3, because $\mid -5.4 \mid > \mid 1.8 \mid$.

$$\left[\begin{array}{ccc|c} 5 & 1 & 1 & 12 \\ 0 & \boxed{\text{-5.4}} & -3.4 & -0.8 \\ 0 & 1.8 & -2.2 & -16.4 \end{array}\right] \to \left[\begin{array}{ccc|c} 5 & 1 & 1 & 12 \\ 0 & -5.4 & -3.4 & -0.8 \\ 0 & 0 & -3.3333 & -16.6666 \end{array}\right]$$

With back substitution we get the solution $x_3 = 5$, $x_2 = -3$, $x_1 = 2$, as before.

2.1.2 Gauss-Jordan Method and Matrix Inversion

This method is used in case of square coefficient matrices. The aim is to transform A into a diagonal matrix, then the solution of the system can be read off easily.

Consider again the problem given Example 2.1. After the first pivoting step we got

$$\left[\begin{array}{ccc|c} 2 & -5 & -3 & 4 \\ 0 & \boxed{4.5} & -0.5 & -16 \\ 0 & 13.5 & 8.5 & 2 \end{array}\right]$$

The next pivotal element is $a_{22} = 4.5$. Write zeros in the column of the pivotal element, and use pivoting backwards, in the first row as well:

$$\left[\begin{array}{ccc|c} 2 & 0 & -3.5555 & -13.7777 \\ 0 & 4.5 & -0.5 & -16 \\ 0 & 0 & \boxed{10} & 50 \end{array}\right]$$

The last pivotal element is 10. Write zeros over it, and apply the rectangle rule in the fourth column:

$$\left[\begin{array}{ccc|c} 2 & 0 & 0 & 4 \\ 0 & 4.5 & 0 & -13.5 \\ 0 & 0 & 10 & 50 \end{array}\right]$$

As the left-hand-side matrix is diagonal, we obtain the solution easily. The first row means $2x_1 = 4$, from which $x_1 = 2$. From the second row $x_2 = \frac{-13.5}{4.5} = -3$, and finally $x_3 = \frac{50}{10} = 5$.

The Gauss-Jordan method is suitable for computing the inverse of a nonsingular matrix.

Let $e^{(i)}$ be the ith unit vector and consider the system $Ax = e^{(i)}$, where A is a nonsingular matrix. Multiplying both sides with A^{-1} we have $x = A^{-1}e^{(i)}$, which means that x is the ith column of the inverse matrix A^{-1}. Therefore in order to obtain the columns of A^1, we have to solve the linear systems

$$Ax = e^{(1)}; \qquad Ax = e^{(2)}; \quad \dots \quad Ax = e^{(n)}.$$

We can solve these systems simultaneously, using the Gauss-Jordan method.

Example 2.3 Find the inverse of matrix $A = \begin{bmatrix} 5 & 2 \\ 4 & 2 \end{bmatrix}$.

We solve two linear systems at the same time with $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$ on the right hand side. We make the operations simultaneously by the two systems as follows:

$$\begin{bmatrix} \boxed{5} & 2 & | & 1 & 0 \\ 4 & 2 & | & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 5 & 2 & | & 1 & 0 \\ 0 & \boxed{0.4} & | & -0.8 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} 5 & 0 & | & 5 & -5 \\ 0 & 0.4 & | & -0.8 & 1 \end{bmatrix}$$

To obtain the solution of the systems (the elements of the inverse matrix) we have to divide the elements of the ith row by the ith pivotal element. The inverse matrix is the result of this division on the right hand side:

$$A^{-1} = \begin{bmatrix} 1 & -1 \\ -2 & 2.5 \end{bmatrix}.$$

We can check the solution by testing the equation $A \cdot A^{-1} = I_2$, where I_2 is the identity matrix of order 2:

$$\begin{bmatrix} 5 & 2 \\ 4 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 & -1 \\ -2 & 2.5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

2.1.3 Triangular Factorization (LU Factorization)

Let $A \in \mathbb{R}^{n \times n}$. The triangular (or LU) factorization of A is a representation of A as the product of a lower and an upper triangular matrices: $A = L \cdot U$.

Theorem 2.1 *A has an LU factorization if and only if all of its main minors are nonsingular that is* $\det(A(1:r, 1:r)) \neq 0, \quad r = 1, 2, \dots, n,$ *where* $A(1:r, 1:r)$ *denotes the matrix containing the elements of matrix A from the first r rows and columns.*

Theorem 2.2 *If A is a nonsingular matrix, then there exits a P permutation matrix for which the product PA has an LU factorization. (It means, that after row exchanges any nonsingular matrix can be factorized.)*

Example 2.4 The matrix $A = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ has no LU factorization, because its left upper element is zero, contradicting the condition $\det(A(1:1, 1:1)) \neq 0$ as required in Theorem 2.1. However, since A is a nonsingular matrix,

$$PA = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} = L \cdot U$$

can be represented in LU-form.

Remark 2.1 *The LU factorization of a matrix is not unique, but if it exists, we have a representation, where L is a unit triangular matrix, i.e. its diagonal contains only ones.*

Example 2.5 Give the LU factorization of matrix $A = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 0 & 2 \end{bmatrix}$.

The U-part can be obtained by Gauss elimination:

$$\begin{bmatrix} \boxed{5} & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 0 & 2 \end{bmatrix} \to \begin{bmatrix} 5 & 2 & 3 \\ 0 & \boxed{0.2} & 2.8 \\ 0 & -1.2 & 0.2 \end{bmatrix} \to \begin{bmatrix} 5 & 2 & 3 \\ 0 & 0.2 & 2.8 \\ 0 & 0 & \boxed{17} \end{bmatrix}$$

The L-part can be determined as follows. Consider the above steps of Gauss elimination and divide the first column of A by the pivotal element 5. The resulting vector

$$\begin{bmatrix} 1 \\ 0.4 \\ 0.6 \end{bmatrix}$$

is the first column of L.

In the next step the pivotal element is 0.2. Divide the second column by 0.2, excepting the element above the pivotal element, instead of it write 0 and we have the second column of L as

$$\begin{bmatrix} 0 \\ 1 \\ -6 \end{bmatrix}.$$

The third column can be obtained from the third step of the Gauss elimination: write zeros above the pivotal element 17 and divide by 17 the pivotal element itself. Therefore the third column of L is

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

So, $L = \begin{bmatrix} 1 & 0 & 0 \\ 0.4 & 1 & 0 \\ 0.6 & -6 & 1 \end{bmatrix}$ is a unit triangular matrix and $A = L \cdot U$ (check it!).

2.1.4 Using LU Factorization in Solving Linear Systems

Suppose that $A = LU$, and we have to solve the system $Ax = b$. We can reduce our work for solving two triangular systems as follows:

- Replace $A = LU$ in $Ax = b$ and we have $LUx = b$.
- Denote Ux by y, and solve the system $Ly = b$ for y.
- Solve the system $Ux = y$ for x.

If A has no LU factorization, but PA has, then first exchange the rows of the system regarding to P: $PAx = Pb$. Since $PA = LU$, we have $LUx = Pb$. Find the solution y of $Ly = Pb$, then compute the solution x of $Ux = y$.

Example 2.6 Solve the following system using LU factorization.

$$\begin{aligned} 5x_1 + 2x_2 + 3x_3 &= 5 \\ 2x_1 + x_2 + 4x_3 &= 7 \\ 3x_1 \qquad + 2x_3 &= 7 \end{aligned}$$

The LU factorization of matrix A was computed previously:

$$A = LU = \begin{bmatrix} 5 & 2 & 3 \\ 2 & 1 & 4 \\ 3 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0.4 & 1 & 0 \\ 0.6 & -6 & 1 \end{bmatrix} \begin{bmatrix} 5 & 2 & 3 \\ 0 & 0.2 & 2.8 \\ 0 & 0 & 17 \end{bmatrix}$$

First solve $Ly = b$:

$$\begin{aligned} y_1 \qquad\qquad &= 5 \\ 0.4y_1 + y_2 \qquad &= 7 \\ 0.6y_1 - 6y_2 + y_3 &= 7 \end{aligned}$$

The solution is: $y_1 = 5,\ y_2 = 5,\ y_3 = 34.$
Then solve $Ux = y$:

$$\begin{aligned} 5x_1 + 2x_2 + 3x_3 &= 5 \\ 0.2x_2 + 2.8x_3 &= 5 \\ 17x_3 &= 34 \end{aligned}$$

The solution of the system is: $x_3 = 2;\ x_2 = -3,\ x_1 = 1.$

2.1.5 Cholesky Decomposition (a Special LU Factorization)

Suppose that A is a symmetric and positive definite matrix (it means that $A^T = A$ and $x^T Ax > 0$ for all nonzero x), then there exists a lower triangular matrix L, for which $A = L \cdot L^T$.

Example 2.7 The matrix $A = \begin{bmatrix} 25 & 15 & 5 \\ 15 & 13 & 5 \\ 5 & 5 & 18 \end{bmatrix}$ is symmetric and positive definite. It can be represented in the form $A = L \cdot L^T$, where $L = \begin{bmatrix} 5 & 0 & 0 \\ 3 & 2 & 0 \\ 1 & 1 & 4 \end{bmatrix}$.

Remark 2.2 *Cholesky decomposition can be effectively used for solving the least square's problem (see Chapter 4.2), where the system to be solved has a symmetric, positive definite matrix of coefficients.*

2.2 Iterative Methods for Solving Systems of Linear Equations

Denote by x^* the exact solution of the system $Ax = b$. The aim is to construct a sequence of vectors $x^{(0)}, x^{(1)}, \ldots, x^{(k)} \ldots$ such that $x^{(k)} \to x^*$ as $k \to \infty$, which means that

$$\lim_{k\to\infty} \| x^* - x^{(k)} \| = 0$$

in some vector norm.

2.2.1 Jacobian Method

First transform the system $Ax = b$ into the form $x = Bx + c$ (for example by expressing the unknown x_i from equation i).

Theorem 2.3 *If* $||B||_\infty = q < 1$, *then the sequence* $x^{(k)} = Bx^{(k-1)} + c$ *converges the unique solution* x^* *of the system* $Ax = b$ *for arbitrarily selected vector* $x^{(0)}$. *In addition we have the error estimation*

$$|| x^* - x^{(k)} ||_\infty \leq \frac{q}{1-q} || x^{(k)} - x^{(k-1)} ||_\infty .$$

Remark 2.3 *The condition* $||B||_\infty < 1$ *is fulfilled, if* A *is strictly diagonally dominant by rows, that is*

$$| a_{ii} | > \sum_{j=1, j\neq i}^{n} | a_{ij} | \qquad\qquad i = 1, 2, \ldots, n.$$

Similar condition can be formulated by columns.

Example 2.8 Consider the system

$$\begin{aligned} 5x_1 + x_2 &= 14 \\ x_1 + x_2 &= 10 \end{aligned}$$

The matrix of the system is $A = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix}$.
This matrix is strictly diagonally dominant, since $5 > 1$ and $2 > 1$. This guarantees the convergence of the Jacobian method.
First express x_1 from the first, and x_2 from the second equation:

$$\begin{aligned} x_1 &= \qquad -0.2x_2 + 2.8 \\ x_2 &= -0.5x_1 \qquad + 5, \end{aligned}$$

therefore

$$B = \begin{bmatrix} 0 & -0.2 \\ -0.5 & 0 \end{bmatrix} \text{ and } c = \begin{bmatrix} 2.8 \\ 5 \end{bmatrix}$$

in the transformed system $x = Bx + c$. The norm of matrix B is $q = ||B||_\infty = 0.5 < 1$.

We choose $x^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ for the initial vector.

$$x^{(1)} = Bx^{(0)} + c = \begin{bmatrix} 0 & -0.2 \\ -0.5 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2.8 \\ 5 \end{bmatrix} = \begin{bmatrix} 2.6 \\ 4.5 \end{bmatrix},$$

$$x^{(2)} = Bx^{(1)} + c = \begin{bmatrix} 0 & -0.2 \\ -0.5 & 0 \end{bmatrix} \begin{bmatrix} 2.6 \\ 4.5 \end{bmatrix} + \begin{bmatrix} 2.8 \\ 5 \end{bmatrix} = \begin{bmatrix} 1.9 \\ 3.7 \end{bmatrix},$$

$$x^{(3)} = Bx^{(2)} + c = \begin{bmatrix} 2.06 \\ 4.05 \end{bmatrix}, \quad x^{(4)} \begin{bmatrix} 1.99 \\ 3.97 \end{bmatrix}, \quad x^{(5)} = \begin{bmatrix} 2.006 \\ 4.005 \end{bmatrix}.$$

Estimating the error of $x^{(5)}$:

$$|| x^* - x^{(5)} ||_\infty \leq \frac{0.5}{1 - 0.5} || x^{(5)} - x^{(4)} ||_\infty = \left|\left| \begin{matrix} 2.006 - 1.99 \\ 4.005 - 3.97 \end{matrix} \right|\right|_\infty = 0.035.$$

The true solution is $x_1 = 2, \ x_2 = 4$, so the real error is 0.006.

2.2.2 Gauss-Seidel Method

The ith component of vector $x^{(k)}$ in the Jacobian iteration $x^{(k)} = Bx^{(k-1)} + c$ is

$$x_i^{(k)} = \sum_{j=1}^{n} b_{ij} x_j^{(k-1)} + c_i.$$

Divide the sum into two parts as follows:

$$x_i^{(k)} = \sum_{j=1}^{i-1} b_{ij} x_j^{(k-1)} + \sum_{j=i}^{n} b_{ij} x_j^{(k-1)} + c_i.$$

Observe, that in the first sum instead of $x_j^{(k-1)}$ we can use the known kth iterations for the components x_j, obtaining

$$x_i^{(k)} = \sum_{j=1}^{i-1} b_{ij} x_j^{(k)} + \sum_{j=i}^{n} b_{ij} x_j^{(k-1)} + c_i. \tag{1}$$

Consider Example 2.8 again. The rules defining variables $x_1, \ x_2$:

$$\begin{aligned} x_1 &= -0.2x_2 + 2.8 \\ x_2 &= -0.5x_1 + 5. \end{aligned}$$

Using (1), these equations can be written in the iterative form

$$\begin{aligned} x_1^{(k)} &= -0.2x_2^{(k-1)} + 2.8 \\ x_2^{(k)} &= -0.5x_1^{(k)} + 5. \end{aligned}$$

If the initial vector is $x^{(0)} = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$, then

$$
\begin{aligned}
x_1^{(1)} &= -0.2x_2^{(0)} + 2.8 = 2.6 \\
x_2^{(1)} &= -0.5x_1^{(1)} + 5 = -0.5 \cdot 2.6 + 5 = 3.7, \text{ so } x^{(1)} = \begin{bmatrix} 2.6 & 3.7 \end{bmatrix}^T. \\
x_1^{(2)} &= -0.2 \cdot 3.7 + 2.8 = 2.06 \\
x_2^{(2)} &= -0.5 \cdot 2.06 + 5 = 3.97, \qquad x^{(2)} = \begin{bmatrix} 2.06 & 3.97 \end{bmatrix}^T. \\
x_1^{(3)} &= -0.2 \cdot 3.97 + 2.8 = 2.006 \\
x_2^{(3)} &= -0.5 \cdot 2.006 + 5 = 3.997, \qquad x^{(3)} = \begin{bmatrix} 2.006 & 3.997 \end{bmatrix}^T.
\end{aligned}
$$

It can be stated that for this system the Gauss-Jordan method converges faster than the Jacobian method.

3 Eigenvalues and Eigenvectors

Definition 3.1 Let $A \in \mathbb{R}^{n \times n}$. A scalar $\lambda \in \mathbb{C}$ is called an eigenvalue of A if there exists a nonzero vector v, such that

$$Av = \lambda v. \tag{2}$$

The vector v is called an eigenvector of A corresponding to λ.

In order to determine the eigenvalues of a matrix A we have to find real or complex numbers $\lambda_1, \lambda_2 \dots$ for which equation (2) has nonzero solution v. This equation can be transformed in the form

$$Av - \lambda v = 0 \rightarrow (A - \lambda I)v = 0,$$

where I denotes the identity matrix. This is a homogenous system of linear equations which has nontrivial solutions if $\det(A - \lambda I) = 0$.

Definition 3.2 The equation

$$\phi(\lambda) = \det(A - \lambda I) = \det \left(\begin{bmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} - \lambda \end{bmatrix} \right) = 0$$

is called the characteristic equation of matrix A.

Remark 3.1 *The solutions of the characteristic equation are the eigenvalues of A.*

Theorem 3.1 *A matrix $A \in \mathbb{R}^{n\times n}$ has exactly n eigenvalues (considering their multiplicities).*

Remark 3.2 *If v is an eigenvector corresponding to the eigenvalue λ, then tv is also an eigenvector corresponding to the same eigenvector for all nonzero $t \in \mathbb{R}$. So, there are infinitely many eigenvectors belonging to an eingenvalue, that's why we usually give the normalized eigenvector v, for which $|v| = 1$.*

Example 3.1 a) Give the eigenvalues of matrix $A = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix}$.

The matrix $(A - \lambda I)$:

$$(A - \lambda I) = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix} = \begin{bmatrix} 3-\lambda & 1 & 0 \\ 1 & 2-\lambda & 1 \\ 0 & 1 & 3-\lambda \end{bmatrix}.$$

The characteristic equation of A:

$$0 = \det(A - \lambda I) = (3-\lambda)\det\left(\begin{bmatrix} 2-\lambda & 1 \\ 1 & 3-\lambda \end{bmatrix}\right) - 1 \cdot \det\left(\begin{bmatrix} 1 & 1 \\ 0 & 3-\lambda \end{bmatrix}\right)$$
$$= (3-\lambda)\left((2-\lambda)(3-\lambda) - 1\right) - (3-\lambda) = (3-\lambda)(\lambda^2 - 5\lambda + 4)$$

The solutions of the characteristic equation are the eigenvalues of A:

$$\lambda_1 = 3,\ \lambda_2 = 1,\ \lambda_3 = 4.$$

b) Give the normalized eigenvector corresponding to $\lambda_3 = 4$.
Using the definition of the eigenvalue we have that all of the eigenvectors corresponding to λ_3 satisfy the equation $Av = 4v$. This equation in matrix-vector form:

$$\begin{bmatrix} 3 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} 4v_1 \\ 4v_2 \\ 4v_3 \end{bmatrix},$$

from which we get the following system of linear equations:

$$\begin{aligned} 3v_1 + v_2 \qquad &= 4v_1 && \rightarrow && v_2 = v_1 \\ v_1 + 2v_2 + v_3 &= 4v_2 && \rightarrow && v_1 + v_3 = 2v_2 \\ v_2 + 3v_3 &= 4v_3 && \rightarrow && v_2 = v_3 \end{aligned}$$

From these equations we obtain $v_1 = v_2 = v_3$, which means, that the eigenvectors can be written in the general form $v = \begin{bmatrix} t; & t; & t \end{bmatrix}^T$ where $t \in \mathbb{R}$. The normalized eigenvector can be computed by the division $\frac{v}{|v|}$. Here $|v| = \sqrt{t^2 + t^2 + t^2} = \sqrt{3} \cdot t$ if $t > 0$, so

$$\frac{v}{|v|} = \begin{bmatrix} \frac{1}{\sqrt{3}}; & \frac{1}{\sqrt{3}}; & \frac{1}{\sqrt{3}} \end{bmatrix}^T$$

Remark 3.3 *The eigenvalue decomposition of A is a product $A = RDR^T$, where D is a diagonal matrix containing the eigenvalues of A in its diagonal, and R has the corresponding normalized eigenvectors as its columns. The matrices providing the eigenvalue decomposition of the matrix in Example 3.1 are the following:*

$$D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \qquad R = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \\ -\frac{2}{\sqrt{6}} & 0 & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} \end{bmatrix}.$$

Theorem 3.2 (Properties of the eigenvalues) *The following statements are valid:*

1) *The determinant of any matrix is equal to the product of its eigenvalues.*

2) *The trace of any matrix is equal to the sum of its eigenvalues.*

3) *A is nonsingular if and only if all eigenvalues of A are nonzero. The eigenvalues of A^{-1} are the reciprocals of the eigenvalues of A.*

4) *A symmetric matrix Q is positive definite if and only if all the eigenvalues of A are positive.*

5) *A symmetric matrix Q is negative definite if and only if all the eigenvalues of A are negative.*

6) *The eigenvalues of a diagonal or a triangular matrix coincide with the elements in the main diagonal.*

The concept of the eigenvalue appears in the definition of 2-norm of matrices.

Definition 3.3 The 2-norm (or spectral norm) of matrix A is defined by

$$||A||_2 = \sqrt{\text{the largest eigenvalue of } A^T A}.$$

Remark 3.4 *Instead of A^TA we can compute the largest eigenvalue of AA^T, because they are the same. Choose the product of less dimension for the sake of easier calculation.*

Example 3.2 Give $||A||_2$, if $A = \begin{bmatrix} 2 & 3 & -1 \\ 4 & 0 & 2 \end{bmatrix}$.

A^TA is a 3×3 type matrix, while AA^T has 2×2 size, so we calculate the eigenvalues of $AA^T = \begin{bmatrix} 14 & 6 \\ 6 & 20 \end{bmatrix}$:

$$\det(AA^T - \lambda I) = \det\left(\begin{bmatrix} 14-\lambda & 6 \\ 6 & 20-\lambda \end{bmatrix}\right) = (14-\lambda)(20-\lambda) - 36$$
$$= \lambda^2 - 34\lambda + 244.$$

The characteristic equation is $\lambda^2 - 34\lambda + 244 = 0$. We need the larger root of this quadratic equation:

$$\lambda_{\max} = \frac{34 + \sqrt{34^2 - 4 \cdot 244}}{2} = 17 + 3\sqrt{5} \approx 23.7082,$$

and $||A||_2 = \sqrt{\lambda_{\max}} \approx 4.8691$.

3.1 The Power Method: Approximation of the Dominant Eigenvalue

Given a matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$; by its dominant eigenvalue we mean an eigenvalue λ_i such that $|\lambda_i| > |\lambda_j|$ for all $j \neq i$ (if such an eigenvalue exists).

Assume that for a matrix $A \in \mathbb{R}^{n\times n}$ the following holds:

- it has n linearly independent eigenvectors $v_1, v_2, \ldots, v_n$,
- it has a unique dominant eigenvalue, indexed by 1:

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|. \tag{3}$$

We start with an arbitrary initial guess $x^{(0)}$ of the eigenvector corresponding to the dominant eigenvalue, then construct a sequence of vectors $x^{(k)}$ as follows:

$$x^{(1)} = Av^{(0)}, \text{ where } v^{(0)} = \frac{x^{(0)}}{||x^{(0)}||_2},$$

$$x^{(2)} = Av^{(1)}, \text{ where } v^{(1)} = \frac{x^{(1)}}{||x^{(1)}||_2},$$

in general

$$x^{(k+1)} = Av^{(k)}, \text{ where } v^{(k)} = \frac{x^{(k)}}{||x^{(k)}||_2}.$$

The vector $v^{(k)}$ is formed by normalizing the vector $x^{(k)}$, in order to avoid too large components in $x^{(k)}$. In addition form the sequence of real numbers μ_k as $\mu_k = (v^{(k)})^T \cdot x^{(k+1)}$. As a result, we obtain the sequence of vectors $v^{(k)}$ converging to the dominant eigenvector, and a sequence of numbers μ_k, converging to the dominant eigenvalue.

Example 3.3 In case of $A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}$ show the algorithm of the power method.

We need an initial vector $x^{(0)}$. For example choose $x^{(0)} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$. First we normalize this vector, forming

$$v^{(0)} = \frac{x^{(0)}}{||x^{(0)}||_2} = \begin{bmatrix} \frac{3}{5} \\ \frac{4}{5} \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix}.$$

Then

$$x^{(1)} = Av^{(0)} = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix} = \begin{bmatrix} 3.4 \\ 2.2 \end{bmatrix},$$

$$\mu_1 = (v^{(0)})^T \cdot x^{(1)} = \begin{bmatrix} 0.6 & 0.8 \end{bmatrix} \cdot \begin{bmatrix} 3.4 \\ 2.2 \end{bmatrix} = 3.8.$$

The next iteration is:

$$v^{(1)} = \frac{x^{(1)}}{||x^{(1)}||_2} = \begin{bmatrix} 0.8396 \\ 0.5433 \end{bmatrix},$$

$$x^{(2)} = Av^{(1)} = \begin{bmatrix} 3.6052 \\ 1.9261 \end{bmatrix}, \text{ and } \mu_2 = (v^{(1)})^T \cdot x^{(2)} = 4.0732.$$

Following the definitions of $v^{(k)}, x^{(k+1)}$ and μ_{k+1}, we get:

$$v^{(2)} = \begin{bmatrix} 0.8820 \\ 0.4712 \end{bmatrix}, \quad x^{(3)} = \begin{bmatrix} 3.5885 \\ 1.8244 \end{bmatrix}, \quad \mu_3 = 4.0248;$$

$$v^{(3)} = \begin{bmatrix} 0.8914 \\ 0.4532 \end{bmatrix}, \quad x^{(4)} = \begin{bmatrix} 3.5806 \\ 1.7978 \end{bmatrix}, \quad \mu_4 = 4.0066;$$

$$v^{(4)} = \begin{bmatrix} 0.8937 \\ 0.4487 \end{bmatrix}, \quad x^{(5)} = \begin{bmatrix} 3.5785 \\ 1.7911 \end{bmatrix}, \quad \mu_5 = 4.0017;$$

and so on ...

We check the approximation by exact calculation:

$$A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} \quad \to \quad \det(A - \lambda I) = \det\left(\begin{bmatrix} 3-\lambda & 2 \\ 1 & 2-\lambda \end{bmatrix}\right).$$

The characteristic equation is $(3-\lambda)(2-\lambda) - 2 = 0$, that is $\lambda^2 - 5\lambda + 4 = 0$. Its roots are $\lambda_1 = 4$ and $\lambda_2 = 1$.

The eigenvector corresponding to the dominant eigenvalue $\lambda_1 = 4$:

$$\begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = 4 \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

which yields

$$\begin{aligned} 3v_1 + 2v_2 &= 4v_1 & \quad \to \quad & 2v_2 = v_1 \\ v_1 + 2v_2 &= 4v_2 & \quad \to \quad & v_1 = 2v_2 \end{aligned}$$

Normalizing the vector $\begin{bmatrix} 2t \\ t \end{bmatrix}$ we get

$$\begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \end{bmatrix} = \begin{bmatrix} 0.8944 \\ 0.4472 \end{bmatrix}.$$

The exact calculation show that $\lambda = 4$ is the dominant eigenvalue of matrix A, and $v = \begin{bmatrix} \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} \end{bmatrix}^T$ is the corresponding normalized eigenvector.

We obtained the estimations of these values by the power method generating the sequences $v^{(k)}$ and μ_k, for which $v^{(k)} \to v$ and $\mu_k \to \lambda$ as $k \to \infty$.

3.2 Inverse Power Method

If A is a nonsingular matrix and its eigenvalues are $\lambda_1, \ldots, \lambda_n$, then the inverse A^{-1} has the eigenvalues $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$. Suppose that $\frac{1}{\lambda_1}$ is dominant, that is:

$$\frac{1}{|\lambda_1|} > \frac{1}{|\lambda_2|} \geq \frac{1}{|\lambda_3|} \geq \cdots \geq \frac{1}{|\lambda_n|}.$$

We determine the dominant eigenvalue $\frac{1}{\lambda_1}$ of A^{-1} by the power method and its reciprocal λ_1 is the eigenvalue of A with the least absolute value.

Example 3.4 Give the least eigenvalue of $A = \begin{bmatrix} 3 & 2 \\ 1 & 2 \end{bmatrix}$ using the inverse power method.

The inverse of A is $A^{-1} = \begin{bmatrix} 0.5 & -0.5 \\ -0.25 & 0.75 \end{bmatrix}$.

Starting again from the vector $x^{(0)} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$, we have $v^{(0)} = \begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix}$.

$$x^{(1)} = A^{-1}v^{(0)} = \begin{bmatrix} -0.1 \\ 0.45 \end{bmatrix}, \qquad \mu_1 = (v^{(0)})^T \cdot x^{(1)} = 0.3;$$

$$v^{(1)} = \begin{bmatrix} -0.2169 \\ 0.9762 \end{bmatrix}, \qquad x^{(2)} = \begin{bmatrix} -0.5966 \\ 0.7864 \end{bmatrix}, \qquad \mu_2 = 0.8971.$$

Some further elements of the sequence μ_k:

$$\mu_3 = 1.0198,\ \mu_4 = 1.0075,\ \mu_5 = 1.0020,\ \mu_6 = 1.0005,$$

so $\lim\limits_{k\to\infty} \mu_k = 1$, which is the dominant eigenvalue of A^{-1}. The least eigenvalue of A is its reciprocal: $\lambda = \frac{1}{1} = 1$.

3.3 QR method for Finding All Eigenvalues of a Matrix

The method is based on the QR factorization of square matrices.

Theorem 3.3 (QR factorization) *Every matrix $A \in \mathbb{R}^{n\times n}$ can be represented in the form $A = QR$, where Q is an orthogonal matrix (which means that $Q^T = Q^{-1}$) and R is an upper triangular matrix.*

The factors Q and R are usually determined by the Gram-Schmidt orthogonalization method.

Definition 3.4 (Similarity of matrices) Given $A, B \in \mathbb{R}^{n\times n}$ matrices are said to be similar, if there exists a nonsingular matrix $T \in \mathbb{R}^{n\times n}$, for which $B = T^{-1}AT$.

It can be easily proved that the eigenvalues of similar matrices coincide.

The essence of the QR method can be summarized as follows. Starting from the original matrix A, we construct a sequence of matrices A_k using the QR factorization technique. The elements of the sequence are similar having the same eigenvalues. The sequence A_k tends to a matrix A^*, which is triangular. The eigenvalues of A^* can be easily determined because the eigenvalues of a triangular matrix coincide with the elements of its main diagonal.

Construction of the sequence A_k:

- $A_1 = A$, where A is the matrix whose eigenvalues to be determined.
- We make the QR factorization process for A_1, i. e. $A_1 = Q_1 R_1$, where Q_1 is an orthogonal and R_1 is an upper triangular matrix.
- We form A_2 by multiplying the factors in reverse order: $A_2 = R_1 Q_1$. The obtained product A_2 is similar to $A_1 = A$, because $R_1 = Q_1^{-1} A_1$ and $A_2 = Q_1^{-1} A_1 Q_1$ as required in the definition of similarity.
- The general step of the algorithm: determine the QR factors of A_k: $A_k = Q_k R_k$, and form the next element of the sequence as follows: $A_{k+1} = R_k Q_k$. This rule provides a sequence of similar matrices, so the eigenvalues of the elements ar the same.

Example 3.5 Consider the matrix $A = \begin{bmatrix} 6 & 1 & 1 \\ -8 & 0 & 4 \\ 12 & 6 & 2 \end{bmatrix}$. Give an approximation of the eigenvalues of A using the QR method.

Let $A_1 = A$, and make the QR factorization for A_1. Using Octave, command qr() yields:

$$Q_1 = \begin{bmatrix} -0.38 & 0.26 & -0.88 \\ 0.51 & -0.73 & -0.44 \\ -0.77 & -0.62 & 0.15 \end{bmatrix}, \qquad R_1 = \begin{bmatrix} -15.62 & -4.99 & 0.12 \\ 0 & -3.47 & -3.93 \\ 0 & 0 & -2.36 \end{bmatrix}.$$

We form the product

$$A_2 = R_1 Q_1 = \begin{bmatrix} 3.34 & -0.53 & 16.05 \\ 1.24 & 5.00 & 0.96 \\ 1.81 & 1.47 & -0.35 \end{bmatrix},$$

and repeating the process: $A_2 = Q_2 R_2,\ A_3 = R_2 Q_2, \ldots$, and so on.

After 10 iterations we get the following matrix:

$$A_{10} = \begin{bmatrix} 8.0086 & 0.8196 & -12.4086 \\ -0.0038 & 3.9992 & 7.1526 \\ 0.0086 & 0.0017 & -4.0078 \end{bmatrix}.$$

It can be considered as an upper triangular matrix, and the elements in its main diagonal are good estimates of the eigenvalues of the original matrix A: $\lambda_1 = 8,\ \lambda_2 = 4,\ \lambda_3 = -4$.

4 Interpolation

Suppose that we have a table of measured or computed values of a function $f(x)$: $y_0 = f(x_0), y_1 = f(x_1), \ldots, y_n = f(x_n)$, for a set of points $x_0 < x_1 < \cdots < x_n$ in some interval $[a; b]$:

x_0	x_1	$\cdots$	x_n
y_0	y_1	$\cdots$	y_n

Given such a table, our goal is to approximate $f(x)$ with a function, say a polynomial $p(x)$, such that $p(x_i) = f(x_i), i = 0, 1, \ldots, n$. Then for any $x \in [a; b]$ we can approximate the value $f(x)$ with $p(x)$: $f(x) \approx p(x)$. If $x_0 < x < x_n$, the approximation $p(x)$ is called an interpolated value, otherwise it is an extrapolated value.

4.1 Polynomial Interpolation

Polynomials represent a natural choice for approximation of more complicated functions, since their simple structure facilitates easy manipulation.
A polynomial $p(x)$ in the power form is given as follows:

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

If $a_n \neq 0$, then $p(x)$ has degree n.

The simplest case of polynomial interpolation, if we have two given points and need the equation of the line which contains the points. For example, let $A = (3; -4)$ and $B = (7; 2)$ the two points, and the equation of the polynomial interpolating them $p(x) = a_0 + a_1 x$. The following equations must be satisfied: $p(3) = -4$ and $p(7) = 2$. We have a linear system of equations:

$$\begin{aligned} a_0 + 3a_1 &= -4 \\ a_0 + 7a_1 &= 2 \end{aligned}$$

The solution is $a_0 = -8.5$ and $a_1 = 1.5$, so the equation of the interpolating polynomial is $p(x) = -8.5 + 1.5x$.

The next theorem shows that independently the number of the points, the interpolating polynomial always exists.

Theorem 4.1 *Given* $(x_0, y_0), \ldots, (x_n, y_n)$, *where* $x_i \neq x_j$ *if* $i \neq j$, *there exists a unique interpolating polynomial* $p(x)$ *of degree no greater than* n, *such that* $p(x_i) = y_i,\ i = 0, \ldots, n$.

Proof We give a constructive proof here providing the interpolating polynomial by the Lagrange method.

First define the n^{th} degree Lagrange polynomials for $j = 0, 1, \ldots, n$ using the points $x_0, x_1, \ldots, x_n$:

$$\begin{aligned} l_j(x) &= \frac{(x - x_0)(x - x_1)\cdots(x - x_{j-1})(x - x_{j+1})\cdots(x - x_n)}{(x_j - x_0)(x_j - x_1)\cdots(x_j - x_{j-1})(x_j - x_{j+1})\cdots(x_j - x_n)} \\ &= \prod_{i=0, i\neq j}^{n} \frac{(x - x_i)}{(x_j - x_i)}. \end{aligned}$$

The Lagrange polynomial l_j has the following property:

$$l_j(x_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Hence the following polynomial

$$p(x) = y_0 l_0(x) + y_1 l_1(x) + \cdots + y_n l_n(x) = \sum_{j=0}^{n} y_j \prod_{i=0, i\neq j}^{n} \frac{(x - x_i)}{(x_j - x_i)}$$

has the desired interpolating property $p(x_i) = y_i$, for $i = 0, \ldots, n$.

For the uniqueness, assume that there are two polynomials of degree $\leq n$, $p(x)$ and $q(x)$, such that $p(x_i) = q(x_i) = y_i, i = 0, \ldots, n$. Then $r(x) = p(x) - q(x)$ is a polynomial of degree $\leq n$, which has $n + 1$ distinct roots $x_0, x_1, \ldots, x_n$. We obtain a contradiction with the fundamental theorem of algebra.

Example 4.1 Use the Lagrange method to find the second-degree polynomial $p(x)$ interpolating the data

x	-1	1	2
y	12	4	9

Solution: The Lagrange polynomials:

$$l_0(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{(x-1)(x-2)}{(-1-1)(-1-2)} = \frac{1}{6}(x^2-3x+2),$$
$$l_1(x) = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{(x-(-1))(x-2)}{(1-(-1))(1-2)} = -\frac{1}{2}(x^2-x-2),$$
$$l_2(x) = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{(x-(-1))(x-1)}{(2-(-1))(2-1)} = \frac{1}{3}(x^2-1).$$

The interpolating polynomial:

$$\begin{aligned} p(x) &= y_0 l_0(x) + y_1 l_1(x) + y_2 l_2(x) \\ &= 12 \cdot \frac{1}{6}(x^2-3x+2) + 4 \cdot (-\frac{1}{2})(x^2-x-2) + 9 \cdot \frac{1}{3}(x^2-1) \\ &= 3x^2 - 4x + 5. \end{aligned}$$

Remark 4.1 *We could solve the problem similarly to the introductory linear problem using the method of undetermined coefficients. In this case the form of the searched polynomial is* $p(x) = a_0 + a_1 x + a_2 x^2$, *and we have to solve the linear system of equations:*

$$\begin{aligned} a_0 - a_1 + a_2 &= 12 \\ a_0 + a_1 + a_2 &= 4 \\ a_0 + 2a_1 + 4a_2 &= 9 \end{aligned}$$

The solution is: $a_0 = 5, a_1 = -4, a_2 = 3.$

Though the method of undetermined coefficients seems less complicated than the Lagrange method, the latter has a very useful property, namely, it is suitable for interpolating $f(x)$ at a given $x_0 < x < x_n$ point without calculating the coefficients of the interpolating polynomial.

For example, let see again the table of Example 4.1, and give an approximation of $f(1.2)$ using second-degree polynomial interpolation.

We use the Lagrange method:

$$\begin{aligned} f(1.2) \approx p(1.2) &= y_0 l_0(1.2) + y_1 l_1(1.2) + y_2 l_2(1.2) \\ &= 12 \cdot \frac{(1.2-1)(1.2-2)}{(-1-1)(-1-2)} + 4 \cdot \frac{(1.2-(-1))(1.2-2)}{(1-(-1))(1-2)} \\ &\quad + 9 \cdot \frac{(1.2-(-1))(1.2-1)}{(2-(-1))(2-1)} \\ &= 4.52 \end{aligned}$$

Check the result by substituting $x = 1.2$ in the known interpolating polynomial $p(x) = 3x^2 - 4x + 5$:

$$p(1.2) = 3 \cdot 1.2^2 - 4 \cdot 1.2 + 5 = 4.52.$$

Example 4.2 Find the cubic polynomial $p(x)$ that interpolates the data:

x	1	2	3	5
y	4	0	0	36

Solution: $p(x) = x^3 - 4x^2 + x + 6$.

4.2 Linear Least Squares Models

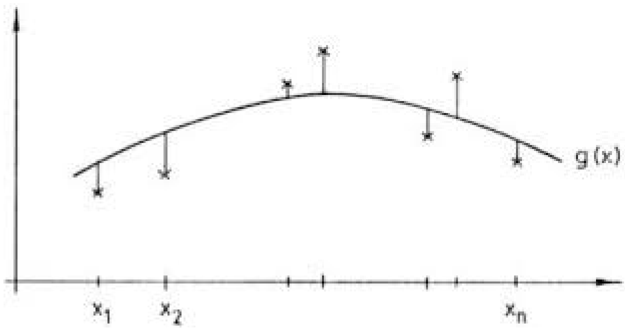
Suppose that we have a table of the values of a function $f(x)$, $y_1 = f(x_1), \ldots$, $y_n = f(x_n)$, for a set of points $x_1 < x_2 < \cdots < x_n$ in some interval $[a; b]$. Our goal is to determine the unknown parameters $a_0, a_1, \ldots, a_k, (k < n)$ of the function

$$g(x) = a_0 g_0(x) + a_1 g_1(x) + \ldots a_k g_k(x) \tag{4}$$

such that

$$\sum_{i=1}^{n} (g(x_i) - y_i)^2$$

should be minimal. This problem is called the linear least squares problem. The terminology linear only refers to the model's dependence on its a_i parameters, the functions g_j-s can be highly nonlinear. The next figure shows that in this problem we do not make curve-fitting. We are looking for a function $g(x)$ which approximates the points in the sense that the sum of squared residuals (the vertical distances on the figure) are to be minimized.



A very frequently arising problem in statistics is the *linear regression problem* which can be formulated as follows: given n points $(x_1; y_1), (x_2; y_2), \ldots (x_n; y_n)$

on the plane find the line $l(x) = a_0 + a_1 x$ (defined by the coefficients a_0 and a_1) that minimizes the sum

$$\sum_{i=1}^{n} (a_0 + a_1 x_i - y_i)^2.$$

Example 4.3 Solve the linear regression problem for the given points

x	0	1	2
y	2	2	3

Solution: The line $l(x) = a_0 + a_1 x$ takes the following values at points 0,1,2:

$$l(0) = a_0, \qquad l(1) = a_0 + a_1, \qquad l(2) = a_0 + 2a_1$$

We have to minimize the sum

$$F(a_0, a_1) = (a_0 - 2)^2 + ((a_0 + a_1) - 2)^2 + ((a_0 + 2a_1) - 3)^2$$

Compute the partial derivatives of function F:

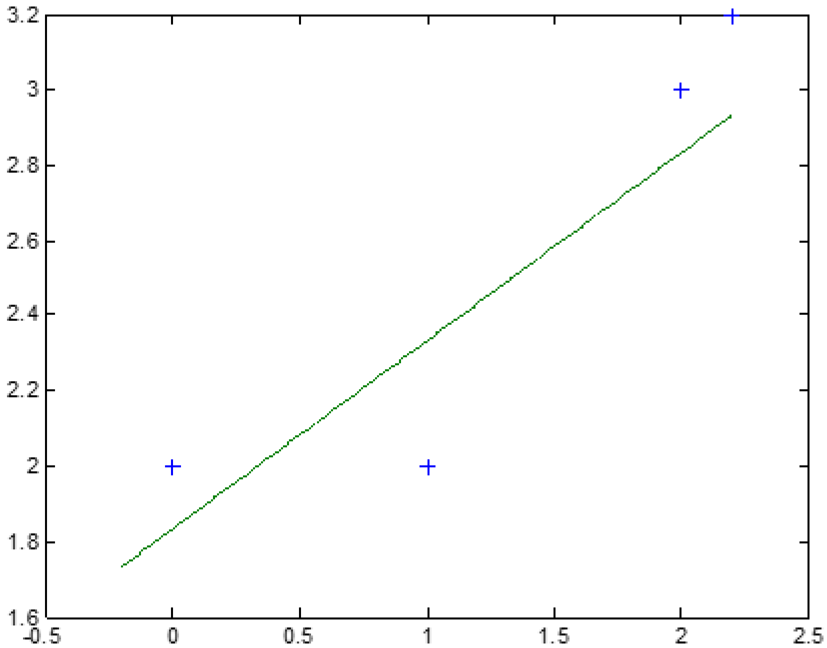
$$F'_{a_0} = 2(a_0 - 2) + 2(a_0 + a_1 - 2) + 2(a_0 + 2a_1 - 3) = 6a_0 + 6a_1 - 14$$

$$F'_{a_1} = 2(a_0 + a_1 - 2) + 2(a_0 + 2a_1 - 3) \cdot 2 = 6a_0 + 10a_1 - 16$$

For finding the possible extrema we have to solve the linear system:

$$\begin{aligned} 6a_0 + 6a_1 - 14 &= 0 \\ 6a_0 + 10a_1 - 16 &= 0 \end{aligned}$$

The solution is $a_0 = \frac{11}{6}$, $a_1 = \frac{1}{2}$. The study of the second derivatives shows that F has a minimum at this point, so the equation of the desired line is $l(x) = \frac{11}{6} + \frac{1}{2}x$, see the next figure:



In general it is not necessary to solve the problem by classical calculus tools. It can be proved, that the coefficients of the function $g(x)$ defined in (4) can be determined by solving a linear system of equations (often referred as the *normal equation*).

Denote by $a = (a_0, a_1, \ldots, a_k)^T$ the column vector of the unknown coefficients, and by $y = (y_1, y_2, \ldots, y_n)^T$ the column vector of the given values of function f. Form a matrix A of type $n \times (k+1)$ whose jth column contains the values $g_j(x_1), g_j(x_2), \ldots g_j(x_n)$, $j = 0, 1, \ldots, k$, where g_j is the function standing with the coefficient a_j in equation (4). Using the introduced notations we define the normal equation as follows:

$$A^T A a = A^T y.$$

Theorem 4.2 *If a^* solves the system*

$$A^T A a = A^T y,$$

then a^* is the solution of the linear least squares problem.

Using the normal equation we solve the problem given in Example 3. We have three points x_1, x_2, x_3 and two unknown parameters a_0, a_1. So the matrix A

has three rows and two columns, as follows:

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$$

The first column contains only ones, because the coefficients of a_0 equal 1 in the equation of the line, and the second column contains the number x_i-s equalling the coefficients of a_1. We form the transposed matrix of A and make the necessary operations:

$$A^T = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \quad A^T A = \begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \quad A^T y = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

The normal equation we have to solve:

$$\begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

which is

$$\begin{aligned} 3a_0 + 3a_1 &= 7 \\ 3a_0 + 5a_1 &= 8 \end{aligned}$$

Its solution is $a_0 = \frac{11}{6}$, $a_1 = \frac{1}{2}$, so the equation of the desired line is $l(x) = \frac{11}{6} + \frac{1}{2}x$.

Example 4.4 Solve the linear least squares problem for the function $g(x) = a_0 + a_1\sqrt{x}$ using the table of points

x	0	1	4	9
y	2	2	3	5

Solution: The matrix $A = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$. The elements in the second column are the square roots of numbers x_i. The normal equation using this matrix is:

$$\begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 12 \\ 23 \end{bmatrix},$$

hence we have to solve

$$\begin{aligned} 4a_0 + 6a_1 &= 12 \\ 6a_0 + 14a_1 &= 23 \end{aligned}$$

Its solution is $a_0 = 1.5,\ a_1 = 1$, so the equation of the function which solves the least square problem is $g(x) = 1.5 + \sqrt{x}$.

Example 4.5 Consider the points

x	-2	-1	0	1	2
y	2	2	4	3	1

Solve the least square problem for the polynomials of degree n, $(n = 1, 2, 3)$. What is happened if $n = 4$?

Solution:
$n = 1:$

$$A = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} 12 \\ -1 \end{bmatrix}, \quad p_1(x) = 2.4 - 0.1x$$

$n = 2:$

$$A = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}, \quad \begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 12 \\ -1 \\ 17 \end{bmatrix}, \quad p_2(x) = 3.4 - 0.1x - 0.5x^2$$

$n = 3:$

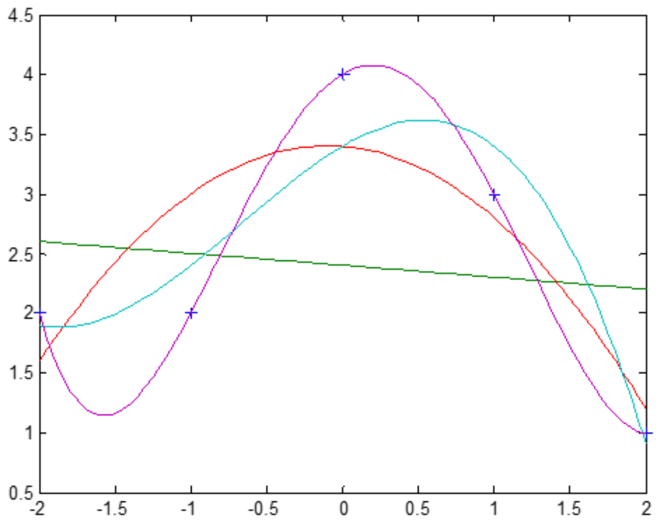
$$A = \begin{bmatrix} 1 & -2 & 4 & -8 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{bmatrix}, \quad \begin{bmatrix} 5 & 0 & 10 & 0 \\ 0 & 10 & 0 & 34 \\ 10 & 0 & 34 & 0 \\ 0 & 34 & 0 & 130 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 12 \\ -1 \\ 17 \\ -7 \end{bmatrix},$$

$p_3(x) = 3.4 + 0.75x - 0.5x^2 - 0.25x^3.$

In case of $n = 4$, the problem is an interpolating problem, because there is a unique polynomial of degree 4, for which the given points fit. Its equation:

$$p_4(x) = 4 + 0.75x - 1.7917x^2 - 0.25x^3 + 0.2917x^4.$$

The next figure shows the graphs of the curves which solve the different least squares problems. The lilac graph represents the interpolating polynomial, it contains all the given points.



4.3 Spline Interpolation

In the previous chapter nth-order polynomials were used to interpolate between $n + 1$ points. For example, for ten points, we can derive a perfect ninth-order polynomial. However, there are cases where these higher degree functions can lead to erroneous result. An alternative approach is to apply lower-order polynomials to subsets of data points. Such connecting polynomials are called *spline functions*.

4.3.1 Linear Splines

The simplest version of spline interpolation where the connection between two points is given by a straight line. Linear splines or first-order splines for a group of ordered data points can be defined as a set of linear functions as follows:

$$\begin{aligned}
f(x) &= f(x_0) + m_0(x - x_0) \quad x_0 \le x \le x_1, \\
f(x) &= f(x_1) + m_1(x - x_1) \quad x_1 \le x \le x_2, \\
&\vdots \\
f(x) &= f(x_{n-1}) + m_{n-1}(x - x_{n-1}) \quad x_{n-1} \le x \le x_n,
\end{aligned}$$

where m_i is the slope of the straight line connecting the points:

$$m_i = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i}.$$

Example 4.6 Fit the data in the following table with first-order splines:

x	3	4.5	7	9
$f(x)$	2.5	1	2.5	0.5

Evaluate the function at $x = 5$.

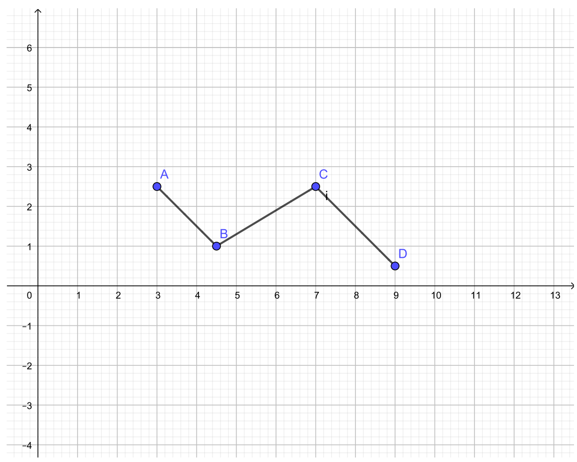
Solution: We compute the slopes of the lines:

$$m_0 = \frac{1 - 2.5}{4.5 - 3} = -1, \quad m_1 = \frac{2.5 - 1}{7 - 4.5} = 0.6, \quad m_2 = \frac{0.5 - 2.5}{9 - 7} = -1$$

and write the definition of the linear spline:

$$f(x) = \begin{cases} 2.5 - (x - 3) = 5.5 - x & \text{if } 3 \le x \le 4.5 \\ 1 + 0.6(x - 4.5) = 0.6x - 1.7 & \text{if } 4.5 \le x \le 7 \\ 2.5 - (x - 7) = 9.5 - x & \text{if } 7 \le x \le 9 \end{cases}$$

The value at $x = 5$ is $0.6 \cdot 5 - 1.7 = 1.3$. The graph of the linear spline:



The primary disadvantage of first-order splines is that they are not smooth. At data points, where two pieces of the spline meet (called a *knot*), the slope changes abruptly. In formal terms, the first derivative of the function is discontinuous at these points. This deficiency is overcome by using higher-order polynomial splines that ensure smoothness at the knots by equating derivatives at these points.

4.3.2 Cubic Splines

Third-order polynomials or cubic splines that ensure continuous first and second derivatives are most frequently used in practice.

The objective in cubic splines to derive a third-order polynomial for each interval between knots, as in

$$f_i(x) = A_i + B_i x + C_i x^2 + D_i x^3.$$

For $n + 1$ data points there are n intervals and $4n$ unknown constants to evaluate. In order to determine the coefficients of $f_i(x)$ $4n$ conditions are required. These are:
The function values must be equal at the interior knots ($2n - 2$ conditions).
The first and last functions must pass through the end points (2 conditions).
The first derivatives at the interior knots must be equal ($n - 1$ conditions).
The second derivatives at the interior knots must be equal ($n-1$ conditions).
The second derivatives at the end knots are zero (2 conditions).

The specification is given in the last condition leads to what is termed a "natural" spline. It is given, because the drafting spline naturally behaves in this fashion. Spline originally means a thin, flexible strip to draw smooth curves through a set of points.

Example 4.7 Fit cubic spline to the data given in the table

x	-1	0	1
f(x)	1	0	1

We have three points and two intervals between them. Two third-order polynomials are defined on these intervals:

$$f(x) = \begin{cases} f_1(x) & = A_1 + B_1 x + C_1 x^2 + D_1 x^3, \text{ if } x \in [-1; 0], \\ f_2(x) & = A_2 + B_2 x + C_2 x^2 + D_2 x^3, \text{ if } x \in [0; 1]. \end{cases}$$

We have 8 unknows, so 8 conditions are required:

1) $x = -1$: $f_1(-1) = 1$ implies $A_1 - B_1 + C_1 - D_1 = 1$.

2) $x = 0$: $f_1(0) = 0$ implies $A_1 = 0$.

3) $x = 0$: $f_2(0) = 0$ implies $A_2 = 0$.

4) $x = 1$: $f_2(1) = 1$ implies $A_2 + B_2 + C_2 + D_2 = 1$.

5) The first derivatives must be equal at the interior knot. Since $f_i'(x) = B_i + 2C_i x + 3D_i x^2$ $(i = 1, 2)$, equation $f_1'(0) = f_2'(0)$ implies $B_1 = B_2$.

6) The second derivatives must be equal at the interior knot. Since $f_i''(x) = 2C_i + 6D_i x$ $(i = 1, 2)$, equation $f_1''(0) = f_2''(0)$ yields $2C_1 = 2C_2$, that is $C_1 = C_2$.

7)-8) The second derivatives at the endpoints are zero:
$f_1''(-1) = 0$ implies $2C_1 - 6D_1 = 0$ and $f_2''(1) = 0$ implies $2C_2 + 6D_2 = 0$.

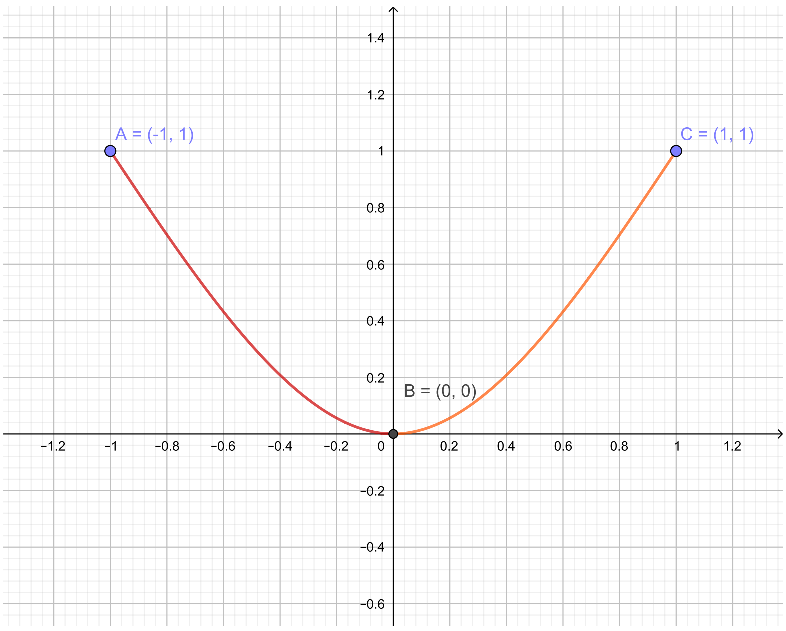
The solution of the system containing linear equations formulated in 1)-8):

$$A_1 = A_2 = 0;\ B_1 = B_2 = 0;\ C_1 = C_2 = \frac{3}{2};\ D_1 = \frac{1}{2};\ D_2 = -\frac{1}{2}.$$

Therefore the equation of the cubic spline:

$$f(x) = \begin{cases} f_1(x) = \frac{3}{2}x^2 + \frac{1}{2}x^3, & \text{if } x \in [-1; 0], \\ f_2(x) = \frac{3}{2}x^2 - \frac{1}{2}x^3, & \text{if } x \in [0; 1], \end{cases}$$

and its graph is:



5 Solving Nonlinear Equations and Nonlinear Systems

Let f be a given real function. We would like to solve the equation

$$f(x) = 0,$$

or, in other words to find a root $x^* \in \mathbb{R}$ of the function $f(x)$ for which $f(x^*) = 0$.

There are only a few types of equations that can be solved by simple analytical methods. We have formulas for solving linear, quadratic, cubic equations and polynomials of degree 4. Abel, a Norwegian mathematician had shown, that no such formulas can be developed for general polynomials of degree higher than four. So, direct solution methods are often not available, instead of them we have to use numerical (iterative) techniques.

An iterative method starts with an initial value x_0 and proceeds by computing improved estimates $x_1, x_2, \dots, x_k$ converging to the exact solution x^*.

5.1 Bisection Method (Half-Interval Method)

The bisection method is one of the bracketing methods, which are based on the observation that if the function $f : [a; b] \to \mathbb{R}$ is continuous and $f(a) \cdot f(b) \leq 0$, then there exists $x^* \in [a; b]$ such that $f(x^*) = 0$. It is the consequence of the intermediate value theorem, which says that a continuous function $f : [a; b] \to \mathbb{R}$ takes all values between $f(a)$ and $f(b)$.

In case of $f(a) \cdot f(b) = 0$ one of the points a and b is the root of the function, otherwise, $f(a) \cdot f(b) < 0$ means that the function has opposite sign values at the endpoints and the bisection method can be started.

The method proceeds as follows:

- We have an interval $[a; b]$ such that $f(a) \cdot f(b) < 0$; set $a_0 = a$ and $b_0 = b$.
- Let $x_0 = \dfrac{a_0 + b_0}{2}$ be the midpoint of the interval $[a_0; b_0]$.
- Compute the value $f(x_0)$ and if $f(a_0) \cdot f(x_0) < 0$, then the root is between a_0 and x_0, otherwise it is between x_0 and b_0. In the first case we go on with $a_1 = a_0$ and $b_1 = x_0$, otherwise we have $a_1 = x_0$ and $b_1 = b_0$.
- We continue the search with the interval $[a_1; b_1]$ which length is the half of the starting interval $[a_0; b_0]$. The next iteration is $x_1 = \frac{a_1+b_1}{2}$,

we compute the value $f(x_1)$ and depending on the signs of $f(a_1) \cdot f(x_1)$ and $f(x_1) \cdot f(b_1)$ we continue the search on the interval $[a_1; x_1]$ or $[x_1; b_1]$ and so on.

We obtain a sequence of intervals

$$[a; b] = [a_0; b_0] \supset [a_1; b_1] \supset \cdots \supset [a_k; b_k] \supset \cdots$$

where for the centers $x_0, x_1, \ldots, x_k, \ldots$ we have $\lim\limits_{k\to\infty} x_k = x^*$.

To prove this result observe that at iteration n the root x^* is in the interval $[a_n; b_n]$. Since x_n is the mid-point of the interval, we have

$$\mid x^* - x_n \mid \leq \frac{1}{2}(b_n - a_n) = \frac{1}{4}(b_{n-1} - a_{n-1}) = \cdots = \frac{1}{2^{n+1}}(b_0 - a_0). \quad (5)$$

Using the fact that $\lim\limits_{k\to\infty} \frac{1}{2^{n+1}} = 0$, the convergence of the bisection method follows.

Equation (5) provides us an easy way to estimate the error of iteration n, or it is suitable to determine the number of iterations required to achieve a given precision ϵ. If we need an approximate solution x_k such that $\mid x^* - x_k \mid \leq \epsilon$, we have to compute k steps, where k satisfies

$$\frac{1}{2^{k+1}}(b_0 - a_0) < \epsilon.$$

That is

$$\frac{b_0 - a_0}{\epsilon} < 2^{k+1}$$

$$\ln \frac{b_0 - a_0}{\epsilon} < \ln 2^{k+1} = (k+1) \cdot \ln 2$$

$$k > \frac{\ln\left(\frac{b_0 - a_0}{\epsilon}\right)}{\ln 2} - 1.$$

For example, if $[a_0; b_0] = [0; 1]$, and $\epsilon = 10^{-3}$, then

$$n > \frac{3 \ln 10}{\ln 2} - 1 \approx 8.9658,$$

meaning that 9 steps is required to achieve the precision ϵ.

Example 5.1 Use the bisection method to approximate a root $x^* \in [1; 3]$ of

$$f(x) = 162x^4 - 405x^3 + 81x - 728.$$

a) Prove that there exists a root in $[1; 3]$.

b) Determine the iterations $x_0, x_1, \ldots, x_5$. Give the error of x_5.

c) If we need a solution with precision $\epsilon = 10^{-6}$, how many steps should we compute?

Solution: a) Since f is continuous, $f(1) < 0$ and $f(3) > 0$, there is at least one root between 1 and 3.

b) We make a table containing the endpoints of the kth interval and the terms of the sequence x_k. The sign indicated after a_k and b_k means the the sign of the evaluated function.

k	$a_k(-)$	$b_k(+)$	$x_k = \frac{a_k+b_k}{2}$	the sign of $f(x_k)$
0	1	3	2	-
1	2	3	2.5	-
2	2.5	3	2.75	$+$
3	2.5	2.75	2.625	-
4	2.625	2.75	2.6875	$+$
5	2.625	2.6875	2.65625	

The error of x_5 can be calculated by the formula:

$$\mid x^* - x_n \mid \leq \frac{1}{2^{n+1}}(b_0 - a_0).$$

So we have

$$\mid x^* - x_5 \mid \leq \frac{1}{2^6}(3-1) = \frac{1}{2^5} \approx 0.03125.$$

Remark: the true solution of the equation $f(x) = 0$ is $\frac{8}{3}$.

c) To answer the question we use the formula

$$k > \frac{\ln\left(\frac{b_0-a_0}{\epsilon}\right)}{\ln 2} - 1.$$

We have

$$k > \frac{\ln\left(\frac{3-1}{10^{-6}}\right)}{\ln 2} - 1 \approx 19.93.$$

thus 20 iteration is needed to achieve the required precision.

5.2 Newton's Method

In the case of Newton's method we use a linear approximation of the function, using the tangent lines to $f(x)$ to obtain the current estimate x_k of the root.

Consider the first-order Taylor's series approximation of $f(x)$ about x_k, which gives the equation of the tangent line belonging to the point $(x_k; f(x_k))$:

$$y = f(x_k) + f'(x_k)(x - x_k).$$

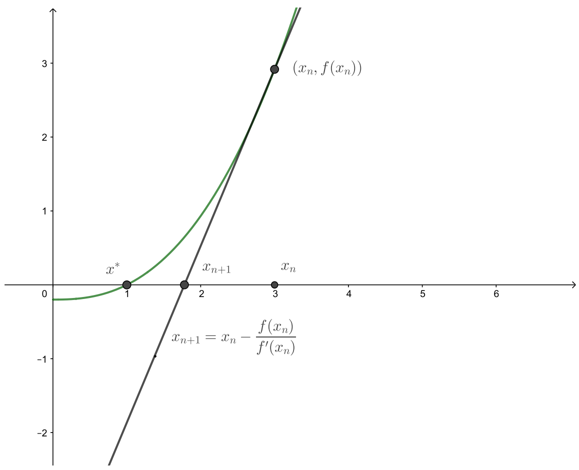
Instead of solving $f(x) = 0$ we solve the linear equation $y = 0$, finding the intersection of the tangent line with the x-axis, because their common point will be the next iteration x_{k+1}:

$$\begin{aligned} f(x_k) + f'(x_k)(x - x_k) &= 0 \\ f'(x_k)(x - x_k) &= -f(x_k) \\ x - x_k &= -\frac{f(x_k)}{f'(x_k)} \\ x &= x_k - \frac{f(x_k)}{f'(x_k)}, \end{aligned}$$

so the next iteration of Newton's method can be written as

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

An illustration of Newton's method:



Theorem 5.1 (Convergence of Newton's Method) *Assume that function f satisfies the following conditions:*

1) f is continuously differentiable twice on the interval $[a;b]$,

2) f has a root on the interval $[a;b]$, for example $f(a) \cdot f(b) \leq 0$ holds,

3) $f'(x) \neq 0$ and $f''(x) \neq 0$ for $x \in [a;b]$,

4) $x_0 \in [a;b]$ is a point for which the signs of $f(x_0)$ and $f''(x_0)$ are the same.

Then the sequence

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

converges to the unique solution x^* in $[a;b]$ of the equation $f(x) = 0$, and we have the following formula for estimating the error of x_k:

$$| x_k - x^* | \leq \frac{M}{2m}(x_k - x_{k-1})^2 \tag{6}$$

where M is an upper bound of $|f''(x)|$ and m is a positive lower bound of $|f'(x)|$ on the interval $[a;b]$.

Remark 5.1 *Inequation (6) shows that Newton's method quadratically convergent.*

Example 5.2 Using Newton's method give an approximation for solving the equation $\sin x = \ln x$ on the interval $[1;3]$.

Solution: We search the root of the function $f(x) = \sin x - \ln x$. Since f is continuous (moreover continuously differentiable twice) on the interval $[1;3]$ and $f(1) > 0$, $f(3) < 0$, there exists a root between 1 and 3. The derivatives

$$f'(x) = \cos x - \frac{1}{x}, \qquad f''(x) = -\sin x + \frac{1}{x^2}$$

do not vanish in the interval $[1;3]$. We checked the items 1)-3) of Theorem 5.1 and we need an initial point to start the iteration. Using condition 4), we choose $x_0 = 1$, because $f(1)$ and $f''(1)$ are both positive numbers. The first iteration:

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 1 - \frac{\sin 1 - \ln 1}{\cos 1 - \frac{1}{1}} \approx 2.830487722$$

We obtain x_2 by replacing $x_1 = 2.830487722$ in the recursive formula:

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \approx 2.267902211,$$

And so on, the result of further iterations:

$x_3 = 2.219744453$; $x_4 = 2.219107263$; $x_5 = 2.219107149$; $x_6 = 2.219107149$.

The last two iterations are the same, showing, that the result is exact for 9 decimal digits.

Remark 5.2 *The above results were obtained by using a pocket calculator in the following simple way (obviously better estimation can be achieved by computer). First give the initial element 1, then press the button =. With this operation we saved the value of x_0 in the Ans memory of the calculator. Then formulate the iteration as follows:*

$$Ans - \frac{\sin Ans - \ln Ans}{\cos Ans - \frac{1}{Ans}}$$

Be careful, because we have to use radians instead of degrees in case of the trigonometric functions. Now we have to press = to calculate x_1, and x_1 appears on the display. Pressing again the button = the calculation is repeated with x_1 providing the next iteration x_2, and so on, after some more clicks on button = and the result does not change any more.

Example 5.3 Use Newton's method for finding the pth root of a positive number a. For example give an approximation for $\sqrt[3]{100}$.

Solution: First we solve the special problem $\sqrt[3]{100}$. We have to find a number x for which $x^3 = 100$, in other words we have to determine the root of the function $f(x) = x^3 - 100$. We know, that the root is between 4 and 5, because $4^3 = 64 < 100 < 125 = 5^3$. The derivatives are $f'(x) = 3x^2$ and $f''(x) = 6x$. Since $f(5)$ and $f''(5)$ have the same signs, we start the iteration from the point $x_0 = 5$.

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 5 - \frac{5^3 - 100}{3 \cdot 5^2} = \frac{14}{3} = 4.666666667$$

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} \approx 4.641723356$$

$$x_3 \approx 4.641588838 \qquad x_4 \approx 4.641588834 \qquad x_5 \approx 4.641588834$$

The last two iterations coincide showing that we achieved the best possible results by using our calculator.

The general formula for approximating $\sqrt[p]{a}$:

$$x_k = \frac{1}{p}\left((p-1)x_{k-1} + \frac{a}{x_{k-1}^{p-1}}\right).$$

For example in case of square root ($p = 2$), the formula is simplified to

$$x_k = \frac{1}{2}\left(x_{k-1} + \frac{a}{x_{k-1}}\right).$$

Obviously we have to guess a good initial estimate for the root. The least integer x for which $x^p > a$ is usually a suitable choice.

5.3 Fixed Point Method

The fixed point method is often referred as simple one-point iteration or method of successive substitutions.

We consider the equation $f(x) = 0$, where $f : [a; b] \to \mathbb{R}$. First we have to transform the equation in the form $x = g(x)$ such that the two equations are equivalent. It is always possible as the following calculation shows.
Let $g(x) = x - \alpha(x)f(x)$, where $\alpha(x) \neq 0$ if $x \in (a; b)$. Then

$$x = g(x) \Leftrightarrow \alpha(x)f(x) = 0 \Leftrightarrow f(x) = 0.$$

The fixed point iteration is defined by the sequence

$$x_{k+1} = g(x_k),\ k = 0, 1, \dots,$$

where $x_0 \in [a; b]$ is an arbitrarily chosen number.

Definition 5.1 A fixed point of a function g is a real number x^* such that

$$g(x^*) = x^*.$$

Clearly, if x^* is a fixed point of g, then $f(x^*) = 0$, and thus x^* is a root of f.
Denote by $C[a; b]$ the set of the continuous functions on the interval $[a; b]$.

Theorem 5.2 *Suppose that the function $g \in C[a; b]$ and $a \leq g(x) \leq b$ for all $x \in [a; b]$, then function $g(x)$ has a fixed point on the interval $[a; b]$.*

Definition 5.2 Let $g \in C[a; b]$. Function g is called a contraction mapping on interval $[a; b]$, if there exists $0 \leq q < 1$ such that

$$|g(x) - g(y)| \leq q|x - y| \qquad \text{for all } x, y \in [a; b].$$

Theorem 5.3 (Banach's fixed point theorem) *If g is a contraction mapping on $[a; b]$ and $a \leq g(x) \leq b$ for all $x \in [a; b]$, then g has exactly one fixed point, which is the limit of the fixed point iteration $x_{k+1} = g(x_k),\ k = 0, 1, \dots,$ where $x_0 \in [a; b]$.*

Theorem 5.4 *Function $g \in C[a;b]$ is a contraction mapping on $[a;b]$, if*

$$\max_{x \in [a;b]} \mid g'(x) \mid = q < 1.$$

Steps of the fixed point method:

1) Equation $f(x) = 0$ is to be transformed in the form $x = g(x)$, where g has the following properties:

   - $g \in C[a;b]$
   - $a \leq g(x) \leq b$
   - $\max\limits_{x \in [a;b]} \mid g'(x) \mid = q < 1.$

2) Starting from any number $x_0 \in [a;b]$, form the fixed point iteration $x_{k+1} = g(x_k)$.

3) The process can be stopped by using the following error estimating formula:

$$\mid x_k - x^* \mid \leq \frac{q}{1-q} \mid x_k - x_{k-1} \mid,$$

   where q is the upper bound of $\mid g'(x) \mid$.

Example 5.4 Using the fixed point method solve $f(x) = 0$ on the interval $[1;3]$ with precision $\epsilon = 10^{-3}$, where $f(x) = x - \sqrt[5]{x} - 1$.

Solution: It is easy to express x from the equation $x - \sqrt[5]{x} - 1 = 0$ by $x = \sqrt[5]{x} + 1$, so

$$g(x) = \sqrt[5]{x} + 1.$$

Function g is continuous and strictly monotone increasing function on the interval $[1;3]$ implying

$$1 < g(1) < g(x) < g(3) < 3 \qquad \text{for all } x \in [1;3]$$

In addition

$$\mid g'(x) \mid = \left| \frac{1}{5\sqrt[5]{x^4}} \right| \leq \left| \frac{1}{5\sqrt[5]{1^4}} \right| = \frac{1}{5} = q < 1.$$

Since g satisfies all conditions, the fixed point iteration converges. If we start it from the point $x_0 = 1$ we obtain the following further iterations:

k	x_k	$\mid x_k - x^* \mid \leq \frac{q}{1-q} \mid x_k - x_{k-1} \mid$
1	$x_1 = g(x_0) = 2$	$\lvert x_1 - x^* \rvert \leq 2.5 \cdot 10^{-1}$
2	$x_2 = g(x_1) = 2.1487$	$\lvert x_2 - x^* \rvert \leq 3.7 \cdot 10^{-2}$
3	$x_3 = g(x_2) = 2.1653$	$\lvert x_3 - x^* \rvert \leq 4.1 \cdot 10^{-3}$
4	$x_4 = g(x_3) = 2.1671$	$\lvert x_4 - x^* \rvert \leq 4.5 \cdot 10^{-4}$

The error calculated to x_4 is less than 10^{-3}, the process terminates.

Similarly to Newton's method we have an opportunity to compute the iterations easily using the Ans button of our calculator. In this case the formula $\sqrt[5]{Ans+1}$ to be repeated, and after 10 iterations we get the best result which can be achieved in this way: $x_{10} = 2.167303978$.

Remark 5.3 *The fixed point method is linearly convergent, that is why it is slower than Newton's method.*

5.4 Solution of Nonlinear Systems

In this section we extend the Newton's method for solving single-variable equations to nonlinear systems of n equations with n variables which is given in the form:

$$\begin{aligned} f_1(x_1, x_2, \ldots, x_n) &= 0 \\ \vdots \quad & \quad \vdots \\ f_n(x_1, x_2, \ldots, x_n) &= 0. \end{aligned}$$

It can be written briefly as

$$F(x) = 0,$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$, x is the vector of the unknowns and 0 denotes the n-dimensional zero-vector. Similarly to the single-variable case we start from a suitable initial vector $x^{(0)} \in \mathbb{R}^n$ and construct a sequence $x^{(k)}$ which converges to the exact solution vector x^* of the system.

We linearise the scalar equations $f_i(x_1, x_2, \ldots, x_n) = 0$ at the point $x^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}]^T \in \mathbb{R}^n$, which means that the expressions on the left hand side are replaced by the first-order Taylor polynomials of them:

$$\begin{aligned} & f_i(x_1, x_2, \ldots, x_n) \approx \\ & \qquad f_i(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}) + \sum_{j=1}^{n} \frac{\partial}{\partial x_j} f_i(x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}) \left(x_j - x_j^{(k)} \right). \end{aligned}$$

This approximation can be expressed more briefly by using the gradient vector:

$$\begin{aligned} f_1(x) &\approx f_1(x^{(k)}) + \nabla f_1(x^{(k)})^T (x - x^{(k)}) \\ &\ \ \vdots \\ f_n(x) &\approx f_n(x^{(k)}) + \nabla f_n(x^{(k)})^T (x - x^{(k)}), \end{aligned} \quad ,$$

where $\nabla f_i(x^{(k)})$ denotes the gradient vector of f_i (it contains the partial derivatives of f_i at the point $x^{(k)}$).

Instead of solving the equation $F(x) = 0$, we solve the following system of linear equations:

$$\begin{aligned} f_1(x^{(k)}) + \nabla f_1(x^{(k)})^T(x - x^{(k)}) &= 0 \\ &\vdots \\ f_n(x^{(k)}) + \nabla f_n(x^{(k)})^T(x - x^{(k)}) &= 0 \end{aligned} \tag{7}$$

The solution of system (7) provides the vector $x^{(k+1)}$, which is the next term of the sequence converging to x^*.

The Jacobian matrix of the function F is given by

$$J(x) = \left[\frac{\partial f_i(x)}{\partial x_j}\right]_{i,j=1}^{n} = \begin{bmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_n(x)^T \end{bmatrix}$$

Using the Jacobian matrix, system (7) can be formulated in a very simple way:

$$F(x^{(k)}) + J(x^{(k)})(x - x^{(k)}) = 0. \tag{8}$$

The solution $x^{(k+1)}$ can be obtained by expressing the variable x:

$$x^{(k+1)} = x^{(k)} - \left[J\big(x^{(k)}\big)\right]^{-1} F(x^{(k)}) \qquad (k = 0, 1, \ldots). \tag{9}$$

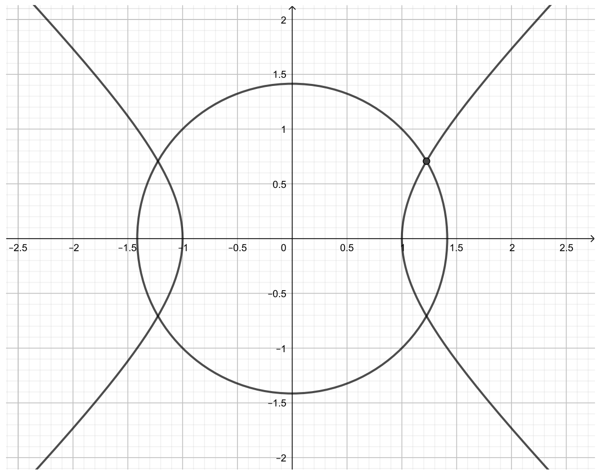
The iteration $\left\{x^{(k)}\right\}$ using formula (9) is called Newton's method for nonlinear system of equations.

Remark 5.4 *a) Formula (9) is analogue to the equation*

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$$

of the single-variable case if the product by the inverse Jacobian matrix is considered as the same operation of division by the derivative of $f(x)$.
b) In practice instead of inverting the Jacobian matrix $J(x^{(k)})$ we solve the linear system of equations (8) by a suitable solution technique.

Example 5.5 Using Newton's method find the common point of the circle $x^2+y^2 = 2$ and the hyperbola $x^2 - y^2 = 1$ in the first quarter $(x \geq 0,\ y \geq 0)$.



Solution: We introduce new variables instead of x and y, and transform the system in the form $F(x) = 0$:

$$\begin{aligned} x_1^2 + x_2^2 - 2 &= 0 \\ x_1^2 - x_2^2 - 1 &= 0. \end{aligned}$$

The system

$$\begin{aligned} f_1 &= x_1^2 + x_2^2 - 2 \\ f_2 &= x_1^2 - x_2^2 - 1 \end{aligned}$$

has the following Jacobian matrix:

$$J(x_1, x_2) = \begin{bmatrix} 2x_1 & 2x_2 \\ 2x_1 & -2x_2 \end{bmatrix}.$$

Since we search a solution in the first quarter we choose positive coordinates for the initial vector:

$$x^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The current form of the linear system (8):

$$F(x^{(0)}) + J(x^{(0)})(x - x^{(0)}) = 0.$$

Replacing vector $x^{(0)}$ in the system we have:

$$\begin{bmatrix} 0 \\ -1 \end{bmatrix} + \begin{bmatrix} 2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 - 1 \\ x_2 - 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

which is equivalent to the following linear system:

$$2x_1 + 2x_2 = 4$$
$$2x_1 - 2x_2 = 1.$$

The solution provides the next iteration $x^{(1)} = \begin{bmatrix} 1,25 \\ 0,75 \end{bmatrix}$.

Now, system (8) has the form:

$$F(x^{(1)}) + J(x^{(1)})(x - x^{(1)}) = 0.$$

Substituting $x^{(1)}$:

$$\begin{bmatrix} 0,125 \\ 0 \end{bmatrix} + \begin{bmatrix} 2,5 & 1,5 \\ 2,5 & -1,5 \end{bmatrix} \begin{bmatrix} x_1 - 1,25 \\ x_2 - 0,75 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

After completing the vector-matrix operations we obtain the linear system

$$2,5x_1 + 1,5x_2 = 4,125$$
$$2,5x_1 - 1,5x_2 = 2$$

and its solution is $x^{(2)} = \begin{bmatrix} 1,225 \\ 0,7083 \end{bmatrix}$.

Then $x^{(2)}$ is replaced in the system (8) again, and its solution provides the next iteration: $x^{(3)} = \begin{bmatrix} 1,2247 \\ 0,7071 \end{bmatrix}$.

This vector is an approximate solution of the problem, which is exact for 4 decimal digits. To prove this, solve the system in an elementary way, as follows. By adding the equations of the circle and the hyperbola we obtain $2x^2 = 3$, that is $x^2 = \frac{3}{2}$, from which $x = \pm\sqrt{\frac{3}{2}}$, and after back substitution we have: $y = \pm\sqrt{\frac{1}{2}}$. Combining the signs in all possible ways we got four solutions; we need the solution of positive coordinates:

$$x^* = \begin{bmatrix} \sqrt{\frac{3}{2}} \\ \sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} 1,22474487 \\ 0,70710678 \end{bmatrix}.$$

Example 5.6 Consider the system

$$\begin{aligned} -x_1^2 + \quad\quad &+ x_3 + 3 = 0 \\ -x_1 + 2x_2^2 &- x_3^2 - 3 = 0 \\ x_2 &- 3x_3^2 + 2 = 0. \end{aligned}$$

Starting from the initial vector $x^{(0)} = \begin{bmatrix} 2 & 2 & 1 \end{bmatrix}^T$ apply two iterations of Newton's method.

Solution:

$$x^{(1)} = \begin{bmatrix} 2,0328 \\ 1,7869 \\ 1,1311 \end{bmatrix}, \qquad x^{(2)} = \begin{bmatrix} 2,0302 \\ 1,7732 \\ 1,1215 \end{bmatrix}.$$

6 Numerical Solution of Differential Equations

Differential equations are composed of an unknown function and its derivatives. Such equations play a fundamental role in science and engineering and have a wide range of applications. Differential equations can be classified as to the number of the independent variables of the unknown function. An ordinary differential equation (ODE) relates a function of a single variable to its ordinary derivatives. Alternatively, a partial differential equation (PDE) deals with a function of several independent variables and its partial derivatives. Differential equations are also classified as to their order. A differential equation is said to be of the nth order if it involves a derivative of the nth order, but no derivative of an order higher than this. We focus on numerical methods for solving first-order ODEs, which are given in the form

$$y' = f(x, y). \tag{10}$$

A function $y = y(x)$ satisfying equation (10) is called a solution of (10). In general an ODE has infinitely many solutions which differ from each other in an additive constant c. For example, the functions $y = \sin x + c$, where c is a real constant, represent all solutions of the differential equation $y' = \cos x$. In this example the solution is obtained analytically, using a known rule of integration. In contrast, most of the numerical methods are not capable to find the solution as a continuous function, they provide only approximate values of y corresponding to some finite set of variables.

Definition 6.1 (IVP and its solution) An initial value problem (IVP) is given in the form

$$y' = f(x, y) \text{ with } y(x_0) = y_0, \tag{11}$$

where x_0 and y_0 are given initial values of the independent variable x and the function y, respectively. A solution to problem (11) on an interval $[x_0, b]$ is a differentiable function $y = y(x)$ such that

$$y'(x) = f(x, y) \text{ for all } x \in [x_0, b] \text{ and } y(x_0) = y_0.$$

Example 6.1 Consider the IVP

$$y' = \frac{2y}{x} - 2 \text{ with } y(1) = 3.$$

Its solution is $y(x) = x^2 + 2x$ for $x \geq 1$, since $y'(x) = 2x + 2 = \dfrac{2y(x)}{x} - 2$ for all x, and $y(1) = 1^2 + 2 \cdot 1 = 3$.

6.1 Taylor Series Solution

Consider an IVP in the form

$$y' = f(x, y), \qquad y(x_0) = y_0.$$

The Taylor polynomial of degree n for the function $y(x)$ about x_0 is given by the sum

$$\begin{aligned} y_n(x) &= y(x_0) + y'(x_0)(x - x_0) + \frac{y''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{y^{(n)}(x_0)}{n!}(x - x_0)^n \\ &= \sum_{k=0}^{n} \frac{y^{(k)}(x_0)}{k!}(x - x_0)^k, \end{aligned}$$

where n is a nonnegative integer. We use the IVP for calculating the coefficients of the sequence $y_n(x)$ as illustrated in the following example.

Example 6.2 Given the IVP

$$y' = 3x + y^2, \qquad y(0) = 1.$$

Here $x_0 = 0$ and $y(x_0) = y_0 = 1$, and from the differential equation $y'(x_0) = 3x_0 + y_0^2 = 3 \cdot 0 + 1 = 1$, so the Taylor polynomial of degree 1 is $y_1(x) = 1 + x$. To compute the coefficient $y''(x_0)$ we differentiate the equation:

$$y'' = (3x + y^2)' = 3 + 2yy' \Longrightarrow y''(x_0) = 3 + 2y(x_0)y'(x_0) = 3 + 2 \cdot 1 \cdot 1 = 5.$$

So the Taylor polynomial of degree 2 is $y_2(x) = 1 + x + \frac{5}{2}x^2$. Now, we differentiate y'':

$$y''' = 2y'y' + 2yy'' \Longrightarrow y'''(x_0) = 2 \cdot 1^2 + 2 \cdot 1 \cdot 5 = 12,$$

and we obtain the next approximation of y as $y_3(x) = 1 + x + \frac{5}{2}x^2 + \frac{12}{6}x^3$. We can continue in this manner to get better approximations of the unknown function y, or we can use the obtained functions to estimate the values y near x_0. For example, using the Taylor polynomial of degree 3, we have $y(0.1) \approx 1 + 0.1 + 2.5 \cdot 0.1^2 + 2 \cdot 0.1^3 = 1.127$.

6.2 Picard's Method

Similarly to the Taylor series solution the following iteration provides a sequence $y_n(x)$ which can be used as an approximation of the solution to the initial value problem. A function $y(x)$ is a solution to the IVP given in (11) if and only if it is a solution to the integral equation

$$y(x) = y(x_0) + \int_{x_0}^{x} f(u, y(u))\, du. \tag{12}$$

A solution of the integral equation (12) can be constructed using Picard's method as follows:

$$\begin{aligned}
y_0(x) &= y_0, \\
y_1(x) &= y_0 + \int_{x_0}^{x} f(u, y_0(u))\, du, \\
y_2(x) &= y_0 + \int_{x_0}^{x} f(u, y_1(u))\, du, \\
&\vdots \\
y_n(x) &= y_0 + \int_{x_0}^{x} f(u, y_{n-1}(u))\, du.
\end{aligned}$$

Example 6.3 Solve the IVP

$$y' = y + x, \qquad y(0) = 0$$

using Picard's method.

$$\begin{aligned}
&y_0(x) = y_0 = 0, \\
&y_1(x) = y_0 + \int_{x_0}^{x} f(u, y_0(u))\, du = 0 + \int_{x_0}^{x} (u + 0)\, du = \left[\frac{u^2}{2}\right]_0^x = \frac{x^2}{2}, \\
&y_2(x) = y_0 + \int_{x_0}^{x} f(u, y_1(u))\, du = 0 + \int_{x_0}^{x} (u + \frac{u^2}{2})\, du = \left[\frac{u^2}{2} + \frac{u^3}{6}\right]_0^x = \frac{x^2}{2} + \frac{x^3}{6}, \\
&y_3(x) = y_0 + \int_{x_0}^{x} f(u, y_2(u))\, du = 0 + \int_{x_0}^{x} (u + \frac{u^2}{2} + \frac{u^3}{6})\, du = \left[\frac{u^2}{2} + \frac{u^3}{6} + \frac{u^4}{24}\right]_0^x \\
&\quad = \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24}.
\end{aligned}$$

Note, that we have obtained members from the Taylor series expansion

$$e^x = \sum_{k=1}^{\infty} \frac{x^k}{k!},$$

so the solution of the differential equation is $y(x) = e^x - x - 1$.

6.3 Euler's Method

Consider the IVP

$$y' = f(x, y), \qquad y(x_0) = y_0.$$

We use the Taylor polynomial of degree 1 to approximate $y(x)$:

$$y(x) \approx y(x_0) + y'(x_0)(x - x_0).$$

Using the notation $h = x - x_0$ and replace the term $y'(x_0)$ by $f(x_0, y(x_0))$ from the IVP, we have

$$y_1 = y(x_0 + h) = y_0 + hf(x_0, y(x_0)),$$

which is the formula of a single step of Euler's method. Contrary to the Taylor series and Picard's method Euler's method does not provide a sequence of functions $y_n(x)$ converging to the solution of the IVP. Instead of it a table can be obtained over interval $[x_0, b]$, containing the approximate values of $y(x)$. Divide the interval $[x_0, b]$ into n equal subintervals, each of length $h = \frac{b-x_0}{n}$. h is called the step size. We have $n + 1$ mesh points $x_k = x_0 + kh$, $k = 0, 1, \cdots, n$, in which the approximate solution is given as

$$\begin{aligned}
& & y_0 &= y(x_0) \\
x_1 &= x_0 + h, & y_1 &= y(x_1) = y_0 + hf(x_0, y_0)) \\
x_2 &= x_1 + h, & y_2 &= y(x_2) = y_1 + hf(x_1, y_1)) \\
& & &\vdots \\
x_{k+1} &= x_k + h, & y_{k+1} &= y(x_{k+1}) = y_k + hf(x_k, y_k))
\end{aligned} \tag{13}$$

Example 6.4 Apply Euler's method three times with $h = 0.1$ for the IVP

$$y' = x \cdot \sqrt[3]{y}, \qquad y(2) = 1.$$

We use the scheme given in (13):

$y_0 = y(2) = 1$

$y_1 = y(2.1) = 1 + hf(2, 1)) = 1 + 0.1 \cdot 2 \cdot \sqrt[3]{1} = 1.2$

$y_2 = y(2.2) = 1.2 + hf(2.1, 1.2) = 1.2 + 0.1 \cdot 2.1 \cdot \sqrt[3]{1.2} = 1.4232$

$y_3 = y(2.3) = 1.4232 + hf(2.2, 1.4232) = 1.4232 + 0.1 \cdot 2.2 \cdot \sqrt[3]{1.4232} = 1.6707.$

6.4 Runge-Kutta Methods

The Runge-Kutta (RK) methods allow one to achieve a good accuracy without computing the higher-order derivatives. As before, we consider the IVP

$$y' = f(x, y), \qquad y(x_0) = y_0.$$

Similarly to the Euler's method, assuming that the value y_k of the approximate solution at point x_k has been computed, we want to find a proper value y_{k+1} of the numerical solution at the next point $x_{k+1} = x_k + h$. We generalize Euler's method formula for y_{k+1} as follows:

$$y_{k+1} = y_k + hG(x_k, y_k, h),$$

where $G(x_k, y_k, h)$ is called an increment function and can be written in a general form as

$$G(x_k, y_k, h) = c_1 f_1 + c_2 f_2 + \cdots + p_m f_m.$$

In this expression, m is some positive integer called the order of the method, $p_1, p_2, \ldots, p_m$ are constants and the numbers $f_1, f_2, \ldots, f_m$ are given by the formulas

$$\begin{aligned}
f_1 &= f(x_k, y_k), \\
f_2 &= f(x_k + \alpha_1 h, y_k + \alpha_1 h f_1), \\
f_3 &= f(x_k + \alpha_2 h, y_k + \alpha_2 h f_2), \\
&\vdots \\
f_m &= f(x_k + \alpha_{m-1} h, y_k + \alpha_{m-1} h f_{m-1}),
\end{aligned} \tag{14}$$

The coefficients $\alpha_1, \ldots, \alpha_{m-1}, p_1, \ldots, p_m$ are chosen such that the precision of the approximation is sufficiently good. Given y_k, one can recursively compute $f_1, \ldots f_m$ and then y_{k+1}. The first-order Runge-Kutta method ($m = 1$) is, in fact, the previously discussed Euler method.

Among the Runge-Kutta methods the most popular are the fourth-order methods. Its classical version uses the coefficients

$$\alpha_1 = \alpha_2 = \frac{1}{2},\ \alpha_3 = 1 \ \text{ and } \ p_1 = p_4 = \frac{h}{6},\ p_2 = p_3 = \frac{h}{3},$$

so y_{k+1} can be computed as $y_{k+1} = y_k + h\dfrac{f_1 + 2f_2 + 2f_3 + f_4}{6}$, where

$$\begin{aligned}
f_1 &= f(x_k, y_k), \\
f_2 &= f(x_k + 0.5h, y_k + 0.5hf_1), \\
f_3 &= f(x_k + 0.5h, y_k + 0.5hf_2), \\
f_4 &= f(x_k + h, y_k + hf_3).
\end{aligned}$$

Example 6.5 Using the classical fourth-order Runge-Kutta method give an approximation for y_1 with $h = 0.1$ in case of the IVP discussed in Example 6.2:

$$y' = 3x + y^2, \qquad y(0) = 1.$$

First compute the coefficients $f_i,\ i = 1, 2, 3, 4$:

$$\begin{aligned}
f_1 &= f(x_0, y_0) = 3 \cdot 0 + 1^2 = 1,\\
f_2 &= f(x_0 + 0.5h, y_0 + 0.5hf_1) = f(0.05, 1.05) = 1.2525,\\
f_3 &= f(x_0 + 0.5h, y_0 + 0.5hf_2) = f(0.05, 1.062625) = 1.2792,\\
f_4 &= f(x_0 + h, y_0 + hf_3) = f(0.1, 1.12792) = 1.5722.
\end{aligned}$$

Replacing these numbers into the formula $y_1 = y_0 + h\frac{f_1+2f_2+2f_3+f_4}{6}$ we obtain

$$y_1 = y(0.1) = 1 + 0.1 \cdot \frac{1 + 2 \cdot 1.2525 + 2 \cdot 1.2792 + 1.5722}{6} \approx 1,1273.$$

Remember that similar approximation was achieved by using the Taylor series method in Example 6.2, but in this case higher order derivatives are not required.

7 Basic Concepts of Optimization

Optimization is a methodology aiming to find the best among available alternatives. The available alternatives are referred to as *feasible solutions*, and their quality is measured using some numerical function called the *objective function*. A feasible solution that yields the best (minimum or maximum) objective function value is called an *optimal solution*.

Steps of formulating an optimization model:

1. Introduce the *decision variables*, which are the parameters whose values need to be determined in order to solve the problem.

2. State the *objective function*, i.e., the quantity that we need to optimize as a function of the decision variables.

3. Specify the *constraints*, i.e., the conditions that the design parameters are required to satisfy.

7.1 Mathematical Description of Optimization Models

Let $X \subseteq \mathbb{R}^n$ be a set of n-dimensional real vectors in the form $x = [x_1, \ldots, x_n]^T$, and let $f : X \to \mathbb{R}$ be a given function. In mathematical terms, an optimization problem has the general form

$$\begin{array}{rcr} \text{maximize } f(x) & \quad\text{or}\quad & \text{minimize } f(x) \\ \text{subject to } x \in X & & \text{subject to } x \in X. \end{array}$$

In the above, each x_j, $j = 1, 2 \ldots n$, is called a *decision variable*, X is called the *feasible (admissible) region*, and f is the *objective function*.

Note that a maximization problem $\max\limits_{x\in X} f(x)$ can be easily converted into an equivalent minimization problem $\min\limits_{x\in X} (-f(x))$ and vice versa.

We will typically use a *functional form* of an optimization problem, in which the feasible region is described by a set of *constraints* given by equalities and inequalities in the form

$$\begin{aligned} h_i(x) &= 0, \quad i = 1, 2, \ldots p, \\ g_j(x) &\leq 0, \quad j = 1, 2, \ldots r, \end{aligned}$$

where $h_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, 2, \ldots p,\ g_j : \mathbb{R}^n \to \mathbb{R},\ j = 1, 2, \ldots r$ are given functions.

Definition 7.1 (Global Minimum) The point $x^* \in X$ is a point of global minimum (global minimizer) for the problem $\min_{x \in X} f(x)$ if

$$f(x^*) \leq f(x) \text{ for all } x \in X.$$

Definition 7.2 (Local Minimum) The point $x^* \in X$ is a point of local minimum (local minimizer) for the problem $\min_{x \in X} f(x)$ if there exists $\epsilon > 0$ such that

$$f(x^*) \leq f(x) \text{ for any } x \in X \text{ with } ||x - x^*|| \leq \epsilon.$$

Remark 7.1 *The concepts of global and local maximizers can be easily given by writing*

$$f(x^*) \geq f(x)$$

instead of $f(x^*) \leq f(x)$ *in Definitions 7.1 and 7.2.*

7.2 Classification of Optimization Models

We can group the models

- by the domain of the decision variables
  - continuous models: the variables can be any points of an interval,
  - discrete models: the variables can be only determined numbers, e.g. integers;
- by the functions of the model
  - linear model: all constraints and the objective function are linear,
  - nonlinear model: at least one of the functions is nonlinear;
- by the dependence of randomness
  - deterministic model: all the parameters are constant,
  - stochastic model: some parameters are random variables.

8 Solving Linear Programming (LP) Problems

Linear programming is a methodology for solving linear optimization problems, in which one wants to optimize a linear objective function subject to constraints on its variables expressed in terms of linear equalities and/or inequalities. Ever since the introduction of the simplex method by George Dantzig in the late 1940s, linear programming has played a major role in shaping the modern horizons of the field of optimization and its applications.

8.1 Solving Two-Variable LP Problems Graphically

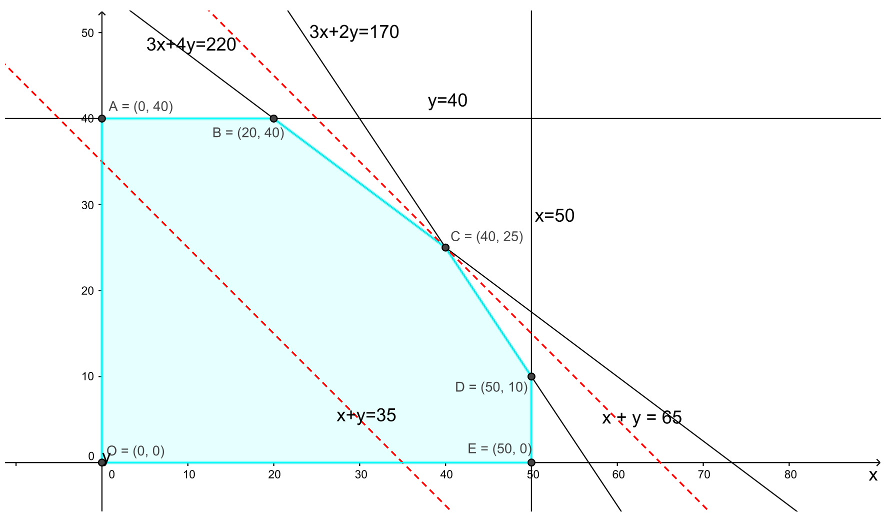
In linear programming problems the objective function and all the constraints are linear. In this chapter we study simple LP problems with two decision variables, x_1 and x_2. In this case we can solve the problem graphically after plotting the set of feasible points in two-dimensional space. Instead of x_1 and x_2 we denote the variables by x and y as usual in coordinate geometry.

Example 8.1 (Sandwich problem: a maximization problem) A mother makes two types of sandwiches for the birthday party of her son. A salami sandwich consists of 3 g butter, 3 slices of egg and two slices of salami. A ham sandwich consists of 4 g butter, 2 slices of egg and a slice of ham. (Of course these ingredients are in between two slices of bread in both cases.) For making sandwiches 100 slices of salami, 40 slices of ham, 170 slices of egg and 220 g butter are available. There is plenty of bread, so no restriction is needed for bread. How many sandwiches of each type should be made to maximize the total number of sandwiches?

Solution: First we have to set the mathematical model of the problem, which means that after introducing suitable variables we give the objective function and formulate the constraints. Suppose that the number of salami sandwiches to be made is x, and the number of ham sandwiches is y. A salami sandwich requires $3x$ g butter, a ham sandwich requires $4y$ g butter, but only 220 g is available, so the inequality $3x + 4y \leq 220$ must be valid. Similarly, we need altogether $3x + 2y$ slices of egg, which amount can be at most 170. We have the constraint $2x \leq 100$ for the salami, and $y \leq 40$ in the case of ham. In addition, neither x or y can be negative, that is, $x \geq 0$ and $y \geq 0$. We look for the maximal value of the amount $x + y$ subject to the previous constraints. Summarizing, the model can be formulated as follows:

$$\begin{aligned} x + y &\longrightarrow \max \\ 3x + 4y &\leq 220 \\ 3x + 2y &\leq 170 \\ 2x &\leq 100 \\ y &\leq 40 \\ x &\geq 0 \\ y &\geq 0 \end{aligned}$$

We solve the problem graphically, plotting the points satisfying all the six constraints. For example points $(x;y)$ for which $3x + 4y \leq 220$ are in the half-plane under the line $3x + 4y = 220$, and the points satisfying the inequality $y \geq 0$ are above the x axis. The shadowed region in the next figure illustrates the set L of the feasible solutions, its points satisfy all constraints. We have to select the point in L which corresponds to the maximal objective function value: for which the sum of its coordinates is maximal. We fix a number z and consider the line $x + y = z$. (The line where the objective function is constant is often called the *isoprofit line*.) We need the maximal z, for which the line $x + y = z$ and the polygon L have at least one common point. The value z can be read from the y axis, so we have to move a line with slope -1 parallel to itself and over the set L, and find where it intersects the y axis at the highest point. In our case we could push the line to the point (40;25) of L, z is maximal in this point and its value is 65. The solution of the problem: 40 salami sandwiches and 25 ham sandwiches can be made for the party, and the optimum is 65.

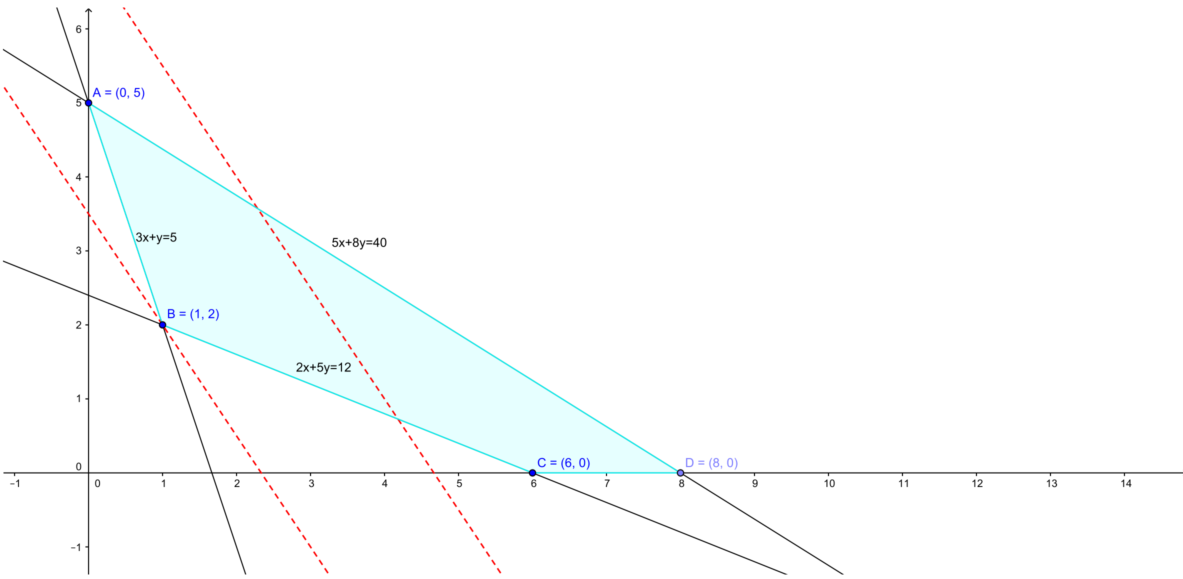


Example 8.2 (Diet problem: a minimization problem) A person is on a diet and eats two types of fruit, lemon and apple, to cover his vitamin C and iron needs. Suppose that the minimum recommended daily intake is 500 units of vitamin C, and 1.2 mg from iron. 100 g lemon contains 300 units of vitamin C and 0.2 mg iron, and there are 100 units of vitamin C and 0.5 mg iron in 100 g apple. The person wants to take in at most 400 calories by eating the fruits. 100 g lemon provides 50 calories and 100 g apple provides 80 calories. At the grocery's shop one kg lemon costs 600 Ft and one kg apple is 400 Ft. What combination of fruit should be bought to cover the needs, not to exceed the calories and to minimize the price of the fruit?

Solution: Suppose that the person buys $x \cdot 100$ g lemon and $y \cdot 100$ g apple in the shop. The price is $60x + 40y$,it is to be minimized. Both variables are nonnegative and we have three further constraints:

$$60x + 40y \longrightarrow \min$$
$$300x + 100y \geq 500$$
$$0.2x + 0.5y \geq 1.2$$
$$50x + 80y \leq 400$$
$$x \geq 0$$
$$y \geq 0$$

We plot the set of the feasible solutions in two-dimensional space. Similarly to the sandwich problem, we have again a polygon in the plane (see figure). If we fix the value of the objective function in z, and express variable y, the equation $y = -\frac{3}{2}x + z$ is obtained. In this case we have to move a line with slope -1.5 over the polygon, and seek the lowest point, where the line intersects the y axis. This line leaves the polygon at point (1;2) and intersects the y axis at 3.5. This means that the person needs 100 g lemon and 200 g apple a day, and it costs 140 Ft. (Some explanation to the optimum: we obtain 140 by multiplying the value 3.5 by 40, because in the beginning, the objective function was divided by 40.)

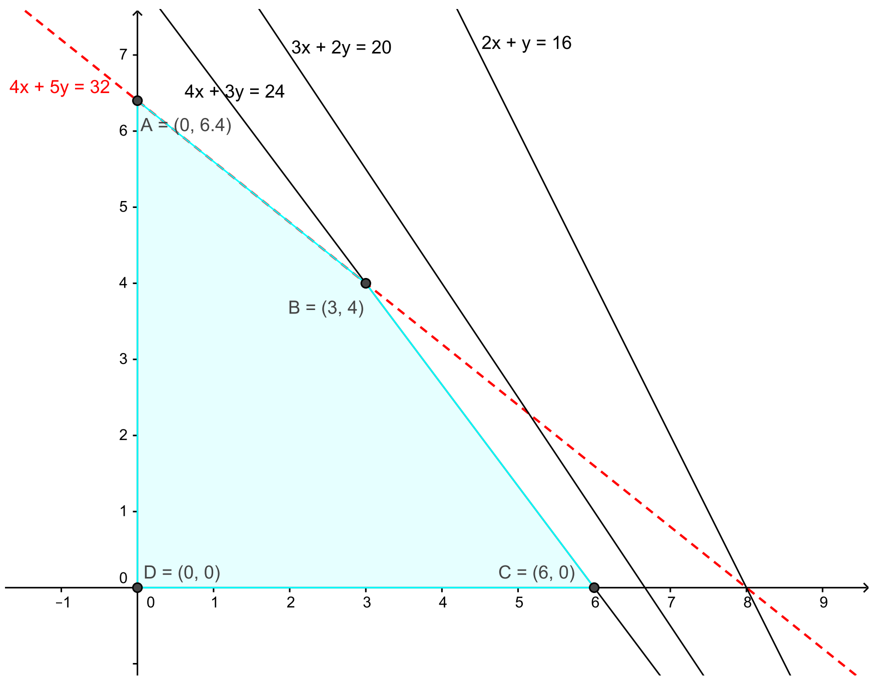


Example 8.3 (A problem with multiple optimal solutions) A builder

finds he is commonly asked to build two types of buildings, A and B. The profits per building are \$4,000 and \$5,000 for A and B, respectively. There are some restrictions on available materials. Building A requires 4,000 m^2 of timber, 4 t of steel, 300 m^2 of roofing iron and 200 $m^3$of concrete. Building B requires 5000 m^2 timber, 3 t of steel, 200 m^2 of roofing iron and 100 m^3 of concrete. However only 32,000 m^2 of timber, 24 t of steel, 2,000 m^2 of roofing iron and 1600 m^3 of concrete are available per year. What combination of A and B should be built per year to maximize profit?

Solution: Let x the number of building A, and y the number of building B to be constructed. The problem can now be modeled as follows:

$$\begin{aligned} 4000x + 5000y &\longrightarrow \max \\ 4000x + 5000y &\leq 32000 \\ 4x + 3y &\leq 24 \\ 300x + 200y &\leq 2000 \\ 200x + 100y &\leq 1600 \\ x &\geq 0 \\ y &\geq 0 \end{aligned}$$



After plotting the set of feasible solutions, we can observe that the constraints related to the roofing iron and the concrete are not redundant, in the sense that they do not define part of the boundary of the feasible region. Otherwise, the coefficients of the objective function are the same as the coefficients of the first constraint, so when the objective function is drawn at the optimal level, it coincides with a boundary line of the feasible region. This means that all points on this border line represent optimal solutions. The only integer solution is the point (3;4) with $32,000 as an optimum. But the builder has the same profit if he builds 2 of building A, and 4.8 of building B, which means that the fifth building is ready up to 80% at the end of the year. (When we accept only integer solutions we need other techniques for searching optimal solutions, see the topic of Integer Programming.)

Example 8.4 (No feasible solutions) A small clothing factory makes shirts and skirts. A profit of $4 and $3 is made from a shirt and a skirt respectively. A shirt requires 3 meters of material and a skirt 4, with only 12 meters available daily for a worker. It takes 5 hours of total time to make a shirt, and 2 hours to make a skirt. Supposing 10 working hours daily for a worker, the boss expects at least 4 garments per worker a day. How is it possible to maximize profit?

Solution: Let x be the number of the shirts and y be the number of the skirts. The problem can be formulated as follows:

$$\begin{aligned} 4x + 3y &\longrightarrow \max \\ 3x + 4y &\leq 12 \\ 5x + 2y &\leq 10 \\ x + y &\geq 4 \\ x &\geq 0 \\ y &\geq 0 \end{aligned}$$

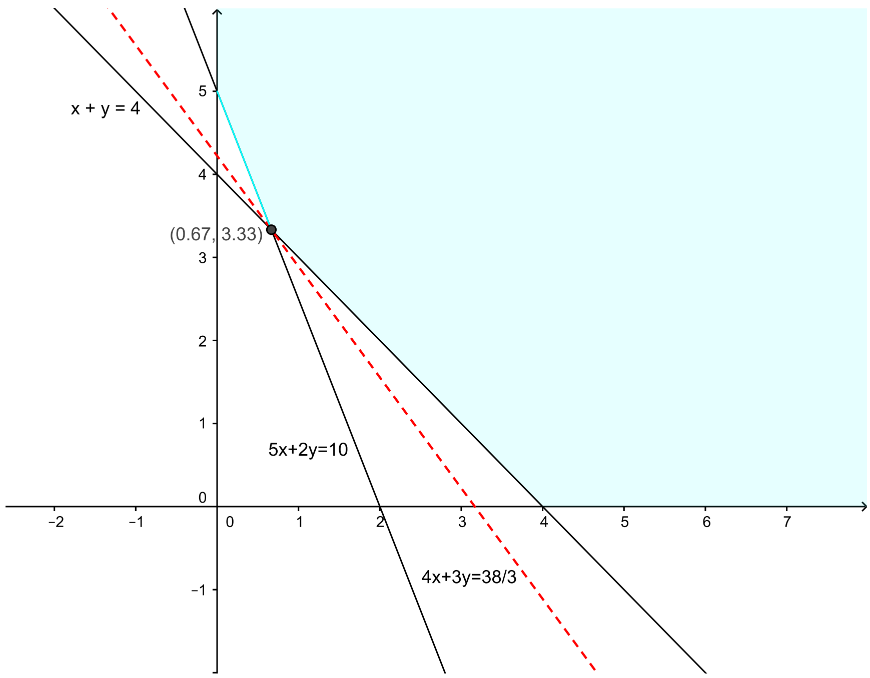
When we try to draw the feasible region of the problem, we find, that there does not exist a point which will satisfy all constraints simultaneously. Hence the problem does not have a feasible solution, and so it can not have any optimal solution: the boss has unrealistic expectations for his workers.

Example 8.5 (Unbounded feasible region) We reformulate the clothing factory problem studied before. Suppose that the boss of the factory got a lot of material at a bargain price, and lifted the limit concerning the usage of material. However, he forces every employee to work at least 10 hours (and there is no upper bound for the working hours). The constraint of making at least 4 garments remains. In this new situation is it possible to achieve maximal profit?

Solution:

$$\begin{aligned} 4x + 3y &\longrightarrow \max \\ 5x + 2y &\geq 10 \\ x + y &\geq 4 \\ x &\geq 0 \\ y &\geq 0 \end{aligned}$$

The feasible region of the problem is not bounded. The line representing the objective function can be moved parallel to itself an arbitrary distance from the origin and still coincide with feasible points. Therefore the problem has no bounded maximal solution. (If the workers are able to work for infinitely long, the profit can be arbitrarily large). However, it is interesting to see, that the minimization problem subject to the constraints can be solved: the minimum profit is $\$\frac{38}{3}$ which can be achieved by making $\frac{2}{3}$ shirts and $\frac{10}{3}$ skirts.



8.2 The Standard Form of a Linear Programming Problem

In linear programming problems the objective function and all the constraints are linear functions of the decision variables $x_1, x_2, ..., x_n$. It is obvious that there are many variations of LP problems. Their objective can be maximization or minimization, the constraints can be given in equalities or inequalities form, and the variables can be restricted or unrestricted in sign. Fortunately it is not necessary to develop different methods for each class of problems. We will present a method which is suitable for solving the problems of one common and wide class: the problems which are given in standard form.

An LP problem is in *standard form* if it can be expressed as:

$$\begin{aligned} c^T x &\longrightarrow \max \\ Ax &= b, \quad (b \geq 0) \\ x &\geq 0 \end{aligned}$$

where

$c = (c_1, c_2, ..., c_n)^T$ the coefficients of the objective function,

$x = (x_1, x_2, ..., x_n)^T$ the vector of the decision variables,

$A = (a_{i,j})_{m \times n}$ the matrix containing the coefficients of the constraints,

$b = (b_1, b_2, ..., b_m)^T$ the right-hand-side constants,

$0 = (0)_{n \times 1}$ a null vector with n components.

The steps that transform an arbitrary LP problem into standard form are as follows:

1. A minimizing problem can be transformed into a maximizing problem by changing the signs of all coefficients of the objective function, because the function $f(x)$ has a minimum in the point x, when the function $-f(x)$ has its maximum.

2. Each negative right-hand-side constant can be made positive by multiplying the entire equality or inequality by -1.

3. An unrestricted decision variable x_i is replaced by $x_i' - x_i''$, where both new variables are nonnegative. If the constraint $x_j \leq 0$ is prescribed, we introduce a new nonnegative variable with $\widetilde{x}_j = -x_j$.

4. Each inequality can be made an equality by adding a nonnegative *slack variable* to a $\leq$ constraint, or subtracting a nonnegative *excess variable*

from a $\geq$ constraint. More precisely, if a constraint is given in the form $a_{k1}x_1 + a_{k2}x_2 + ... + a_{kn}x_n \leq b_k$, we transform it into the equality

$$a_{k1}x_1 + a_{k2}x_2 + ... + a_{kn}x_n + u_k = b_k,$$

and if a constraint is given in the form $a_{l1}x_1 + a_{l2}x_2 + ... + a_{\mathrm{ln}}x_n \geq b_l$, instead of it we write

$$a_{l1}x_1 + a_{l2}x_2 + ... + a_{\mathrm{ln}}x_n - v_l = b_l$$

where the new constraints $u_k \geq 0$ and $v_l \geq 0$ are added to the problem.

Example 8.6 Determine the standard form of this general LP problem:

$$\begin{aligned} 2x_1 + 3x_2 - 4x_3 &\longrightarrow \min \\ x_1 - x_2 \qquad\quad &\leq 5 \\ -x_1 + 2x_2 + x_3 &\leq -2 \\ 3x_1 - x_2 + x_3 &= 10 \\ x_1 \geq 0,\ x_2 \text{ unrestricted},&\ x_3 \leq 0 \end{aligned}$$

Solution:

1. Transform into a maximizing problem by multiplying the objective function by -1:
$$-2x_1 - 3x_2 + 4x_3 \longrightarrow \max$$

2. The right-hand side of the second constraint is negative, so the entire row is multiplied by -1: $x_1 - 2x_2 - x_3 \geq 2$.

3. We write $x_2' - x_2''$ instead of x_2, and $-\widetilde{x}_3$ instead of x_3.

4. A new slack variable u_1 is added to the first constraint and an excess variable v_2 is subtracted from the second one.

The result is the following standard LP problem:

$$\begin{gathered} -2x_1 - 3x_2' + 3x_2'' - 4\widetilde{x}_3 \longrightarrow \max \\ x_1 - x_2' + x_2'' + u_1 = 5 \\ x_1 - 2x_2' + 2x_2'' + \widetilde{x}_3 - v_2 = 2 \\ 3x_1 - x_2' + x_2'' - \widetilde{x}_3 = 10 \\ x_1 \geq 0,\ x_2' \ \geq 0, x_2'' \ \geq 0,\ \widetilde{x}_3 \geq 0, u_1 \geq 0, v_2 \geq 0 \end{gathered}$$

8.3 Geometrical Background of LP Problems

We studied the graphical solution of simple LP problems in Chapter 8.1. We found that the feasible region of a two-variable problem is a polygon or an unbounded set with a finite number of corners. The objective function reaches its optimum in a corner of the feasible set or all the points of a boundary line provide optimal solutions of the problem. Similar geometrical considerations can be used in three-dimensional space, but a problem with 4 or more variables is not solvable in this way. In order to obtain a general solution method for the high-dimensional problems we introduce some basic notions and facts for extending the concepts applied in two-dimensional space.

Definition 8.1 (line segment) Let x_1 and x_2 be two given points in the n-dimensional space $\mathbb{R}^n$. The line segment joining x_1 and x_2 contains all the points $x = \lambda x_1 + (1-\lambda)\, x_2$, where $0 \leq \lambda \leq 1$.

Definition 8.2 (convex set) A nonempty set C of $\mathbb{R}^n$ is called convex if the line segment joining any two points of C is also in C. That is, C is convex if $\lambda x_1 + (1-\lambda)\, x_2 \in C$, for all $x_1, x_2 \in C$ and $0 \leq \lambda \leq 1$.

Definition 8.3 (bounded set) A set $H \subset \mathbb{R}^n$ is called bounded if there exists a real number k, for which $||x_1 - x_2|| \leq k$ for every $x_1, x_2 \in H$.

Definition 8.4 (extreme point) Let $C \subset \mathbb{R}^n$ be a convex set. A point $E \in C$ is called an extreme point of C if E is not an inner point of some line segment in C, that is, E cannot be expressed as $E = \lambda x_1 + (1-\lambda)\, x_2$, for some $0 < \lambda < 1$, $x_1 \neq x_2$, x_1, $x_2 \in C$.

Consider the feasible region $X = \{x \in \mathbb{R}^n \mid Ax = b,\ x \geq 0\}$ of an LP in the standard form. It can be proved, that the set X is convex. Assume, that in the system of linear equations $Ax = b$ we have n variables and k equations. In general $n \geq k$, because when transforming the LP into standard form, we introduce new variables, and the number of constraints is fixed. There are $\binom{n}{k}$ different ways of choosing a set of k *basic variables* and solve the system leaving the other variables to be 0. The solution, where we have k variables having nonzero values and $n-k$ variables with zero values is called a *basic solution* to the system. If all variables have nonnegative values in a basic solution, then the solution is called a *basic feasible solution.*

Theorem 8.1 *Any basic feasible solution of the system $Ax = b$ represents an extreme point of the feasible region $X = \{x \in \mathbb{R}^n \mid Ax = b,\ x \geq 0\}$.*

Definition 8.5 An LP is called

- feasible, if it has at least one feasible solution and infeasible, otherwise;
- optimal, if it has an optimal solution;
- unbounded, if it is feasible and its objective function is not bounded (from above for a maximization problem and from below for a minimization problem) in the feasible region.

Theorem 8.2 *The following statements hold about the optimality of an LP problem:*

- *If an LP is not optimal, it is either unbounded or infeasible.*
- *If the feasible region $X = \{x \in \mathbb{R}^n \mid Ax = b,\ x \geq 0\}$ of an LP is bounded, then the LP is optimal.*
- *If an LP is optimal, it either has a unique optimal solution or infinitely many optimal solutions.*
- *If an LP is optimal, then an optimal solution of the problem corresponds to an extreme point of X.*

Example 8.7 Consider the convex set given by the inequations

$$\begin{aligned} x_1 + 2x_2 &\leq 8 \\ 3x_1 + x_2 &\leq 9 \\ x_1, x_2 &\geq 0 \end{aligned} \tag{15}$$

Transform it in the form $Ax = b,\ x \geq 0$ and determine the basic feasible solutions of the system.

Solution: Two slack variables u_1 and u_2 are introduced, and we have the system:

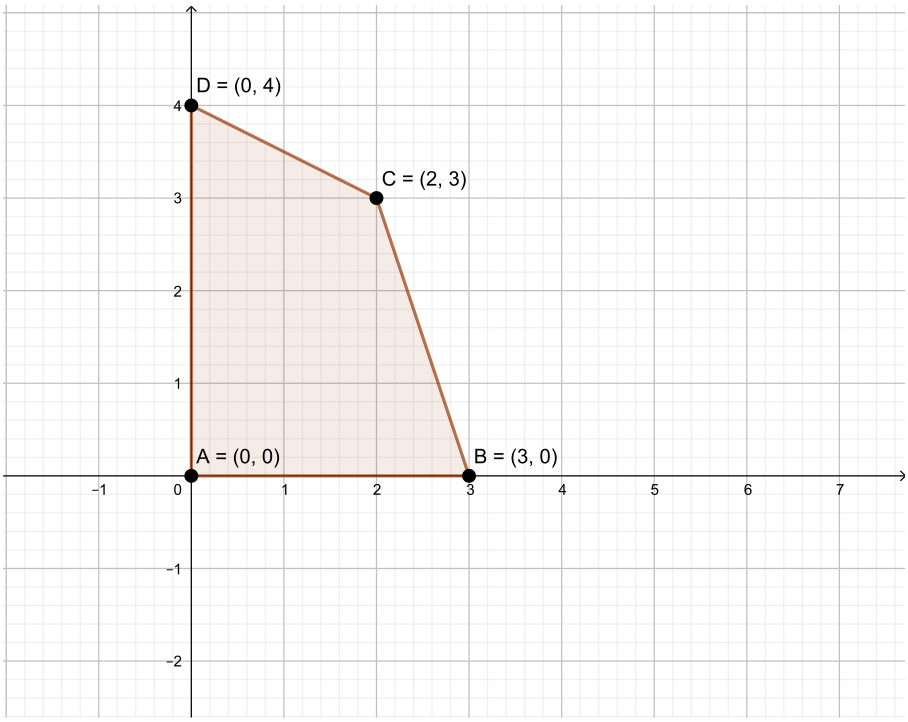
$$\begin{aligned} x_1 + 2x_2 + u_1 &= 8 \\ 3x_1 + x_2 + u_2 &= 9 \\ x_1, x_2, u_1, u_2 &\geq 0 \end{aligned}$$

We have four variables. Two of them are chosen to be basic variables, the other two are nonbasic variables. The simplest way to determine the basic solutions is to delete two variables from the system (giving zero value to them) and solve the remaining system. For example, if $x_1 = 0$ and $x_2 = 0$, the value of the other two variables can be easily obtained as $u_1 = 8,\ u_2 = 9$.

Similarly, if $x_1 = 0$ and $u_1 = 0$, then $x_2 = 4$ and $u_2 = 5$. If $x_1 = 0$ and $u_2 = 0$, then $x_2 = 9$ and $u_1 = -10$. If $x_2 = 0$ and $u_1 = 0$, then $x_1 = 8$ and $u_2 = -15$. If $x_2 = 0$ and $u_2 = 0$, then $x_1 = 3$ and $u_1 = 5$, and finally, if $u_1 = 0$ and $u_2 = 0$, then $x_1 = 2$ and $x_2 = 3$.
The results are summarized in the following table:

x_1	0	0	0	8	3	2
x_2	0	4	9	0	0	3
u_1	8	0	-10	0	5	0
u_2	9	5	0	-15	0	0

We obtained $\binom{4}{2} = 6$ basic solutions, from which 4 are basic feasible solutions containing only nonnegative coordinates. The basic feasible solutions correspond to the extreme points of the convex set (15) as it is demonstrated in the following figure:



9 The Simplex Algorithm

In the previous chapter it was shown that if a standard LP problem has an optimal solution it corresponds to an extreme point of the feasible set X. The extreme points of X can be obtained by determining the basic feasible solutions of the system $Ax = b$. In order to determine the optimal solution of an LP problem we will use a very efficient technique, the simplex method. Its main steps are the following:

1. Select a basic feasible solution. This means geometrically that one of the corners (extreme points) of the feasible set X is chosen.

   Check whether the corner is optimal. If so, you have found the optimal solution and the algorithm terminates. Otherwise go to Step 2.

2. Choose a new basic feasible solution by replacing one of the basic variables by a nonbasic variables. The entering variable should be selected so as to improve the objective function value. Geometrically saying: move to an adjacent corner of X where a better objective function value is expected than it was previously.

This procedure is repeated until no improvement in objective function value can be made. When this happens the optimal solution has been found.

9.1 Simplex Method for Solving Normal LP Problems

For the sake of better understanding the simplex algorithm is first applied to a special LP problem which we call the *normal problem.* A normal problem is a maximizing problem with non-negative decision variables where every constraint is given by "$\leq$". When this problem is converted into standard form, we have to add a new u_i variable to the ith constraint. These new variables are called *slack variables.* The slack variables provide us with an initial basic solution of the problem, so we can start the simplex algorithm without any preparations.

When calculating manually the steps and operations of the algorithm are represented by the simplex tableau. This is actually the basis table of the system $Ax = b$, extended by a new row which contains the coefficients of the objective function at the beginning of the algorithm. This row will be called the *check row,* because its elements help in selecting the pivotal elements and identifying the end of the algorithm.

Let us consider the sandwich problem which is solved graphically in Chapter 8.1: A mother makes two types of sandwiches for the birthday party of her son. A salami sandwich consists of 3 g butter, 3 slices of egg and two slices of salami. A ham sandwich consists of 4 g butter, 2 slices of egg and a slice of ham. (Of course these ingredients are in between two slices of bread in both cases.) For making sandwiches 100 slices of salami, 40 slices of ham, 170 slices of egg and 220 g butter are available. There is plenty of bread, so no restriction is needed for bread. How many sandwiches of each type should be made to maximize the total number of sandwiches?

Denoting the variables by x_1 and x_2, the model of the problem is:

$$
\begin{aligned}
x_1 + x_2 &\longrightarrow \max \\
3x_1 + 4x_2 &\leq 220 \\
3x_1 + 2x_2 &\leq 170 \\
2x_1 &\leq 100 \\
x_2 &\leq 40 \\
x_1, x_2 &\geq 0
\end{aligned}
$$

This is a normal problem, since every constraint is given by $\leq$. We need 4 slack variables to transform the problem into standard form:

$$
\begin{aligned}
x_1 + x_2 &\longrightarrow \max \\
3x_1 + 4x_2 + u_1 &\leq 220 \\
3x_1 + 2x_2 + u_2 &\leq 170 \\
2x_1 + u_3 &\leq 100 \\
x_2 + u_4 &\leq 40 \\
x_1, x_2, u_1, u_2 &\geq 0
\end{aligned}
$$

In the initial simplex tableau the slack variables form the basis:

	x_1	x_2	b
u_1	3	4	220
u_2	3	2	170
u_3	2	0	100
u_4	0	$\boxed{1}$	40
$-z$	1	1	0

The check row is labelled $-z$, because the last element of this row will show the opposite of the objective function value. Initially this value is 0, because all basic variables have zero objective value coefficients. In other words, at the beginning the variables x_1, x_2 are nonbasic variables, their values equal zero, so the objective function value $x_1 + x_2$ is also zero. Geometrically speaking: we are in the origin when we start examining the corners of the feasible set. The initial solution of the system $Ax = b$ can be read from the right-hand side, namely: $u_1 = 220, u_2 = 170, u_3 = 100, u_4 = 40$.

The next step is replacing one of the basic variables by a nonbasic variable so as to improve the objective function value. For this purpose we have to choose a suitable pivotal element. There are three important rules for selecting a pivotal element in the simplex tableau:

- The pivotal element must be positive. (A negative pivotal element would imply non-feasible solution, because the entire row is divided by the pivotal, resulting in a negative number in the right-hand side.)
- Choose the pivotal element from the column where the check row contains a positive number. (This choice guarantees the improvement of the objective function value.)
- After deciding the column of the pivotal element, we have to select the element which wins the ratio test. This means that for every element of the column we form the ratio of the right-hand-side constant to the element itself, and choose the element with the smallest ratio. (With this choice we avoid appearing negative elements in column b.)

If there is more than one option, try to choose the version which promises the easiest calculations.

Rules of pivoting:

- write the reciprocal of the pivotal element p instead of p,
- the row of p is divided by p,
- the column of p is divided by $-p$,
- other elements are calculated the rectangle-rule: $x - \dfrac{ab}{p}$.

In our example, the check row allows the choice from both columns. If we decide on the first column, 2 must be the pivotal element, because the last is the smallest of the ratios $\frac{220}{3}, \frac{170}{3}, \frac{100}{2}$. In the second column 1 should be the winner of the ratio test: $\min\{\frac{220}{4}, \frac{170}{2}, \frac{40}{1}\} = \frac{40}{1}$. It is easier to calculate with 1, so we replace the variable u_4 by x_2. We use the rules for the calculation as previously discussed, extending the operations to the last row:

	x_1	u_4	b
u_1	$\boxed{3}$	-4	60
u_2	3	-2	90
u_3	2	0	100
x_2	0	1	40
$-z$	1	-1	-40

Note that if there is a 0 in the row of the pivotal element, the column of this 0 is not changed. Similarly, if there is a 0 in the column of the pivot element, the row of this 0 remains unchanged. (Explanation: when we apply the rectangle rule, 0 is to be subtracted from the number in question.)

The simplex tableau now shows a better solution then the initial one. We obtained the solution $x_1 = 0, x_2 = 40$, so 0 salami sandwiches and 40 ham sandwiches are made. (We are in the extreme point (0;40) of the polygon). However, the objective function value can be improved, because there is a positive number in the check row. We bring x_1 into the basis instead of u_1, because the first one is the smallest of the ratios $\frac{60}{3}$, $\frac{90}{3}$ and $\frac{100}{2}$:

	u_1	u_4	b
x_1	$\frac{1}{3}$	$-\frac{4}{3}$	20
u_2	-1	$\boxed{2}$	30
u_3	$-\frac{2}{3}$	$\frac{8}{3}$	60
x_2	0	1	40
$-z$	$-\frac{1}{3}$	$\frac{1}{3}$	-60

The present solution is $x_1 = 20, x_2 = 40$, the objective function value: $z = 60$. Since the check row contains again a positive number, the value z can be increased by replacing u_4 by u_2 :

	u_1	u_2	b
x_1	$-\frac{1}{3}$	$\frac{2}{3}$	40
u_4	$-\frac{1}{2}$	$\frac{1}{2}$	15
u_3	$-\frac{2}{3}$	$-\frac{4}{3}$	60
x_2	$\frac{1}{2}$	$-\frac{1}{2}$	25
$-z$	$-\frac{1}{6}$	$-\frac{1}{6}$	-65

This tableau is optimal, because there are no more positive elements in the check row. The algorithm terminates with the optimal solution $x_1 = 40, x_2 = 25$, the optimum is 65.

9.2 Further Examples of Solving Normal Problems by the Simplex Algorithm

The next problem contains three decision variables, so it cannot be represented in two-dimensional space.

Example 9.1 A plant manufactures three types of vehicle: automobiles, trucks and vans, on which the company makes a profit of $4000, $6000 and $3000, respectively, per vehicle. The plant has three main departments: parts, assembly and finishing. The next table shows how many hours a vehicle spends in the different departments, and what the weekly capacities in these departments are:

	automobile	truck	van	capacity
parts	50	40	30	240
assembly	40	30	20	200
finishing	20	40	10	160

How many of each type of vehicle should the company manufacture in order to maximize profit for a weekly period?

Solution: We introduce the following decision variables: $x_1 =$ be the number of automobiles manufactured, $x_2 =$ the number of trucks manufactured and $x_3 =$ the number of vans manufactured.

After dividing the profits by 1000, and the constraints by 10 we have the following problem:

$$4x_1 + 6x_2 + 3x_3 \longrightarrow \max$$
$$5x_1 + 4x_2 + 3x_3 \leq 24$$
$$4x_1 + 3x_2 + 2x_3 \leq 20$$
$$2x_1 + 4x_2 + x_3 \leq 16$$
$$x_1, x_2, x_3 \geq 0$$

We introduce three slack variables and the problem becomes:

$$4x_1 + 6x_2 + 3x_3 \longrightarrow \max$$
$$5x_1 + 4x_2 + 3x_3 + u_1 = 24$$
$$4x_1 + 3x_2 + 2x_3 + u_2 = 20$$
$$2x_1 + 4x_2 + x_3 + u_3 = 16$$
$$x_1, x_2, x_3, u_1, u_2, u_3 \geq 0$$

The initial simplex tableau shows the first solution: the plant manufactures 0 vehicles with 0 profit:

	x_1	x_2	x_3	b
u_1	5	4	3	24
u_2	4	3	2	20
u_3	2	$\boxed{4}$	1	16
$-z$	4	6	3	0

We choose the pivotal element from the second column, because variable x_2 promises the greatest profit:

	x_1	u_3	x_3	b
u_1	3	-1	$\boxed{2}$	8
u_2	$\frac{5}{2}$	$-\frac{3}{4}$	$\frac{5}{4}$	8
x_2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	4
$-z$	1	$-\frac{3}{2}$	$\frac{3}{2}$	-24

In the next step we replace u_1 by x_3:

	x_1	u_3	u_1	b
x_3	$\frac{3}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	4
u_2	$\frac{5}{8}$	$-\frac{1}{8}$	$-\frac{5}{8}$	3
x_2	$\frac{1}{8}$	$\frac{3}{8}$	$-\frac{1}{8}$	3
$-z$	$-\frac{5}{4}$	$-\frac{3}{4}$	$-\frac{3}{4}$	-30

The algorithm terminates, because there is no positive element in the check row. The optimal solution is: $x_1 = 0, x_2 = 3, x_3 = 4$. The optimum is 30 in the tableau, that is the maximum profit is \$30,000, which can be achieved by manufacturing 3 trucks and 4 vans in a week. The value of the slack variable u_2 equals 3 in the optimal tableau, meaning that there are 30 hours of unused capacity in assembly. The other two departments work at full capacity because the slack variables u_1 and u_2 are zeros.

Example 9.2 This problem has multiple optimal solutions. Now we examine how this case can be identified when applying the simplex algorithm.

$$2x_1 + 2x_2 \longrightarrow \max$$
$$x_1 + x_2 \leq 70$$
$$x_1 \leq 50$$
$$x_2 \leq 40$$
$$x_1 \geq 0, x_2 \geq 0$$

The steps of the algorithm:

	x_1	x_2	b
u_1	1	1	70
u_2	1	0	50
u_3	0	$\boxed{1}$	40
$-z$	2	2	0

	x_1	u_3	b
u_1	$\boxed{1}$	-1	30
u_2	1	0	50
x_2	0	1	40
$-z$	2	-2	-80

	u_1	u_3	b
x_1	1	-1	30
u_2	-1	1	20
x_2	0	1	40
$-z$	-2	0	-140

There is no more positive number in the check row, the solution $x_1 = 30$, $x_2 = 40$ is optimal and the optimum is 140. However the variable u_3 has a zero check row coefficient, indicating that the objective function value would remain unchanged if u_3 was brought into the basis. In order to check it we replace u_2 by u_3:

	u_1	u_2	b
x_1	0	1	50
u_3	-1	1	20
x_2	1	-1	20
$-z$	-2	0	-140

We have a new optimal solution $x_1 = 50, x_2 = 20$ with the same optimum as before. Actually there are infinitely many optimal solutions: any point on the line segment joining the two basic optimal solutions is optimal. Formally, the optimal solutions are the points $\lambda(30; 40) + (1 - \lambda)(50; 20)$ where $\lambda \in [0; 1]$.

Example 9.3 The following problem does not have an optimal solution, because the objective function is not bounded on the feasible set.

$$
\begin{gathered}
3x_1 + 2x_2 \longrightarrow \max \\
\tfrac{1}{2}x_1 - x_2 \leq 1 \\
2x_1 - x_2 \leq 7 \\
x_1 \geq 0, x_2 \geq 0
\end{gathered}
$$

Solution by the simplex algorithm:

	x_1	x_2	b
u_1	$\boxed{\frac{1}{2}}$	-1	1
u_2	2	-1	7
$-z$	3	2	0

	u_1	x_2	b
x_1	2	-2	2
u_2	-4	$\boxed{3}$	3
$-z$	-6	8	-6

	u_1	u_2	b
x_1	$-\frac{2}{3}$	$\frac{2}{3}$	4
x_2	$-\frac{4}{3}$	$\frac{1}{3}$	1
$-z$	$\frac{14}{3}$	$-\frac{8}{3}$	-14

From the last table it can be seen that u_1 should enter the basis next. The problem is that all of the coefficients are negative in u_1's column, so it is not possible to choose a pivotal element. A positive check row coefficient whose column entries are all non-positive indicates that the objective function can have arbitrarily large value and there is no bounded optimal solution of the problem.

9.3 The Two-Phase Simplex Method

Until now we used the simplex algorithm for solving normal problems where all constraints were of the $\leq$ type. In this case the introduced slack variables formed an initial basis, and the simplex method can be started conveniently. With constraints $=$ or $\geq$ the procedure is different, because an initial basic solution is not known as it was before. The first phase of the method aims to create an initial feasible basis so that the simplex algorithm can be used. After converting the problem into standard form, a new artificial variable is added to the left-hand side of each constraint equation which was of the $=$ or $\geq$ type. These new variables are denoted by u_i^* and they are also constrained to be non-negative. Then a new objective function is introduced as a sum of the artificial variables. This creates a new problem in which the secondary objective function is to be minimized. If the optimal solution value of the new problem is greater than zero, the original problem does not have a feasible solution. Otherwise we obtain a basic and feasible solution of the original problem because all the artificial variables have value zero. The obtained solution can be used as a starting solution for further iterations of the simplex method. The second phase consists of these iterations.

Example 9.4 Solve the following general LP problem using the two-phase simplex method.

$$\begin{aligned}
2x_1 + 6x_2 + 2x_3 &\longrightarrow \max \\
x_2 + x_3 &\leq 160 \\
-x_1 + x_2 + x_3 &= 100 \\
x_1 + x_2 + x_3 &\geq 180 \\
x_1, x_2, x_3 &\geq 0
\end{aligned}$$

Solution: In order to convert it into standard form, a slack variable u_1 is added to the left-hand side of the first constraint and an excess variable v_3 is subtracted from the third one. In addition two artificial variables are introduced and they are added to the left-hand side of the last two constraints.

The new problem can be formulated as:

$$\begin{aligned} 2x_1 + 6x_2 + 2x_3 &\longrightarrow \max \\ x_2 + x_3 + u_1 &= 160 \\ -x_1 + x_2 + x_3 + u_2^* &= 100 \\ x_1 + x_2 + x_3 - v_3 + u_3^* &= 180 \\ x_1, x_2, x_3, u_1, u_2^*, u_3^*, v_3 &\geq 0 \end{aligned}$$

The initial simplex tableau contains an extra row, labelled by z^*. The elements of this row are obtained by adding the rows of the artificial variables. First we have to find the optimum of the secondary objective function.

	x_1	x_2	x_3	v_3	b
u_1	0	1	1	0	160
u_2^*	-1	$\boxed{1}$	1	0	100
u_3^*	1	1	1	-1	180
$-z$	2	6	2	0	0
z^*	0	2	2	-1	280

We do not need artificial variables in the basis, so variable x_2 is entered and u_2^* is removed:

	x_1	u_2^*	x_3	v_3	b
u_1	1	-1	0	0	60
x_2	-1	1	1	0	100
u_3^*	$\boxed{2}$	-1	0	-1	80
$-z$	8	-6	-4	0	-600
z^*	2	-2	0	-1	80

Now replace u_3^* by x_1:

	u_3^*	u_2^*	x_3	v_3	b
u_1	$-\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	20
x_2	$\frac{1}{2}$	$\frac{1}{2}$	1	$-\frac{1}{2}$	140
x_1	$\frac{1}{2}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$	40
$-z$	-4	-2	-4	4	-920
z^*	-1	-1	0	0	0

The first phase is finished because the artificial variables are removed from the basis (their value is 0) and z^* is also 0. The simplex tableau contains a feasible solution of the original problem as follows: $x_1 = 40, x_2 = 140, x_3 = 0$, and the objective function value is 920. We can leave the last row and the columns belonging to u_3^* és u_2^* from the tableau:

	x_3	v_3	b
u_1	0	$\boxed{\frac{1}{2}}$	20
x_2	1	$-\frac{1}{2}$	140
x_1	0	$-\frac{1}{2}$	40
$-z$	-4	4	-920

The check row contains a positive number, indicating that a program can be improved by bringing v_3 into the basis:

	x_3	u_1	b
v_3	0	2	40
x_2	1	1	160
x_1	0	1	60
$-z$	-4	-8	-1080

Every check row coefficient is negative, the second phase is finished. The optimal solution of the problem: $x_1 = 60, x_2 = 160, x_3 = 0$ and the optimum is 1080.

10 Duality and Sensitivity Analysis

In the previous chapters we studied a lot of LP problems, and it became obvious that the difficulty of their solution mainly depends on the number and the type of their constraints. It would be useful if we could reduce the number of constraints in relatively large problems. This is often possible by constructing a new LP problem, the so-called *dual problem*. When we talk about the dual problem, the original problem is referred as the *primal*. We should decide which problem can be solved more easily, because the final simplex tableau contains the solution of both problems.

10.1 Dual of a Normal Problem

First we construct the dual pair of a normal problem and at the same time give an interpretation of the dual problem. Recall that a normal problem often models the following situation:

A company produces n types of products using m types of sources which are available in a restricted amount. The technological matrix is given, its element a_{ij} is the amount of the source i which is needed for producing a unit of product j. The available amount (capacity) of source i is b_i, the profit from a unit of product j is c_j. Determine the amount of each type of products so as to maximize the company's profit.

Denoting by x_i the amount of product i to be produced, we obtain the following problem:

$$\begin{aligned} c_1x_1 + c_2x_2 + \ldots + c_nx_n &\longrightarrow \max \\ a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &\leq b_1 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &\leq b_m \\ x_1, x_2, \ldots, x_n &\geq 0 \end{aligned}$$

Using vector-matrix denotations the model can be set more briefly:

$$\begin{aligned} c^T x &\longrightarrow \max \\ Ax &\leq b \\ x &\geq 0 \end{aligned}$$

Suppose that a corporation is considering hiring the capacities of the company. Let y_i be the hiring rate of a unit of source i, $i = (1, 2, \ldots, m)$. Of course all hiring costs have to be nonnegative. It is not worth hiring the sources if the revenue is less than the present profit, so the following inequalities must be valid:

$$\begin{aligned} a_{11}y_1 + a_{21}y_2 + \ldots + a_{m1}y_m &\geq c_1 \\ &\vdots \\ a_{1n}y_1 + a_{2n}y_2 + \ldots + a_{mn}y_m &\geq c_n \end{aligned}$$

However, the corporation wishes to minimize the total hiring cost it has to pay, so it wants to keep the amount $b_1y_1 + b_2y_2 + \ldots b_my_m$ as low as possible.

The problem can be summarized as follows:

$$\begin{aligned} b^T y &\longrightarrow \min \\ A^T y &\geq c \\ y &\geq 0 \end{aligned}$$

and this problem is called the *dual problem* of the original normal problem.

Theorem 10.1 (Strong Duality Theorem) *If either the primal or the dual problem has an optimal solution, then so does the other, and the optimums coincide.*

Example 10.1 A company produces 3 types of products (T_1, T_2, T_3) by using 3 types of employees (E_1, E_2, E_3). The next table shows the needs of each employee for producing a unit of each product, the available amounts of employees and the unit prices of the products.

	T_1	T_2	T_3	capacity
E_1	1	1	2	5
E_2	2	1	3	11
E_3	2	3	2	12
price	7	9	5	

Give the model of the problem and find the combination of products so as to maximize profit. Determine the dual of the problem and using the optimal simplex tableau of the primal give the optimal solution to the dual problem.

Solution: We have a normal problem, its model is

$$\begin{aligned} 7x_1 + 9x_2 + 5x_3 &\longrightarrow \max \\ x_1 + x_2 + 2x_3 &\leq 5 \\ 2x_1 + x_2 + 3x_3 &\leq 11 \\ 2x_1 + 3x_2 + 2x_3 &\leq 12 \\ x_1, x_2, x_3 &\geq 0 \end{aligned}$$

There are 3 constraints in the primal problem which indicate 3 variables in the dual. The coefficients in constraint i of the dual are given by the coefficients of the variable x_i in the primal constraints. The dual right-hand-side constants are equal to the primal objective function coefficients, and vice versa. The dual is a minimizing problem and all constraints are of $\geq$ type. Dual variables are also nonnegative. Summarizing these observations the dual problem is expressed as

$$\begin{aligned} 5y_1 + 11y_2 + 12y_3 &\longrightarrow \min \\ y_1 + 2y_2 + 2y_3 &\geq 7 \\ y_1 + y_2 + 3y_3 &\geq 9 \\ 2y_1 + 3y_2 + 2y_3 &\geq 5 \\ y_1, y_2, y_3 &\geq 0 \end{aligned}$$

We solve the primal problem by the simplex method:

	x_1	x_2	x_3	b
u_1	$\boxed{1}$	1	2	5
u_2	2	1	3	11
u_3	2	3	2	12
$-z$	7	9	5	0

	u_1	x_2	x_3	b
x_1	1	1	2	5
u_2	-2	-1	-1	1
u_3	-2	$\boxed{1}$	-2	2
$-z$	-7	2	-9	-35

	u_1	u_3	x_3	b
x_1	3	-1	4	3
u_2	-4	1	-3	3
x_2	-2	1	-2	2
$-z$	-3	-2	-5	-39

The final tableau is optimal, the primal optimal solution is

$$x_1 = 3,\ x_2 = 2,\ x_3 = 0,\ z_{\max} = 39.$$

In this case the dual problem can be solved by the two-phase method, which would need more effort. However, this procedure can be avoided, because the dual optimal solution can be read from the final simplex tableau of the primal. The optimums of the primal and dual are equal, so the minimum of the dual objective function is 39. The dual optimal solution can be determined by the slack variables of the primal: y_i is the opposite of the check row coefficient belonging to u_i. If a variable u_i has remained in the basis, the corresponding y_i has the value of zero in the dual optimal solution. So the optimal solution of the dual problem is $y_1 = 3, y_2 = 0, y_3 = 2$, and the optimum is 39, as mentioned before.

10.2 Dual of a General LP Problem

In general LP problems the constraints can be given by equalities or inequalities and the signs of the variables can be restricted in various ways. Using the following table we can construct the corresponding dual problem to an arbitrary LP problem.

	maximizing problem	minimizing problem	
variables	≥ 0	$\geq$	constraints
	unrestricted	$=$	
	≤ 0	$\leq$	
constraints	$\leq$	≥ 0	variables
	$=$	unrestricted	
	$\geq$	≤ 0	

Some hints to the usage of the table:
The dual of a maximizing problem is a minimizing problem and vice versa.

Each primal constraint corresponds to a dual variable and each primal variable corresponds to a dual constraint. The type of the constraints and the restrictions for the variables are strongly connected. For instance if a constraint in a maximizing problem is given by $\leq$, then the corresponding dual variable must be nonnegative. If a variable in a maximizing problem is unrestricted, then the corresponding constraint is an equality, and so on.

Example 10.2 Construct the dual of the following problem:

$$\begin{aligned} 2x_1 - x_2 + 5x_3 - 4x_4 &\longrightarrow \max \\ x_1 + x_2 - x_3 - 2x_4 &= 60 \\ x_1 + x_2 - 2x_3 + 3x_4 &\geq 100 \\ x_1 + x_3 + 2x_4 &\leq 80 \\ x_1 \geq 0, x_2 \leq 0, x_3 &\text{ unrestricted}, x_4 \geq 0 \end{aligned}$$

Solution: The dual problem is a minimizing problem with 3 variables and 4 constraints, because the primal has 3 constraints and 4 variables.

- $x_1 \geq 0$ and $x_4 \geq 0$ is prescribed in the primal, so the first and fourth dual constraints are of $\geq$ type,
- $x_2 \leq 0$ in the primal, the second dual constraint is of $\leq$ type,
- the primal variable x_3 is unrestricted, implying an equality in the third dual constraint,
- the first primal constraint is an equality, which is why the dual variable y_1 is unrestricted,
- the second primal constraint is given by $\geq$ implying $y_2 \leq 0$,
- because of the $\leq$ in the third primal constraint the dual variable y_3 must be nonnegative.

The obtained dual problem is:

$$\begin{aligned} 60y_1 + 100y_2 + 80y_3 &\longrightarrow \min \\ y_1 + y_2 + y_3 &\geq 2 \\ y_1 + y_2 &\leq -1 \\ -y_1 - 2y_2 + y_3 &= 5 \\ -2y_1 + 3y_2 + 2y_3 &\geq -4 \\ y_1 \text{ unrestricted}, y_2 &\leq 0, y_3 \geq 0 \end{aligned}$$

If a primal problem is solved by the two-phase simplex method, we usually remove the u^* columns before starting the second phase. However, when a dual optimal solution should be determined, it is better to save these columns.

Example 10.3 Give the primal and dual optimal solutions to the problem

$$\begin{aligned} 5x_1 - 2x_2 + 8x_3 &\longrightarrow \max \\ x_1 + 2x_2 + x_3 &\leq 50 \\ x_1 + x_2 &= 40 \\ x_1 + x_2 + 2x_3 &\geq 20 \\ x_1, x_2, x_3 &\geq 0 \end{aligned}$$

Solution: Only the starting and the final tableau are presented, the others are left to the reader for the sake of practice:

	x_1	x_2	x_3	v_3	b
u_1	1	2	1	0	50
u_2^*	1	1	0	0	40
u_3^*	1	1	2	-1	20
$-z$	5	-2	8	0	0
z^*	2	2	2	-1	60

	x_2	u_1	b	u_2^*	u_3^*
x_3	1	1	10	-1	0
v_3	2	2	40	-1	-1
x_1	1	0	40	1	0
-z	-15	-8	-280	3	0

The columns of the variables u^* are copied to the right-hand side of the tableau. The optimal solution to the primal:

$$x_1 = 40,\ x_2 = 0,\ x_3 = 10,\ z_{\max} = 280.$$

The dual problem is formulated as

$$\begin{aligned} 50y_1 + 40y_2 + 20y_3 &\longrightarrow \min \\ y_1 + y_2 + y_3 &\geq 5 \\ 2y_1 + y_2 + y_3 &\geq -2 \\ y_1 + 2y_3 &\geq 8 \\ y_1 \geq 0,\ y_2 \text{ unrestricted},\ y_3 &\leq 0 \end{aligned}$$

The solution to the dual can be read from the optimal tableau of the primal. The variable y_2 was unrestricted, in this case its sign in the optimal solution is determined so as the dual objective function has the optimum equalling the optimum of the primal problem. This remark explains the dual optimal solution, which is

$$y_1 = 8,\ y_2 = -3,\ y_3 = 0 \text{ and } z_{\min} = 280.$$

10.3 Sensitivity Analysis

Duality can be used to answer questions related to the sensitivity of the solution. We would often like to know what happens with the solution if the parameters of the problem are changed. Obviously these questions can be answered by solving the problem from the beginning, but it is not necessary, at least in case of minor changes. We will make the sensitivity analysis based on the dual variables without recalculating all simplex tableaus.

Sensitivity analysis considers the range within which some coefficients of the problem can vary without affecting the optimal solution. It can begin after an LP problem has been solved, and deals with the impact of changing one parameter at a time.

10.3.1 Sensitivity Analysis for the Right-Hand-Side Constants

First we study the case when one of the right-hand-side constants has been modified.

Example 10.4 Consider again Example 10.1 and make sensitivity analysis for each right-hand-side constant.

$$\begin{aligned} 7x_1 + 9x_2 + 5x_3 &\longrightarrow \max \\ x_1 + x_2 + 2x_3 &\leq 5 \\ 2x_1 + x_2 + 3x_3 &\leq 11 \\ 2x_1 + 3x_2 + 2x_3 &\leq 12 \\ x_1, x_2, x_3 &\geq 0 \end{aligned}$$

Solution: Recall that the optimal solution of the problem is $x_1 = 3,\ x_2 = 2,\ x_3 = 0,\ z_{\max} = 39$.

Suppose that the right-hand side value of the first constraint is altered from 5 to $5 + \lambda$. For what value of λ will the present solution remain feasible and optimal? This means that we study the effect of the modification without changing the structure of the solution: x_1 and x_2 should remain the basic variables. In order to answer the question we start the simplex algorithm with $5 + \lambda$ instead of 5:

	x_1	x_2	x_3	b
u_1	$\boxed{1}$	1	2	$5+\lambda$
u_2	2	1	3	11
u_3	2	3	2	12
$-z$	7	9	5	0

	u_1	x_2	x_3	b
x_1	1	1	2	$5+\lambda$
u_2	-2	-1	-1	$1-2\lambda$
u_3	-2	$\boxed{1}$	-2	$2-2\lambda$
$-z$	-7	2	-9	$-35-7\lambda$

	u_1	u_3	x_3	b
x_1	3	-1	4	$3+3\lambda$
u_2	-4	1	-3	$3-4\lambda$
x_2	-2	1	-2	$2-2\lambda$
$-z$	-3	-2	-5	$-39-3\lambda$

Comparing the final tableau with the one obtained in solving Example 10.1, we can observe that only the last column has been modified. For the optimality of this tableau the right-hand side coefficients should be nonnegative, that is, the following inequalities are to be satisfied at the same time:

$$\begin{matrix} 3+3\lambda \geq 0 \\ 3-4\lambda \geq 0 \\ 2-2\lambda \geq 0 \end{matrix} \text{, which means } \begin{matrix} \lambda \geq -1 \\ \lambda \leq \frac{3}{4} \\ \lambda \leq 1 \end{matrix} \quad \text{implying } -1 \leq \lambda \leq \tfrac{3}{4}.$$

We can read the new solution from the final simplex tableau: if $\lambda \in [-1; \frac{3}{4}]$, then the optimal solution of the modified problem is

$$x_1 = 3+3\lambda,\ x_2 = 2-2\lambda,\ x_3 = 0,\ z_{\max} = 39+3\lambda.$$

For example if the available amount of the first labour is increased by 0.5, then we obtain the following optimal solution: $x_1 = 4.5,\ x_2 = 1,\ x_3 = 0,$ $z_{\max} = 40.5$.

As it was mentioned earlier, it is not necessary to do the calculations from the beginning. Suppose that the right-hand-side constant of constraint i is changed from b_i to $b_i + \lambda$. Solve the system of inequalities $b + \lambda u_i \geq 0$, where b and u_i are the corresponding column vectors in the final simplex tableau. The solution gives limits for parameter λ within which the structure of the original solution remains unchanged. The value x_i in the new solution can be determined in the following way: it equals the corresponding component of $b + \lambda u_i$ if x_i is a basic variable, otherwise x_i remains zero.

For example, suppose that the right-hand side value of the third constraint changes to $12+\lambda$ from 12. We have three inequalities to solve:

$$\begin{matrix} 3-\lambda \geq 0 \\ 3+\lambda \geq 0 \\ 2+\lambda \geq 0 \end{matrix} \longrightarrow \begin{matrix} \lambda \leq 3 \\ \lambda \geq -3 \\ \lambda \geq -2 \end{matrix} \longrightarrow -2 \leq \lambda \leq 3.$$

We obtained that if $\lambda \in [-2; 3]$, then the optimal solution of the problem is $x_1 = 3-\lambda,\ x_2 = 2+\lambda,\ x_3 = 0,\ z_{\max} = 39+2\lambda$.

Observe the different range of changes in the objective function value when modifying the first and the third capacities. The change depends on the check row coefficient corresponding to the variable u_i which is actually the optimal solution of the dual problem. In economics this important parameter is referred as the *shadow price.* The shadow price of a constraint is the amount that the objective function value would change if the named constraint changed by one unit. The shadow price is valid up to the allowable increase or decrease in the constraint.

We have not dealt with the case of the second constraint yet. Suppose that the right-hand side constant of the second constraint is changed from 11 to $11 + \lambda$. The variable u_2 is in the basis, its coefficients can be represented by the unit vector $(0, 1, 0)^T$ which is not presented in the simplex tableau. So we have only one inequality for λ, namely $3 + \lambda \geq 0$, yielding $\lambda \geq -3$. This means that the capacity of the second source can be decreased at most 3 units and can be increased by an arbitrary amount. The shadow price is 0, any change in capacity leaves the optimum unchanged, the solution remains: $x_1 = 3,\ x_2 = 2,\ x_3 = 0,\ z_{\max} = 39$.

10.3.2 Sensitivity Analysis for the Objective Function Coefficients

Suppose that one of the objective function coefficients is changed, for example, the price of a product is decreased. What is the range of this change for which the optimal solution of the original problem will remain optimal? Let us consider again Example 10.1:

$$\begin{aligned} 7x_1 + 9x_2 + 5x_3 &\longrightarrow \max \\ x_1 + x_2 + 2x_3 &\leq 5 \\ 2x_1 + x_2 + 3x_3 &\leq 11 \\ 2x_1 + 3x_2 + 2x_3 &\leq 12 \\ x_1, x_2, x_3 &\geq 0 \end{aligned}$$

Suppose that the first coefficient in the objective function changes from 7 to $7 + \lambda$. Follow the impact of this modification by recalculating the simplex tableaus with $7 + \lambda$ instead of 7:

	x_1	x_2	x_3	b
u_1	$\boxed{1}$	1	2	5
u_2	2	1	3	11
u_3	2	3	2	12
$-z$	$7+\lambda$	9	5	0

	u_1	x_2	x_3	b
x_1	1	1	2	5
u_2	-2	-1	-1	1
u_3	-2	$\boxed{1}$	-2	2
$-z$	$-7-\lambda$	$2-\lambda$	$-9-2\lambda$	$-35-5\lambda$

	u_1	u_3	x_3	b
x_1	3	-1	4	3
u_2	-4	1	-3	3
x_2	-2	1	-2	2
$-z$	$-3-3\lambda$	$-2+\lambda$	$-5-4\lambda$	$-39-3\lambda$

Only the check row coefficients have been changed which is not surprising, because the feasible region is unaltered. The original optimal solution remains optimal if the check row has no positive elements. The inequalities

$$-3-3\lambda \leq 0,\ -2+\lambda \leq 0,\ -5-4\lambda \leq 0$$

imply $-1 \leq \lambda \leq 2$. This means that the unit price of product T_1 can be moved between 6 and 9 without changing the optimal solution $x_1 = 3, x_2 = 2, x_3 = 0$. Obviously the optimal value of the objective function should be modified: it changes from 39 to $39 + 3\lambda$.

In general, if coefficient i of the objective function is changed from c_i to $c_i + \lambda$ then solve the system of inequalities $-z - \lambda x_i \leq 0$, where $-z$ is the vector of the check row coefficients and x_i denotes the row vector of variable x_i in the final simplex tableau.

For instance, if the second coefficient of the objective function is changed from 9 to $9 + \lambda$, we have the inequalities

$$-3+2\lambda \leq 0,\ -2-\lambda \leq 0 \ \text{ and } \ -5+2\lambda \leq 0.$$

The solution of this system is $-2 \leq \lambda \leq \frac{3}{2}$. It means, that if the price of T_2 is changed by λ, where $\lambda \in [-2; \frac{3}{2}]$, then the optimal solution of the problem remains $x_1 = 3,\ x_2 = 2,\ x_3 = 0$ and the new optimum changes to $39 + 2\lambda$.

If the third coefficient of the objective function is changed, then the calculations are even simpler, because the coefficients belonging to x_3 form a unit vector, because x_3 is a nonbasic variable. Only the condition $-5-\lambda \leq 0$ should be satisfied, that is, $\lambda \geq -5$. This means that the price of T_3 can be an arbitrary nonnegative value. To explain this result, we need to take into consideration that the structure of the solution must remain unchanged, i.e. product T_3 is not involved in production, so its price has no impact on the optimum.

11 Applications of Linear Programming

11.1 Linear-Fractional Programming

Linear-fractional programming (LFP) can be considered as a generalization of linear programming. The objective function of a linear-fractional programming problem is not linear but a ratio of two linear functions. A linear program can be regarded as a special case of a linear-fractional program in which the denominator is the constant function one. An LFP problem can be transformed into an LP problem, which then can be solved by the simplex method.

Consider a factory which produces x_i units of its product i. The cost of producing a unit of product i is denoted by d_i, and the price for which it can be sold is c_i. The problem is to find the highest ratio of income to cost showing the efficiency of the production. In order to formulate the problem let $x = (x_1, x_2, ..., x_n)^T$, $c = (c_1, c_2, ..., c_n)^T$ and $d = (d_1, d_2, ..., d_n)^T$. The standard form of the linear-fractional programming problem is defined by

$$\begin{aligned} \frac{c^T x + c_0}{d^T x + d_0} &\longrightarrow \max \\ Ax &= b \\ x &\geq 0 \end{aligned}$$

We will assume that the set of the feasible solutions is non-empty and bounded. In addition the denominator of the objective function is assumed to be positive in the feasible region (the coefficients of the denominator can be multiplied by -1, if needed). These assumptions guarantee the existence of an optimal solution of the LFP problem.

In order to convert the LFP problem into an LP problem the nominator and the denominator of the objective function and all constraints should be multiplied by a positive unknown t. After this transformation we have the following problem:

$$\begin{aligned} \frac{c^T xt + c_0 t}{d^T xt + d_0 t} &\longrightarrow \max \\ Axt - bt &= 0 \\ x \geq 0, t &\geq 0 \end{aligned}$$

Next, new variables are introduced by the definition $y = xt$, and the denominator of the objective function is required to be 1. With this restriction it is enough to optimize the nominator of the objective function. The obtained problem is a linear programming problem as follows:

$$\begin{aligned} c^T y + c_0 t &\longrightarrow \max \\ Ay - bt &= 0 \\ d^T y + d_0 t &= 1 \\ y \geq 0, t &\geq 0 \end{aligned}$$

Example 11.1 Solve the linear-fractional programming problem given by

$$\begin{aligned} \frac{2x_1 + x_2 + x_3 - 1}{x_1 + x_2 + 2x_3 + 1} &\longrightarrow \max \\ 2x_1 + 2x_2 &\leq 6 \\ x_1 + x_2 + x_3 &\leq 4 \\ x_1, x_2, x_3 &\geq 0 \end{aligned}$$

Solution: The nominator and the denominator of the objective function and both constraints are multiplied by t. New variables are defined by $y_i = x_i t$ $(i = 1, 2, 3)$. The denominator of the objective function is restricted to be 1. Then the problem is:

$$\begin{aligned} 2y_1 + y_2 + y_3 - t &\longrightarrow \max \\ 2y_1 + 2y_2 \qquad - 6t &\leq 0 \\ y_1 + y_2 + y_3 - 4t &\leq 0 \\ y_1 + y_2 + 2y_3 + t &= 1 \\ y_1, y_2, y_3, t &\geq 0 \end{aligned}$$

This LP problem can be solved by the two-phase simplex method, because there is an equality between the constraints. The tableaus of the first phase:

	y_1	y_2	y_3	t	b
u_1	2	2	0	-6	0
u_2	1	1	1	-4	0
u_3^*	1	1	2	$\boxed{1}$	1
$-z$	2	1	1	-1	0
z^*	1	1	2	1	1

	y_1	y_2	y_3	u_3^*	b
u_1	8	8	12	6	6
u_2	5	5	9	4	4
t	1	1	2	1	1
$-z$	3	2	3	1	1
z^*	0	0	0	-1	0

As the secondary objective function became zero, the first phase is finished; we found an initial solution. The last row and the column of u_3^* are removed, and we can start the second phase:

	y_1	y_2	y_3	
u_1	$\boxed{8}$	8	12	6
u_2	5	5	9	4
t	1	1	2	1
$-z$	3	2	3	1

	u_1	y_2	y_3	
y_1	$\frac{1}{8}$	1	$\frac{3}{2}$	$\frac{3}{4}$
u_2	$-\frac{5}{8}$	0	$\frac{3}{2}$	$\frac{1}{4}$
t	$-\frac{1}{8}$	0	$\frac{1}{2}$	$\frac{1}{4}$
$-z$	$-\frac{3}{8}$	-1	$-\frac{3}{2}$	$-\frac{5}{4}$

All check row elements are negative, the table contains an optimal solution for the transformed problem:

$$y_1 = \frac{3}{4},\ y_2 = 0,\ y_3 = 0,\ t = \frac{1}{4} \text{ and } z_{\max} = \frac{5}{4}.$$

The optimal solution of the original LFP problem:

$$x_1 = \frac{y_1}{t} = 3,\ x_2 = \frac{y_2}{t} = 0,\ x_3 = \frac{y_3}{t} = 0.$$

The optimal value of the objective function is $\frac{5}{4}$, equalling to the optimum of the LP problem.

11.2 The Transportation Problem

The transportation problem is a special linear programming problem. Because of its structure there are more efficient methods for solving it than the simplex algorithm. Among these techniques the so-called distribution method will be discussed here.

The transportation problem can be formulated as follows:
In a supply system there are m factories which must supply the needs of n warehouses. The production capacity of the factories is limited to the amounts $a_1, a_2, ..., a_m$, and the demands of the warehouses are given by $b_1, b_2, ..., b_{n.}$ The unit cost of shipping one item from factory i to warehouse j is also known, and denoted by c_{ij}. Shipping cost is linear, which means that if x unit is shipped from factory i to warehouse j then the cost is $c_{ij} \cdot x$. In addition, we suppose that total supply and total demand is balanced, that is $\sum\limits_{i=1}^{m} a_i = \sum\limits_{j=1}^{n} b_j$ (the situation when this equality is not true will be later discussed). The problem is to find the minimum cost schedule that satisfies the production and demand constraints.

Let x_{ij} = the number of units shipped from factory i to warehouse j.

Then the problem is:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij} \to \min$$

$$\sum_{j=1}^{n} x_{ij} = a_i \quad i = 1, 2, ..., m$$

$$\sum_{i=1}^{m} x_{ij} = b_j \quad j = 1, 2, ..., n$$

$$x_{ij} \geq 0 \quad i = 1, 2, ..., m \;\; j = 1, 2, ..., n$$

It was mentioned earlier that instead of a simplex algorithm the problem will be solved by the distribution method, which is based on the variables of the dual problem. For the sake of better understanding, let us take the dual of the transportation problem.

We associate a dual variable u_i with each of the first m constraints and a dual variable v_j, with each of the next n constraints. Because of the special structure of the primal problem the left-hand side of the dual constraints is a sum of type $u_i + v_j$. The relation of each constraint is of the $\leq$ type, since each primal variable is restricted to be nonnegative. There are only equalities in the primal problem, implying no restrictions for the sign of dual variables. The dual problem is:

$$\sum_{i=1}^{m} a_i u_i + \sum_{j=1}^{n} b_j v_j \to \max$$

$$u_i + v_j \leq c_{ij} \quad i = 1, 2, ..., m \;\; j = 1, 2, ..., n$$

Suppose that we are solving the transportation problem as a regular LP problem using the simplex method. When we have a feasible basic solution of the primal problem, each of the basic variables is positive. This implies the equality $u_i + v_j = c_{ij}$ for each basic variable x_{ij} , since the corresponding slackness variable must be zero in the dual problem. This creates $m + n - 1$ equations in $m + n$ unknowns, which system can be solved by assigning an arbitrary value to one of the unknowns. (We usually set u_1 to zero.) After determining the values u_i and v_j we calculate the differences $\overline{c_{kl}} = c_{kl} - u_k - v_l$ for all nonbasic variables. The value $\overline{c_{kl}}$ shows to what extent the objective function would change if the variable x_{kl} entered the basis. If all $\overline{c_{kl}} \geq 0$, the current basic feasible solution is optimal. If at least one $\overline{c_{kl}} < 0$, we must select the x_{kl} which has the most negative value of $\overline{c_{kl}}$. This variable is entered in the basis, then we test the optimality again and continue the procedure until the optimum is reached.

The following theorem assures that a transportation problem always has an optimal solution.

Theorem 11.1 *In the transportation problem the primal and dual problem both have optimal solutions, and the minimum of the primal objective function is equal to the maximum of the dual objective function.*

11.2.1 Example of a Transportation Problem

Consider a supply system comprising three factories and three warehouses. The capacities of the factories F_1, F_2, F_3 are 20, 15, 10 units, respectively; the needs of the warehouses W_1, W_2, W_3 are 5, 20 and 20, respectively. The problem is balanced, because total supply and total demand is equal. Shipping costs per unit are given by the *transportation tableau* as follows:

		Warehouses			Supply
		W_1	W_2	W_3	
	F_1	0.9	1	1	20
Factories	F_2	1	1.4	0.8	15
	F_3	1.3	1	0.8	10
Demand		5	20	20	

If x_{ij} denotes the number of units transported from factory F_i to warehouse W_j, $(i, j = 1, 2, 3)$, then the regular LP model of the problem is:

$$0.9x_{11} + x_{12} + x_{13} + x_{21} + 1.4x_{22} + 0.8x_{23} + 1.3x_{31} + x_{32} + 0.8x_{33} \to \min$$
$$x_{11} + x_{12} + x_{13} = 20$$
$$x_{21} + x_{22} + x_{23} = 15$$
$$x_{31} + x_{32} + x_{33} = 10$$
$$x_{11} + x_{21} + x_{31} = 5$$
$$x_{12} + x_{22} + x_{32} = 20$$
$$x_{13} + x_{23} + x_{33} = 20$$
$$x_{ij} \geq 0, \quad i, j = 1, 2, 3.$$

The main points of the solution method are similar to the simplex algorithm: starting from an initial feasible solution the solution is improved step by step until the optimal solution is achieved. In order to find an initial solution we use the *northwest corner method.* This method starts by allocating as much as possible in the northwest (upper left) corner of the tableau. In our example it means that 5 units can be tied in the cell (1;1) which corresponds to factory F_1 and warehouse W_1. So 15 units left in factory F_1 and the demand of warehouse W_1 is satisfied. This means that column 1 can be removed from consideration and cell (1;2) becomes the new northwest corner. The maximum that can be allocated to this cell is 15, all that remained in F_1. As there are no more units in factory F_1, row 1 is removed from consideration and cell (2;2) becomes the new northwest corner. 5 units can be tied to this cell, then 10 units to its right-hand neighbor and finally 10 units are left for cell (3;3). The obtained initial feasible solution is the following:

	W_1	W_2	W_3	
F_1	5	15	0	20
F_2	0	5	10	15
F_3	0	0	10	10
	5	20	20	

The cells with positive amounts indicate the basic variables x_{ij} of the initial solution. We associate variables u_i to each row i and v_j to each column j. For each basic cell $(i;j)$ set the equality $u_i + v_j = c_{ij}$:

$$\begin{aligned}
u_1 + v_1 &= 0.9 \\
u_1 + v_2 &= 1 \\
u_2 + v_2 &= 1.4 \\
u_2 + v_3 &= 0.8 \\
u_3 + v_3 &= 0.8
\end{aligned}$$

Let $u_1 = 0$, and determine the values of the remaining dual variables. The result is $v_1 = 0.9,\ v_2 = 1,\ u_2 = 0.4,\ v_3 = 0.4$ and $u_3 = 0.4$.

We now calculate the differences $\overline{c_{kl}} = c_{kl} - u_k - v_l$ for all nonbasic cells. Recall, that $\overline{c_{kl}}$ shows the change in the objective function value for a unit allocation to the nonbasic cell $(k;l)$.

$$\begin{aligned}
\overline{c_{13}} &= c_{13} - u_1 - v_3 = 1 - 0 - 0.4 = 0.6 \\
\overline{c_{21}} &= c_{21} - u_2 - v_1 = 1 - 0.4 - 0.9 = -0.3 \\
\overline{c_{31}} &= c_{31} - u_3 - v_1 = 1.3 - 0.4 - 0.9 \\
\overline{c_{32}} &= c_{32} - u_3 - v_2 = 1 - 0.4 - 1 = -0.4
\end{aligned}$$

The best improvement can be achieved by cell $(3;2)$, so nonbasic variable x_{32} should enter the basis. In order to identify which variable should leave the basis, we have to construct a *loop* that starts and ends at the cell $(3;2)$ and connects only basic cells. In our case this loop consists of the cells

$(3;2) \to (2;2) \to (2;3) \to (3;3) \to (3;2)$. The amount to be allocated to the untied cell is subtracted from and added to the cells of the loop so that the availabilities and requirements remain satisfied and no values in cells can become negative. This implies that when choosing the cell to remove from the basis, we should choose the cell of the minimum value among the cells in an odd number of steps from the starting cell. In our case this minimum value is 5, and the new transportation tableau is:

	F_1	F_2	F_3	
T_1	5	15	0	20
T_2	0	0	15	15
T_3	0	5	5	10
	5	20	20	

We should test the optimality of the new solution. Calculate again the values u_i, v_j corresponding to the tied cells:

$$u_1 + v_1 = 0.9$$
$$u_1 + v_2 = 1$$
$$u_2 + v_3 = 0.8$$
$$u_3 + v_2 = 1$$
$$u_3 + v_3 = 0.8$$

By choosing $u_1 = 0$, we obtain the solution $v_1 = 0.9,\ v_2 = 1,\ u_3 = 0,\ v_3 = 0.8$ and $u_2 = 0$. Next we calculate the values $\overline{c_{kl}}$:

$$\overline{c_{13}} = 1 - 0 - 0.8 = 0.2$$
$$\overline{c_{21}} = 1 - 0 - 0.9 = 0.1$$
$$\overline{c_{22}} = 1.4 - 0 - 1 = 0.4$$
$$\overline{c_{31}} = 1.3 - 0 - 0.9 = 0.4.$$

The nonnegativity of the obtained values implies that the last tableau contains the optimal solution of the problem. The optimum can be determined by multiplying the shipping amounts by the shipping costs:

$$5 \cdot 0.9 + 15 \cdot 1 + 15 \cdot 0.8 + 5 \cdot 1 + 5 \cdot 0.8 = 40.5.$$

11.2.2 Solving a Non-standard Transportation Problem

Sometimes a transportation problem is formulated in a more general way than it was in the previous subchapter. We show that after reformulating the problem the distribution method can be applied for solving it.

1. In case of maximizing (for example an upper bound is needed for the costs) the coefficients of the objective function should be multiplied by -1.

2. There can be prohibited cells in the transportation tableau (for example it is not possible to ship from a certain factory to a certain warehouse). In this case the cell in question is denoted by M, indicating an infinitely large cost. This value remains unchanged during the calculations.

3. If total supply is not equal to total demand, that is $\sum_{i=1}^{m} a_i \neq \sum_{j=1}^{n} b_j$, then a fictitious factory or fictitious warehouse is introduced. Its capacity or demand is defined so as to balance total supply with total demand. All unit transportation costs to or from this fictitious location are defined to be zero (because there will be no transportation here). Sometimes fictitious transportation is not allowed in a location, for example if the demand of a given warehouse must be satisfied by all means. In this case we write M as its cost instead of 0.

Example 11.2 The need for concrete of three buildings (E_1, E_2, E_3) is supplied by three concrete factories (B_1, B_2, B_3). The buildings require 30,40,40 tons of concrete per day, respectively; the daily capacities of the factories are 60, 50 and 20 tons of concrete, respectively. The following table shows the shipping cost of one ton of concrete from each factory to each building:

	E_1	E_2	E_3	capacity
B_1	10	13	6	60
B_2	4	1	9	50
B_3	15	10	6	20
demand	30	40	40	

Find the schedule with minimal transportation cost, if it is required that factories B_1 and B_3 work with at full capacity.

Solution: As total capacity exceeds total demand, a fictitious building E_4 is introduced with a demand of 20 tons. Because of the restrictions, all concrete produced by B_1 and B_3 must be used, so we set the transportation costs infinitely large from these factories to the fictitious building:

	E_1	E_2	E_3	E_4	capacity
B_1	10	13	6	M	60
B_2	4	1	9	0	50
B_3	15	10	6	M	20
demand	30	40	40	20	

This means that the demand of the fictitious building can only be satisfied by factory B_2 (more precisely, the concrete satisfying the fictitious demand will not be produced). Taking into consideration this information we first allocate 20 tons of concrete to cell $(2;4)$ and then start the northwest corner method for finding an initial solution. The result is:

	E_1	E_2	E_3	E_4	
B_1	30	30	0	M	60
B_2	0	10	20	20	50
B_3	0	0	20	M	20
	30	40	40	20	

Based on the tied cells we calculate the dual variables by the equations:

$$\begin{aligned} u_1 + v_1 &= 10 \qquad & u_2 + v_3 &= 9 \\ u_1 + v_2 &= 13 \qquad & u_2 + v_4 &= 0 \\ u_2 + v_2 &= 1 \qquad & u_3 + v_3 &= 6 \end{aligned}$$

By choosing $u_1 = 0$ we have:

$$v_1 = 10,\ v_2 = 13,\ u_2 = -12,\ v_3 = 21,\ v_4 = 12 \text{ and } u_3 = -15.$$

For the nonbasic cells:

$$\begin{aligned} \overline{c_{13}} &= c_{13} - u_1 - v_3 = 6 - 0 - 21 = -15 \\ \overline{c_{21}} &= c_{21} - u_2 - v_1 = 4 + 12 - 10 = 6 \\ \overline{c_{31}} &= c_{31} - u_3 - v_1 = 15 + 15 - 10 = 20 \\ \overline{c_{32}} &= c_{32} - u_3 - v_2 = 10 + 15 - 13 = 12. \end{aligned}$$

It is worthwile to transport from factory B_1 to building E_3. The elements of the loop $(1;3) \to (2;3) \to (2;2) \to (1;2) \to (1;3)$ show that 20 tons of concrete can be reallocated to cell $(1;3)$. The new solution is:

	E_1	E_2	E_3	E_4	
B_1	30	10	20	M	60
B_2	0	30	0	20	50
B_3	0	0	20	M	20
	30	40	40	20	

We have to calculate the dual variables corresponding to the new solution:

$$\begin{aligned} u_1 + v_1 &= 10 & u_2 + v_2 &= 1 \\ u_1 + v_2 &= 13 & u_2 + v_4 &= 0 \\ u_1 + v_3 &= 6 & u_3 + v_3 &= 6 \end{aligned}$$

Let $u_1 = 0$ and we obtain:

$$v_1 = 10,\ v_2 = 13,\ v_3 = 6,\ u_2 = -12,\ v_4 = 12 \text{ and } u_3 = 0.$$

The values $\overline{c_{kl}}$:

$$\begin{aligned} \overline{c_{21}} &= 4 + 12 - 10 = 6 & \overline{c_{31}} &= 15 - 0 - 10 = 5 \\ \overline{c_{23}} &= 9 + 12 - 6 = 15 & \overline{c_{32}} &= 10 - 0 - 13 = -3. \end{aligned}$$

It means that further improvement can be achieved by entering cell $(3;2)$. The corresponding loop is: $(3;2) \to (3;3) \to (1;3) \to (1;2) \to (3;2)$, and 10 units can be tied in the new location:

	E_1	E_2	E_3	E_4	
B_1	30	0	30	M	60
B_2	0	30	0	20	50
B_3	0	10	10	M	20
	30	40	40	20	

We test the optimality of the new solution:

$$\begin{aligned} u_1 + v_1 &= 10 & u_2 + v_4 &= 0 \\ u_1 + v_3 &= 6 & u_3 + v_2 &= 10 \\ u_2 + v_2 &= 1 & u_3 + v_3 &= 6. \end{aligned}$$

The dual variables are

$$u_1 = 0,\ v_1 = 10,\ v_3 = 6,\ u_3 = 0,\ v_2 = 10,\ u_2 = -9 \text{ and } v_4 = 9,$$

so we have

$$\overline{c_{12}} = 13 - 0 - 10 = 3 \qquad \overline{c_{23}} = 9 + 9 - 6 = 12$$
$$\overline{c_{21}} = 4 + 9 - 10 = 3 \qquad \overline{c_{31}} = 15 - 0 - 10 = 5.$$

As all values $\overline{c_{kl}}$ are positive there is no more improvement in the objective function value. The optimal solution can be read from the last tableau. Factory B_2 produces only 30 tons of concrete, the others work with at full capacity. The minimal transportation cost is

$$30 \cdot 10 + 30 \cdot 6 + 30 \cdot 1 + 20 \cdot 0 + 10 \cdot 10 + 10 \cdot 6 = 670.$$

11.3 The Assignment Problem

Consider a collection of n workers and n machines. Each worker must be assigned one and only one machine. Each worker has been rated on each machine and a standardized time for him to complete a standard task is known. These data are given in a cost matrix $C \in \mathbb{R}^{n \times n}$, its element c_{ij} is the standardized time of worker i on machine j.

The aim is to make an assignment of workers to machines so as to minimize the total amount of standardized time of the assignment.

We introduce the following decision variables:

$$x_{ij} = \begin{cases} 1, & \text{if worker } i \text{ is assigned to machine } j \\ 0, & \text{otherwise} \end{cases}$$

The model of the assignment problem:

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} x_{ij} &\to \min \\ \sum_{j=1}^{n} x_{ij} &= 1 \quad i = 1, 2, ..., n \\ \sum_{i=1}^{n} x_{ij} &= 1 \quad j = 1, 2, ..., n \\ x_{ij} &\in \{0; 1\} \quad i, j = 1, 2, ..., n \end{aligned}$$

The feasible solutions of the problem can be given by listing all the quadratic and binary matrices X which contain exactly one 1 in each row and column of X. The total number of these matrices is $n!$, so we have a nonempty, finite feasible set. This means, that the assignment problem always has an optimal solution.

The given model can remind us of the model of the transportation problem, because this formulation is indeed a special case of the transportation problem. In the assignment problem the workers represent factories and the machines represent warehouses, where each factory and each warehouse has a capacity and demand of one unit. That's why the assignment problem can be solved by the distribution method, which was applied in solving transportation problems. However in this chapter we use a more efficient method which exploits the special structure of the assignment problem. This method is called the Hungarian method due to the Hungarian mathematician Dénes Kőnig.

Before introducing the Hungarian method we have two observations.
1) Because there is a one-to-one assignment of workers to machines, the problem reduces to finding a set S of n entries of matrix C with the properties

(i) exactly one entry is selected from each row of C,

(ii) exactly one entry is selected from each column of C,

(iii) the sum of the selected entries is as small as possible.

Example 11.3 Find the minimum time assignment in the following problem, where W_i denotes worker i and M_j denotes machine j, $(1 \leq i, j \leq 4)$, and the numbers in position (i, j) give the time needed for W_i to complete his work on machine M_j.

	M_1	M_2	M_3	M_4
W_1	5	**1**	3	2
W_2	**3**	2	4	5
W_3	6	7	**3**	2
W_4	4	5	4	**1**

In this case it is easy to identify the minimal set S. Because every boldfaced number is the least in its column, and they are located in different rows, the set including the boldfaced entries satisfy properties (i)-(iii).

2) Let C' denote the matrix which is obtained by subtracting a number α_i from every element in row i of C and subtracting a number β_j from every element in column j of C. We will show, that the set of minimal sets S for C and C' are identical. If c'_{ij} denotes the elements of matrix C', we have

$$c'_{ij} = c_{ij} - \alpha_i - \beta_j.$$

If z' and z are the objective function values for the problems represented by C' and C, respectively, then

$$\begin{aligned} z' &= \sum_{i=1}^{n}\sum_{j=1}^{n} c'_{ij}x_{ij} = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(c_{ij}-\alpha_i-\beta_j\right)x_{ij} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i x_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{n}\beta_j x_{ij} \\ &= z - \sum_{i=1}^{n}\alpha_i \underline{\sum_{j=1}^{n} x_{ij}} - \sum_{j=1}^{n}\beta_j \underline{\sum_{i=1}^{n} x_{ij}} \end{aligned}$$

Corresponding to the constraints of the assignment problem the underlined sums equal 1, so we obtain

$$z' = z - \sum_{i=1}^{n}\alpha_i - \sum_{j=1}^{n}\beta_j .$$

Thus z' and z differ only by the total amount subtracted, which is a constant, proving that the problems have identical minimal sets.

Consider again Example 11.3, and make row reductions by subtracting the smallest entry of row i from every element of the row:

$$\begin{array}{cccccccccc} 5 & 1 & 3 & 2 & (-1) & \Longrightarrow & 4 & 0 & 2 & 1 \\ 3 & 2 & 4 & 5 & (-2) & \Longrightarrow & 1 & 0 & 2 & 3 \\ 6 & 7 & 3 & 2 & (-2) & \Longrightarrow & 4 & 5 & 1 & 0 \\ 4 & 5 & 4 & 1 & (-1) & \Longrightarrow & 3 & 4 & 3 & 0 \end{array}$$

Then apply the same operation to the columns. We subtract 1 from every element of the first and third column. The other two columns are unchanged, because their least element equals 0:

$$\begin{array}{cccc} 3 & \mathbf{0} & 1 & 1 \\ \mathbf{0} & 0 & 1 & 3 \\ 3 & 5 & \mathbf{0} & 0 \\ 2 & 4 & 2 & \mathbf{0} \end{array}$$

If we choose the boldfaced zeros we have the optimal solution of the modified problem with zero optimum. Observe that the minimal sets S and S' are identical for the original and the modified problem and the optimums differ by the sum of the numbers with which the rows and columns were reduced.

11.3.1 The Hungarian Method

The algorithm called Hungarian method was published by W. H. Kuhn in 1955. The name is explained by the fact that the paper was inspired by the work of Hungarian mathematicians, Dénes Kőnig and Jenő Egerváry. We need the following definition.

Definition 11.1 Let C be a matrix containing nonnegative elements. A set S of zeros in C is said to be independent, if there are no two elements in S in the same row or in the same column.

Example 11.4 Two sets of zeros are signed in the same matrix as follows:

$$\begin{bmatrix} 3 & \mathbf{0} & 1 & 1 \\ \mathbf{0} & 0 & 1 & 3 \\ 3 & 5 & \mathbf{0} & 0 \\ 2 & 4 & 2 & \mathbf{0} \end{bmatrix} \qquad \begin{bmatrix} 3 & \mathbf{0} & 1 & 1 \\ 0 & \mathbf{0} & 1 & 3 \\ 3 & 5 & \mathbf{0} & \mathbf{0} \\ 2 & 4 & 2 & 0 \end{bmatrix}$$

On the left, the set of the boldfaced zeros is independent, because all zeros are located in different rows and columns. The set of the boldfaced zeros in the right-hand-side matrix is not independent, since it contains more than one zeros from the third row and the second column.

The concept of the independent set of zeros has a key role in solving the assignment problem. If the cost matrix C is in $\mathbb{R}^{n\times n}$ and we find an independent set of zeros S having n elements, we have the optimal solution of the problem, because the objective value corresponding to S equals zero, which is obviously the possible least value of the objective function. The next theorem offers a way to determine the cardinality of the independent sets.

Theorem 11.2 (Kőnig's theorem) *The maximum number of independent zeros is equal to the minimum number of vertical and horizontal lines which are needed for covering all zeros in the matrix.*

In Example 11.4 we need four lines to cover all zeros in matrix A, it has an independent set of zeros containing four zeros. Matrix B has independent sets up to two elements since all zeros in B can be covered by two lines.

Now we give the algorithm of the Hungarian method. It starts from a cost matrix $C \in \mathbb{R}^{n\times n}$ and aims to find an independent set of zeros containing n elements.

1. Make the reduction of matrix C subtracting the smallest element per row and per column.

2. Cover all zeros with minimum number of lines using the following method:

    i) identify a row (column) with exactly one uncrossed zero and draw a vertical (horizontal) line through this zero,

    ii) if all rows and columns with zeros have at least two uncrossed zeros, choose the row or column with the least, identify one of the zeros and proceed as in i).

3. If the minimum number of lines necessary to cross out all the zeros equals n, then a minimal S can be identifed providing the optimal solution of the problem.

4. If the minimum number of lines is less than n, some new zeros should be created as follows:

    a) subtract the smallest uncrossed entry ϵ from all uncrossed entry,

    b) add ϵ to each double crossed entry (elements through which only one line pass are not changed),

    c) remove all lines and proceed as in Step 2.

Example 11.5 Solve the assignment problem represented by the cost matrix

$$C = \begin{bmatrix} 8 & 2 & 14 & 7 \\ 13 & 4 & 10 & 13 \\ 12 & 5 & 3 & 7 \\ 12 & 8 & 8 & 20 \end{bmatrix}.$$

After the row reductions (subtracting 2, 4, 3 and 8) we obtain

$$\begin{bmatrix} 6 & 0 & 12 & 5 \\ 9 & 0 & 6 & 9 \\ 9 & 2 & 0 & 4 \\ 4 & 0 & 0 & 12 \end{bmatrix}.$$

In the first and the last columns column reduction is alse needed, subtracting 4 in both cases:

$$\begin{bmatrix} 2 & 0 & 12 & 1 \\ 5 & 0 & 6 & 5 \\ 5 & 2 & 0 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}.$$

Now we start to cover zeros. The first row contains a single zero, we cover its column.

$$\begin{bmatrix} 2 & 0 & 12 & 1 \\ 5 & 0 & 6 & 5 \\ 5 & 2 & 0 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}$$

The third and fourth rows both have two uncrossed zeros, so we move on the columns. The first column contains a single zero, its row is covered:

$$\begin{bmatrix} 2 & 0 & 12 & 1 \\ 5 & 0 & 6 & 5 \\ 5 & 2 & 0 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}$$

The third column have a single uncrossed zero, its row is covered:

$$\begin{bmatrix} 2 & 0 & 12 & 1 \\ 5 & 0 & 6 & 5 \\ 5 & 2 & 0 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}$$

Three lines is enough to cover all zeros, so we need further reductions to achieve the optimal solution.

The smallest uncrossed element is $\epsilon = 1$. This number is subtracted from all uncrossed entries and added to the double crossed elements. After the transformation we have

$$\begin{bmatrix} 1 & 0 & 11 & 0 \\ 4 & 0 & 5 & 4 \\ 5 & 3 & 0 & 0 \\ 0 & 1 & 0 & 8 \end{bmatrix}.$$

We have to start again the covering process. There is a single zero in the second row, its column is covered:

$$\begin{bmatrix} 1 & 0 & 11 & 0 \\ 4 & \mathbf{0} & 5 & 4 \\ 5 & 3 & 0 & 0 \\ 0 & 1 & 0 & 8 \end{bmatrix}$$

Then we cover the column of the uncrossed zero remained in the first row:

$$\begin{bmatrix} 1 & 0 & 11 & \mathbf{0} \\ 4 & 0 & 5 & 4 \\ 5 & 3 & 0 & 0 \\ 0 & 1 & 0 & 8 \end{bmatrix}$$

Now there is an uncrossed zero in the third row, its column is covered:

$$\begin{bmatrix} 1 & 0 & 11 & 0 \\ 4 & 0 & 5 & 4 \\ 5 & 3 & \mathbf{0} & 0 \\ 0 & 1 & 0 & 8 \end{bmatrix}$$

The last uncrossed zero is in the fourth row and in the first column, we can choose in covering the first column or the last row:

$$\begin{bmatrix} 1 & 0 & 11 & 0 \\ 4 & 0 & 5 & 4 \\ 5 & 3 & 0 & 0 \\ \mathbf{0} & 1 & 0 & 8 \end{bmatrix}$$

All the zeros are covered and four lines required to make it. This means that there is an independent set containing four zeros. These zeros were indicated by boldfacing when they induced the selection of the next covering line. So the optimal solution of the problem is $x_{14} = 1,\ x_{22} = 1,\ x_{33} = 1,\ x_{41} = 1$ and the remained decision variables have the value 0. The optimum is the sum of the corresponding elements in the cost matrix:

$$z = c_{14} + c_{22} + c_{33} + c_{41} = 7 + 4 + 3 + 12 = 26.$$

It is equal to the total of the numbers subtracted during the reductions: $2 + 4 + 3 + 8 + 4 + 4 + 1 = 26$.

11.3.2 The Generalized Assignment Problem

The first option to make the assignment problem more general is to look for the maximum instead of the minimum of the objective function. This problem first appeared in the literature as a marriage game in the 1950s. In the wording of P.R. Halmos and H. Vaughon, the task is as follows:

A colony of settlers of 10 bachelors is supplemented by 10 potential brides. After a short courtship, the young men decide to get married immediately.

Each bridal candidate will receive a list of names with the names of the 10 young men, and on this she will have to score the husband candidates, the one that best suits her with 10 points, the next chosen one with 9 points, and so on. Suppose that for any fixed mate choice, the happiness of the whole colony is proportional to the sum of the scores given by the bridal candidates. Determine a mate choice that will result in maximum happiness for the entire colony.

Example 11.6 We examine the case of a smaller colony where the optimal pairs between 5 men and 5 women have to be found. Suppose the ladies scored the men on a scale of 1 to 5, and summing up the ratings, we obtained the matrix below.:

	M_1	M_2	M_3	M_4	M_5
W_1	5	2	3	1	4
W_2	3	1	4	5	2
W_3	2	1	4	5	3
W_4	4	2	5	3	1
W_5	2	1	3	5	4

This time we do not convert the maximum problem to the minimum problem by multiplying the objective function by (-1) as usual in LP problems. In this case it is better to subtract all the elements from the largest element of the cost matrix (in our case 5) and applying the Hungarian method to this "complementary" cost matrix. The transformed matrix:

$$\begin{matrix} 0 & 3 & 2 & 4 & 1 \\ 2 & 4 & 1 & 0 & 3 \\ 3 & 4 & 1 & 0 & 2 \\ 1 & 3 & 0 & 2 & 4 \\ 3 & 4 & 2 & 0 & 1 \end{matrix}$$

Since each row contains a zero, row reduction is not required. We perform two column reductions: subtract 3 from the elements of the second column and 1 from the elements of the fifth column:

$$\begin{matrix} 0 & 0 & 2 & 4 & 0 \\ 2 & 1 & 1 & 0 & 2 \\ 3 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 2 & 3 \\ 3 & 1 & 2 & 0 & 0 \end{matrix}$$

All zeros in the matrix can be covered by four lines, for example, as follows:

$$\begin{array}{ccccc}
0 & 0 & 2 & 4 & 0 \\
2 & 1 & 1 & 0 & 2 \\
3 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 2 & 3 \\
3 & 1 & 2 & 0 & 0
\end{array}$$

The minimum of the uncovered elements is 1. It is subtracted from every uncovered elements and is added to every double covered elements:

$$\begin{array}{ccccc}
0 & 0 & 2 & 5 & 1 \\
1 & 0 & 0 & 0 & 2 \\
2 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 3 & 4 \\
2 & 0 & 1 & 0 & 0
\end{array}$$

We start again the zero covering process:

$$\begin{array}{ccccc}
0 & 0 & 2 & 5 & 1 \\
1 & 0 & 0 & 0 & 2 \\
2 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 3 & 4 \\
2 & 0 & 1 & 0 & 0
\end{array}$$

After drawing three cover lines, we see that there is an alternative optimum, because the five-element independent zero set can be selected in two ways:

$$\begin{array}{ccccc}
\mathbf{0} & 0 & 2 & 5 & 1 \\
1 & 0 & 0 & \mathbf{0} & 2 \\
2 & \mathbf{0} & 0 & 0 & 1 \\
1 & 0 & \mathbf{0} & 3 & 4 \\
2 & 0 & 1 & 0 & \mathbf{0}
\end{array}
\qquad
\begin{array}{ccccc}
\mathbf{0} & 0 & 2 & 5 & 1 \\
1 & 0 & 0 & \mathbf{0} & 2 \\
2 & 0 & \mathbf{0} & 0 & 1 \\
1 & \mathbf{0} & 0 & 3 & 4 \\
2 & 0 & 1 & 0 & \mathbf{0}
\end{array}$$

The optimal assignment determined by the left matrix:

$$W_1 - M_1;\ W_2 - M_4;\ W_3 - M_2;\ W_4 - M_3;\ W_5 - M_5.$$

The optimum: 5+5+1+5+4=20 (in the other case the optimum is the same).

An assignment problem, similarly to a transportation problem, can be interpreted and solved if the number of workers and machines are not the same, or if one of the assignments is forbidden (for example, if there is a machine that one of the workers cannot manage). In the first case, fictitious actors have to be introduced, and in the second case, we denote the prohibited assignment at an infinitely high cost marked with M.

Example 11.7 A math teacher prepares five students for a math competition where 6 problems are given, but only five need to be solved. The table below shows the test result for each student, the number of minutes they would expect to complete each problem type. Give the optimal assignment between the students and the problems, for which the total time for solving the tasks is the minimum, taking into account the following conditions: student S_5 cannot cope with task P_6, and problem P_5 must be solved!

	P_1	P_2	P_3	P_4	P_5	P_6
S_1	14	13	12	17	15	10
S_2	8	12	9	11	10	12
S_3	3	4	2	6	5	7
S_4	7	8	6	9	9	6
S_5	11	17	14	16	15	M

We take a fictitious student denoted by S_6, to which we will assign the task that will not be solved in the end. However, according to the conditions, this cannot be P_5, so we will insert a prohibitory mark M here and zero time for all the other problems.

	P_1	P_2	P_3	P_4	P_5	P_6
S_1	14	13	12	17	15	10
S_2	8	12	9	11	10	12
S_3	3	4	2	6	5	7
S_4	7	8	6	9	9	6
S_5	11	17	14	16	15	M
S_6	0	0	0	0	M	0

Instead of solving the problem in detail, we only state the final result: the value of the optimum is 41 minutes, which can be achieved, for example, with the following assignment (because there is an alternative optimum here as well):

$$S_1 - P_6;\ S_2 - P_5;\ S_3 - P_3;\ S_4 - P_2;\ S_5 - P_1,$$

The unsolved problem was P_4.

12 Nonlinear Optimization (NLO)

The general form of NLO problems:

$$f(x) \to \min\ (\text{or}\max)$$

subject to

$$h_i(x) = 0 \qquad (i = 1, 2, \ldots m)$$
$$g_j(x) \leq 0 \qquad (j = 1, 2, \ldots r)$$

where $f, h_i, g_j : S \to \mathbb{R}$, and S is a region in the n-dimensional Euclidean space $\mathbb{R}^n$, so f, h_i, g_j are functions of n variables.

12.1 Classical Unconstrained Optimization

In this chapter $f : S \to \mathbb{R}$ is a multi-variabled function on the set $S \subseteq \mathbb{R}^n$. By the definitions given in Chapter 7, we say that f has a global minimum (maximum) at $x^* \in S$ if $f(x^*) \leq f(x)$ $(f(x^*) \geq f(x))$ for all $x \in S$, and f has a local minimum (maximum) at $x^* \in S$ if there exists a $\delta > 0$ such that $f(x^*) \leq f(x)$ $(f(x^*) \geq f(x))$ for all $x \in S$ satisfying $\mid x - x^* \mid < \delta$.

Definition 12.1 (Stationary Point) f has a stationary point at $x^* \in S$ if f is differentiable at x^* and

$$\frac{\partial f}{\partial x_i}(x^*) = 0 \ \text{ for all } i = 1, 2, \ldots n.$$

Remark 12.1 *Using the notion of gradient vector grad* $f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots \frac{\partial f}{\partial x_n}\right)$, *we can write that* f *has a stationary point at* $x^* \in S$ *if* f *is differentiable at* x^* *and*

$$grad\ f(x^*) = 0.$$

Theorem 12.1 (Necessary Condition for Local Extrema) *Let* x^* *be an interior point of* S. *Suppose that* f *is differentiable at* x^* *and* f *has a local extremum (minimum or maximum) at* x^*. *Then* f *has a stationary point at* x^*.

Definition 12.2 (Matrix Definitness) A matrix $H \in \mathbb{R}^{n \times n}$ is said to be positive (negative) definite if

$$x^T H x > 0 \qquad \left(x^T H x < 0\right)$$

for all nonzero vector $x \in \mathbb{R}^n$.

Remark 12.2 (Check of Definiteness) *Compute the determinants of the main minors H_r of H, $(r = 1, 2, \dots n)$.*

a) If $det(H_r) > 0$ for all r, then H is positive definite.

b) If $sgn(det(H_r)) = (-1)^r$ for all r, then H is negative definite.

Definition 12.3 (Hessian Matrix) The set of second partial derivatives of f assembling into a matrix

$$H(x) = \begin{bmatrix} \frac{\partial f^2}{\partial x_1 \partial x_1}(x) & \frac{\partial f^2}{\partial x_1 \partial x_2}(x) & \dots & \frac{\partial f^2}{\partial x_1 \partial x_n}(x) \\ \frac{\partial f^2}{\partial x_2 \partial x_1}(x) & \frac{\partial f^2}{\partial x_2 \partial x_2}(x) & \dots & \frac{\partial f^2}{\partial x_2 \partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f^2}{\partial x_n \partial x_1}(x) & \frac{\partial f^2}{\partial x_n \partial x_2}(x) & \dots & \frac{\partial f^2}{\partial x_n \partial x_n}(x) \end{bmatrix}$$

is called the Hessian matrix of f at x.

Theorem 12.2 (Sufficient and Necessary Condition for Local Extrema) *Let x^* be an interior point of S and suppose that f has a stationary point at x^*.*

a) If $H(x^*)$ is negative definite, then f has a local maximum at x^*.

b) If $H(x^*)$ is positive definite, then f has a local minimum at x^*.

Example 12.1 The function $f(x_1, x_2) = 3x_1x_2 - {x_1}^3 - {x_2}^3$ has a local maximum at point $(1; 1)$.

Example 12.2 The function

$$f(x_1, x_2, x_3) = {x_1}^2 + {x_2}^2 + {x_3}^2 + x_1x_2 + x_1x_3 + x_2x_3 - 7x_1 - 8x_2 - 9x_3 + 10$$

has a local minimum at point $(1; 2; 3)$.

12.2 Multidimensional Optimization with Equality Constraints (the Method of Lagrange)

The problem:

$$f(x) \to \min \text{ (or max)}$$

subject to

$$h_i(x) = 0 \qquad (i = 1, 2, \dots m),$$

where $x \in \mathbb{R}^n$.

An obvious approach for solving the problem is using the equations to eliminate some of the variables from the objective function.

Example 12.3 Find the point in line $y = 3 - 2x$ which is nearest to the origin.

Solution: We have to minimize the distance $\sqrt{x^2 + y^2}$, where $(x; y)$ is a point in the line. The monotonicity of the square root function implies that it is enough to minimize the function $x^2 + y^2$ subject to the constraint $y = 3 - 2x$. Replacing y by $3 - 2x$ in the objective function, we have

$$x^2 + y^2 = x^2 + (3 - 2x)^2 = x^2 + 9 - 12x + 4x^2 = 5x^2 - 12x + 9.$$

The derivative of the function $f(x) = 5x^2 - 12x + 9$ is $f'(x) = 10x - 12$, so f has a stationary point at $x = 1.2$. The second derivative of f is $f''(x) = 10$, which is positive at point 1.2 implying that f has a local minimum at this point. So, the coordinates of the closest point are $x = 1.2$ and $y = 3 - 2x = 0.6$. The minimum distance to the origin is $\sqrt{1.2^2 + 0.6^2} = \sqrt{1.8}$.

If we have more complicated constraints we can use the method of Lagrange. Form the *Lagrange function* of variables $x = (x_1, x_2, \dots x_n)$ and $\lambda = (\lambda_1, \lambda_2, \dots \lambda_m)$ as follows:

$$F(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x).$$

Set the system of equations:

$$\begin{aligned} \frac{\partial F}{\partial x_i} &= 0 \qquad (i = 1, 2, \dots, n) \\ \frac{\partial F}{\partial \lambda_j} &= 0 \qquad (j = 1, 2, \dots, m) \end{aligned} \tag{16}$$

Theorem 12.3 *If f has an extremum at x^*, then x^* is the x-part of some solution of system* (16).

Example 12.4 Find the point in circle $(x - 3)^2 + (y - 4)^2 = 9$ which is nearest to the origin.

Solution: The Lagrange function is

$$F(x, y, \lambda) = x^2 + y^2 + \lambda\left((x - 3)^2 + (y - 4)^2 - 9\right).$$

Its derivatives are

$$F'_x = 2x + 2\lambda(x - 3), \quad F'_y = 2y + 2\lambda(y - 4), \quad F'_\lambda = (x - 3)^2 + (y - 4)^2 - 9.$$

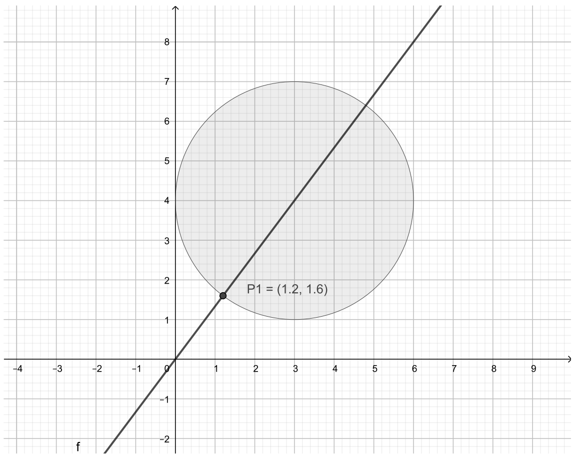
We have to solve the system

$$2x + 2\lambda(x-3) = 0$$

$$2y + 2\lambda(y-4) = 0$$

$$(x-3)^2 + (y-4)^2 - 9 = 0$$

If we express λ from the first and the second equation we obtain $\lambda = -\frac{x}{x-3}$ and $\lambda = -\frac{y}{y-4}$. That is $-\frac{x}{x-3} = -\frac{y}{y-4}$, from which we have $y = \frac{4}{3}x$. y is replaced in equation $(x-3)^2 + (y-4)^2 - 9 = 0$ yielding a quadratic equation with solutions $\frac{6}{5}$ and $\frac{24}{5}$. So we have two points $P_1(\frac{6}{5}; \frac{8}{5})$ and $P_2(\frac{24}{5}; \frac{32}{5})$. P_1 is the nearest point to the origin and P_2 is the farthest point from the origin.



Example 12.5 Using the method of Lagrange find the possible solutions of the following NLO problem:

$$x^2 + 3xy + 2y^2 + 4x + 0,5z^2 + 12 \to \min$$
$$x + y + z = 4$$
$$x - z = 2$$

Solution: The Lagrange function:

$$F(x, y, z, \lambda_1, \lambda_2) = x^2+3xy+2y^2+4x+0,5z^2+12+\lambda_1(x+y+z-4)+\lambda_2(x-z-2)$$

We get the following system of equations forming the partial derivatives:

$$2x + 3y + 4 + \lambda_1 + \lambda_2 = 0 \tag{17}$$
$$3x + 4y + \lambda_1 = 0 \tag{18}$$
$$z + \lambda_1 - \lambda_2 = 0 \tag{19}$$
$$x + y + z - 4 = 0 \tag{20}$$
$$x - z - 2 = 0 \tag{21}$$

First we try to get rid of λ_2, so equation (17) and (19) are added:

$$2x + 3y + z + 4 + 2\lambda_1 = 0, \tag{22}$$

and from this sum the double of equation (18) is subtracted:

$$-4x - 5y + z + 4 = 0. \tag{23}$$

The unique solution of the system consisting equations (20),(21) and (23) is the point $(4; -2; 2)$. This is our only candidate for the optimal solution of the problem.

12.3 Multidimensional Optimization with Inequality Constraints (Karush - Kuhn - Tucker Conditions)

Theorem 12.4 *Suppose that f has a local minimum at x^* in the feasible region of the problem*

$$f(x) \to \min$$

subject to

$$h_i(x) = 0 \qquad (i = 1, 2, \dots m), \quad \textit{and}$$
$$g_j(x) \le 0 \qquad (j = 1, 2, \dots r),$$

where f, h_i and g_j have continuous first derivatives,
then it is necessary that the partial derivatives with respect to x_i of the Lagrange function

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x)$$

are equal to 0 at x^* and

$$\mu_j g_j(x^*) = 0 \qquad (j = 1, 2, \dots r),$$
$$\mu_j \ge 0 \qquad (j = 1, 2, \dots r),$$

for some set of real numbers $\lambda_1, \lambda_2, \dots \lambda_m$, $\mu_1, \mu_2 \dots \mu_r$, (and at least one of them differs from 0).

The points which satisfy all the listed conditions and the original constraints are referred as KKT points (Karush - Kuhn - Tucker conditions).

Example 12.6 Find the minimum of the function

$$f(x, y) = x^2 - 12x + y^2 - 4y + 5$$

subject to the constraints

$$x + 4y - 16 = 0$$

$$x^2 + y^2 - 25 \leq 0.$$

Solution: The Lagrange function is

$$L(x, y, \lambda, \mu) = x^2 - 12x + y^2 - 4y + 5 + \lambda(x + 4y - 16) + \mu(x^2 + y^2 - 25).$$

We have to find the points satisfying the conditions

$$2x - 12 + \lambda + 2\mu x = 0 \quad (24)$$

$$2y - 4 + 4\lambda + 2\mu y = 0 \quad (25)$$

$$\mu(x^2 + y^2 - 25) = 0 \quad (26)$$

$$\mu \geq 0 \quad (27)$$

$$x + 4y - 16 = 0 \quad (28)$$

$$x^2 + y^2 - 25 \leq 0 \quad (29)$$

We start from (26), assuming first that $\mu = 0$. From (24) we have $\lambda = 12 - 2x$, from (25) $\lambda = 1 - 0,5y$. So $12 - 2x = 1 - 0,5y$, that is $y = 4x - 22$. Substituting it to (28), we have $x = \frac{104}{17}$ and $y = \frac{42}{17}$, but this point does not satisfy (29). Now assume $\mu > 0$, then $x^2 + y^2 = 25$ from (26). We have $x = 16 - 4y$ from (28), and we get the quadratic equation

$$(16 - 4y)^2 + y^2 = 25.$$

The first solution is $y = 3$. In this case $x = 4$. The point $(4; 3)$ is a KKT point, because if its coordinates are replaced into (24) and (25), we get the system

$$\lambda + 8\mu = 4$$

$$4\lambda + 6\mu = -2$$

with the positive solution $\mu = \frac{9}{13}$.
The second solution to the quadratic is $y = 4.53$, implying $x = -2.12$. The point $(-2.12; 4.53)$ is not a KKT point, because when checking the sign of μ, we get $\mu = -2.69$, conflicting with (27).

Conclusion: the feasible region is bounded and f is continuous, so f has a minimum at the KKT point $(4; 3)$.

Example 12.7

$$\begin{aligned}(x_1-4)^2+(x_2-4)^2 &\to \min\\ x_1+x_2 &\le 4\\ x_1+3x_2 &\le 9\end{aligned}$$

Solution. We have two inequality constraints, so the Lagrange function:

$L(x_1,x_2,\mu_1,\mu_2)=(x_1-4)^2+(x_2-4)^2+\mu_1(x_1+x_2-4)+\mu_2(x_1+3x_2-9)$

The KKT points must statisfy the following system:

$$\begin{aligned}2(x_1-4)+\mu_1+\mu_2&=0 \qquad (30)\\ 2(x_2-4)+\mu_1+3\mu_2&=0 \qquad (31)\\ \mu_1(x_1+x_2-4)&=0 \qquad (32)\\ \mu_2(x_1+3x_2-9)&=0 \qquad (33)\\ \mu_1&\ge 0 \qquad (34)\\ \mu_2&\ge 0 \qquad (35)\\ x_1+x_2-4&\le 0 \qquad (36)\\ x_1+3x_2-9&\le 0 \qquad (37)\end{aligned}$$

We start from equations (32) and (33) observing that one of the factors of the products must be equal to 0. According to Theorem 12.4 variables μ_1 and μ_2 can not equals zero at the same time, so we study the following three cases:

Case 1: $\mu_1=0$ and $x_1+3x_2-9=0$ (then $\mu_2\ne 0$, so (35) implies $\mu_2>0$). Equations (30) and (31) after replacing $\mu_1=0$ into them:

$$\begin{aligned}2(x_1-4)+\mu_2&=0\\ 2(x_2-4)+3\mu_2&=0\end{aligned}$$

The first equation is multiplied by 3, and the second equation is subtracted from it: $6x_1-2x_2-16=0$, implying $x_2=3x_1-8$. This is substituted in equation $x_1+3x_2-9=0$ we have: $10x_1-33=0$, that is $x_1=3.3$. The value of the other variable: $x_2=1.9$. The point (3.3;1.9) does not satisfy condition (36), so it can not be a KKT point.

Case 2: $\mu_2=0$ and $x_1+x_2-4=0$ (then $\mu_1\ne 0$ and (35) implies $\mu_1>0$). We substitute $\mu_2=0$ into equations (30) and (31):

$$\begin{aligned}2(x_1-4)+\mu_1&=0\\ 2(x_2-4)+\mu_1&=0\end{aligned}$$

The difference of these equations is $x_1 = x_2$. Using this result in equation $x_1+x_2 = 4$ we obtain $x_1 = x_2 = 2$. Point $(2;2)$ satisfies the further conditions as well, but we have to compute the value of μ_1, in order to check its positivity. Equation $2(x_1 - 4) + \mu_1 = 0$ implies $\mu_1 = 4$, so all conditions are fulfilled, point $(2;2)$ is a KKT point of the problem.

Case 3: $x_1 + x_2 - 4 = 0$ and $x_1 + 3x_2 - 9 = 0$ (then $\mu_1 > 0$ and $\mu_2 > 0$ must be also satisfied). The solution of the system

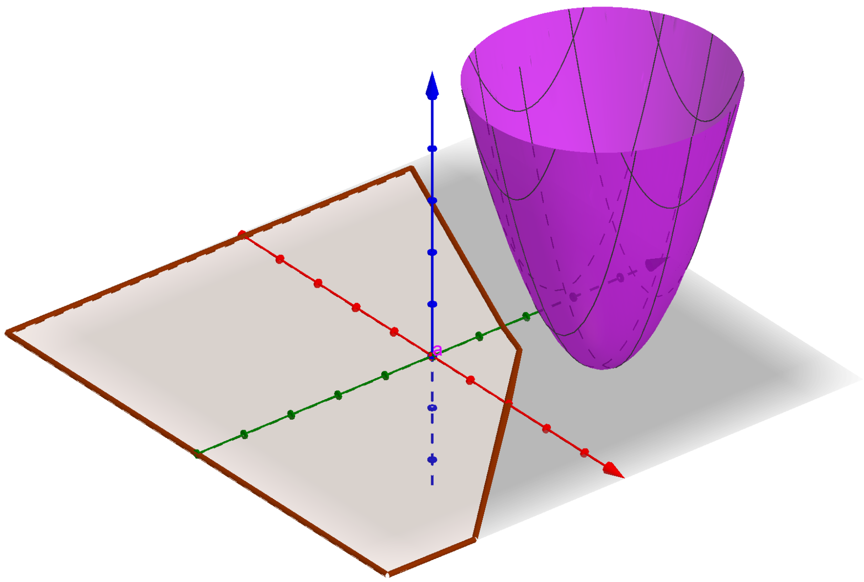
$$\begin{aligned} x_1 + x_2 &= 4 \\ x_1 + 3x_2 &= 9 \end{aligned}$$

is $x_1 = 1.5, \quad x_2 = 2.5$. These numbers are replaced in equations (30) and (31):

$$\begin{aligned} -5 + \mu_1 + \mu_2 &= 0 \\ -3 + \mu_1 + 3\mu_2 &= 0 \end{aligned}$$

The difference of the equations is $2 + 2\mu_2 = 0$ implying $\mu_2 = -1$, which conflicts with condition (35) so point $(1.5;2.5)$ is not a KKT point.

We obtained only one KKT point, which is actually the optimal solution of the problem (observe that we have a convex objective function, which must have a minimizer). The next figure illustrates the surface given by the objective function and the feasible region on the plane xy.



Example 12.8

$$\begin{aligned} xy &\to \max \\ x + y^2 &\leq 2 \\ x &\geq 0 \\ y &\geq 0 \end{aligned}$$

Solution: It is a maximization problem, so we have to use the opposite of the objective function. The restrictions regarding the signs of x and y must be considered as extra conditions, so we have three μ-type variables in the Lagrange function:

$$L(x, y, \mu_1, \mu_2, \mu_3) = -xy + \mu_1(x + y^2 - 2) + \mu_2(-x) + \mu_3(-y)$$

The KKT points must statisfy the following system:

$$-y + \mu_1 - \mu_2 = 0 \tag{38}$$
$$-x + 2y\mu_1 - \mu_3 = 0 \tag{39}$$
$$\mu_1(x + y^2 - 2) = 0 \tag{40}$$
$$\mu_2(-x) = 0 \tag{41}$$
$$\mu_3(-y) = 0 \tag{42}$$
$$\mu_1, \mu_2, \mu_3 \geq 0 \tag{43}$$
$$x, y \geq 0 \tag{44}$$
$$x + y^2 - 2 \leq 0 \tag{45}$$

Assume, that in equation (40) μ_1 equals 0. Then (38) and (39) yield $y+\mu_2 = 0$ and $x + \mu_3 = 0$. These equations contains only nonnegative numbers, so all the variables must be zero to make the equations true. It is easy to see, that the objective function can not have maximum at the point $(0; 0)$, so option $\mu_1 = 0$ can be discarded. Therefore $\mu_1 > 0$ and in equation (40) the other factor implies the zero result, so $x = 2-y^2$. Suppose that $x = 0$, then $y = \sqrt{2}$ implies $\mu_3 = 0$ according to (42). From equation (39) we get $\mu_1 = 0$, which was previously exluded, so $x = 0$ can be also omitted. Therefore $x > 0$, and (41) implies $\mu_2 = 0$. Then $y = \mu_1$ follows from (38). Since $\mu_1 \neq 0$, y can not be zero, so (42) implies $\mu_3 = 0$. These observations are summarized in equation (39) obtaining

$$-(2 - y^2) + 2y^2 - 0 = 0,$$

from which we have $y^2 = \frac{2}{3}$, implying $y = \sqrt{\frac{2}{3}}$ and $x = 2 - y^2 = \frac{4}{3}$.

We have found only one KKT point. Since the feasible set is bounded and the objective function is continuous, there must be a minimizer in the feasible set, so the point $(\frac{4}{3}; \sqrt{\frac{2}{3}})$ is the optimal solution of the problem.

References

[1] Butenko, S., Pardalos, P. M.: Numerical Methods and Optimization an Introduction, Taylor and Francis Group, 2014.

[2] Chapra, S. C., Canale, R. P.: Numerical Methods for Engineers, McGraw-Hill Education, 2015.

[3] Foulds, L. R.: Optimization Techniques an Introduction, Springer-Verlag, New York, 1981.

[4] Nocedal, J., Wright, S. J.: Numerical Optimization, Springer-Verlag, New York, 1999.