Numerical Methods
and
Optimization

Bachelor
Applied Artificial Intelligence
(AAI-B3)
DRAFT

Florian Link
December 19, 2022

Contents

Important NOTE

Please report errors to me directly by email!

Rosenheim, März 2022. Florian Link

1. Introduction

2. Systems of Linear Equations - Direct Methods

2.1. Introduction

Systems of linear equations play a crucial role in almost all applications, not just numerical ones. One often has a complicated and non-linear problem for which a linear approximation, for example the first-order Taylor polynomial, is accepted as a good approximation. These linear approximations then lead to systems of linear equations, which still require a lot of effort to solve when you have many thousands of unknowns, even on modern computers. For this reason we want to deal with the topic of linear equation systems and their numerical treatment first. Linear equation systems often come directly from the applications, as in the following example.

Example 2.1 (complete emptying of warehouses). A solar module producer produces three different types of solar modules M1, M2 and M3. The parts required for production are listed in the table below.

	M1	M2	M3
solar cells	24	48	72
cables	1	1	1
solar glass	1	4	2

The producer currently has 76800 solar cells, 1700 cables and 2850 pieces of solar glass in stock. Is there a production possibility to completely empty the warehouse? So we are looking for a number of modules M1, M2 and M3 that leads to the complete emptying of the stores.

Written as a system of equations, this problem looks as follows (we denote the required amount of M1,M2 and M3 with x_1, x_2 and x_3)

$$\begin{aligned} 24x_1 + 48x_2 + 72x_3 &= 76800 \\ x_1 + x_2 + x_3 &= 1700 \\ x_1 + 4x_2 + 2x_3 &= 2850 \end{aligned} \tag{2.1}$$

In matrix formulation, the problem is briefly written as

$$Ax = b$$

with

$$A = \begin{pmatrix} 24 & 48 & 72 \\ 1 & 1 & 1 \\ 1 & 4 & 2 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad b = \begin{pmatrix} 76800 \\ 1700 \\ 2850 \end{pmatrix},$$

where we (almost always) omit the vector arrows in the following. This is a linear system of equations that can be solved with the Gauss algorithm, for example. Its implementation on the computer is one of the topics of this section. //.

Now briefly on the delimitation to nonlinear systems of equations: We call systems nonlinear in which the unknowns do not only appear as scalar multiples, e.g. is the equation

$$x_1^5 + x_2 - x_3 = 4$$

nonlinear, whereas the equation

$$x_1 - x_2 = \sqrt{\pi}$$

is linear.

Before we develop the first algorithms for solving systems of linear equations, some basics from linear algebra shall be refreshed.

2.2. Fundamentals of linear algebra

We are mainly interested in quadratic systems of equations the corresponding matrix is then quadratic and there are exactly as many unknowns as equations. Nevertheless, let us first reiterate a few facts of the general case of systems of equations with a non-square coefficient matrix.

Let $A \in \mathbb{R}^{m\times n}$. Then we have for

$$Ax = b$$

with given vector $b \in \mathbb{R}^m$ and searched solution vector $x \in \mathbb{R}^n$ for $n \neq m$:

Case $n > m$: Here, there are more unknowns than equations and either there is no solution or infinitely many solutions in the solution set $\mathbb{L}$, where in the last case we have $\dim \mathbb{L} = \dim \ker A \geq n - m$.

Case $n < m$: There are fewer unknowns here than equations and there are none or exactly one solution in the case $\operatorname{rank} A = n$.

Here, $\ker A = \{x \in \mathbb{R}^n \mid Ax = 0\}$, and $\operatorname{rank} A$ is the number of linearly independent columns of A.

We return to square matrices and repeat some important theorems and definitions. Regarding the concept of eigenvalue or eigenvector:

Definition 2.2. *Let $A \in \mathbb{R}^{n\times n}$. A vector $v \in \mathbb{R}^n \backslash \{0\}$ is called eigenvector of A corresponding to the eigenvalue $\lambda \in \mathbb{R}$, if*

$$Av = \lambda v$$

//.

That is, an eigenvector v is only changed in its length by the mapping $v \mapsto Av$, but not in its direction.

Definition 2.3. *Let A be a square matrix. Then A is called singular if $det(A) = 0$, otherwise it is called regular.* //.

The following important Lemma holds true:

Proposition 2.4. *Let $A \in \mathbb{R}^{n\times n}$. Then A is regular, if one of the following equivalent conditions hold.*

(a) $\det A \neq 0$

(b) $\operatorname{rank} A = n$

(c) $\ker A = \{0\}$

(d) The columns of A form a Basis of $\mathbb{R}^n$

(e) The rows of A form a Basis of $\mathbb{R}^n$

(f) All eigenvalues differ from Zero, i.e. $0 \notin \sigma(A)$

(g) *A is invertible, i.e. there exists a Matrix $X \in \mathbb{R}^{n\times n}$, s.t. $AX = E$. Here, $E \in \mathbb{R}^{n\times n}$ denotes the identity Matrix.*

Each matrix $A \in \mathbb{R}^{m\times n}$ can also be viewed as a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^m$, where the image set is given by $\operatorname{im} A = \{Ax \mid x \in \mathbb{R}^n\}$. In this spirit, the matrix representation (with respect to the standard basis) of the composition of two matrices is given by the matrix product

$$(AB)x = A(Bx).$$

Let $A = (a_{ij})$ be an $(m \times n)$-Matrix. The $(n \times m)$-Matrix $B = (b_{ji})$ such that $b_{ji} = a_{ij}$ is called the transpose of A, denoted by A^T. A matrix A is said to be symmetric if it is equal to its transpose, i.e. if $A^T = A$. A symmetric matrix is necessarily a square matrix. To state an important property of symmetric matrices (the spectral theorem), we first introduce some usefull notation: For $i, j = 1, \dots, n$, we let (Kronecker-delta)

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

For $a_1, \dots, a_n \in \mathbb{R}^n$ we define $[a_1, \dots, a_n] \in \mathbb{R}^{n\times n}$ to be the matrix with ith column a_i.

Definition 2.5 (Orthogonal matrix). *A matrix $Q \in \mathbb{R}^{n\times n}$ is called orthogonal, if*

$$Q^T Q = E_n.$$

We let

$$\mathcal{O}_n := \{Q \in \mathbb{R}^{n\times n} \mid Q^T Q = E_n\}. \qquad /\!/.$$

Unless otherwise stated, $\mathbb{R}^n$ is understood to be equipped with the standard scalar product $\langle \cdot\,, \cdot \rangle$.

Lemma 2.6. *Let $Q \in \mathbb{R}^{n\times n}$ and $\langle \cdot\,, \cdot \rangle$ be the canonical inner product on $\mathbb{R}^n$ with induced norm $\|\cdot\|$. Then the following statements are equivalent*

(a) *Q is orthogonal*

(b) *Q^T is orthogonal*

(c) *Q is invertible and $Q^{-1} = Q^T$*

(d) *The columns of Q form an orthonormal Basis of $\mathbb{R}^n$ with respect to $\langle \cdot\,, \cdot \rangle$*

(e) *The rows of Q form an orthonormal Basis of $\mathbb{R}^n$ with respect to $\langle \cdot\,, \cdot \rangle$*

(f) *$\langle Qx, Qy \rangle = \langle x, y \rangle \ \forall x, y \in \mathbb{R}^n$*

(g) *$\|Qx\| = \|x\| \ \forall x \in \mathbb{R}^n$*

For any diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix}$$

we also write $\operatorname{diag}(\lambda_1, \dots, \lambda_n) := D$.

Theorem 2.7 (Spectral theorem). *Let $A \in \mathbb{R}^{n\times n}$ be symmetric. Then the following two equivalent statements hold:*

(a) *There exists an orthonormal Basis of $\mathbb{R}^n$ consisting of eigenvectors of A.*

(b) There exists $V = [v_1, \dots, v_n] \in \mathcal{O}_n$ *and* $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$ *such that*

$$A = VDV^T,$$

where $Av_i = \lambda_i v_i$.

We now look at a few more details on the determination of eigenvalues and properties of the characteristic polynomial.

Lemma 2.8. *The eigenvalues of a matrix* $A \in \mathbb{R}^n$ *are exactly the zeros of the characteristic polynomial*

$$\chi_A(\lambda) = \det(A - \lambda E).$$

The characteristic polynomial has the form

$$\chi_A(\lambda) = (-1)^n \lambda^n + (-1)^{n-1} \operatorname{trace}(A) \lambda^{n-1} + \dots + \det A,$$

where $\operatorname{trace}(A) = \sum_{i=1}^n a_{ii}$ *equals the trace of* $A = (a_{ij})$. *Since similar matrices have the same characteristic polynomial, i.e.* $\chi_A(\lambda) = \chi_{S^{-1}AS}$ *for all regular* S, *the coefficients of* χ_A *are invariants with respect to similarity transformations. In particular, we obtain*

$$\operatorname{trace}(A) = \operatorname{trace}(S^{-1}AS).$$

It also follows from the last equation that the trace of a matrix is the sum of all eigenvalues of A *weighted with (algebraic) multiplicity* $\mu(\chi_A, \lambda)$*:*

$$\operatorname{trace}(A) = \sum_{\lambda \in \sigma(A)} \mu(\chi_A, \lambda)\lambda.$$

Moreover, the determinant of a matrix is the product of the eigenvalues counted with multiplicity, i.e.

$$\det(A) = \prod_{\lambda \in \sigma(A)} \lambda^{\mu(\chi_A, \lambda)}.$$

Here, $\sigma(A)$ *denotes the set of all eigenvalues of* A*, the so-called spectrum of* A.

Proof. Only a few important steps should be shown: The definition of eigenvalues or eigenvectors immediately implies that the eigenvalues are the zeros of the characteristic polynomial, more precisely: If $v \neq 0$ is an eigenvector corresponding to the eigenvalue λ, then it holds

$$Av = \lambda v \quad \Leftrightarrow \quad (A - \lambda E)v = 0 \quad \Leftrightarrow \quad \det(A - \lambda E) = 0.$$

That similarity transformations do not change the characteristic polynomial and thus its coefficients follows immediately for regular S from

$$\chi_A(\lambda) = \det(A - \lambda E) = \det\left(S^{-1}(A - \lambda E)S\right) = \chi_{S^{-1}AS}(\lambda).$$

The last two statements about the relationship between the trace or the determinant and the eigenvalues then follow from the fact that every matrix can be brought into a triangular shape with the eigenvalues on the diagonal by means of a similarity transformation. ■

Definition 2.9 (positive definite Matrix). *A symmetric matrix* A *is called positive definite if*

$$\langle x, Ax \rangle = x^T Ax > 0$$

for all $x \in \mathbb{R}^n \backslash \{0\}$. //.

Lemma 2.10. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Then*

(a) All eigenvalues of A are strictly positive.

(b) All diagonal elements of $A = (a_{ij})$ are strictly positive, i.e.

$$a_{ii} > 0 \quad \text{for } i = 1, \ldots, n.$$

Proof. Exercise. ■

Definition 2.11 (normed linear Space). *Let V be a $\mathbb{K}$-vector space. A map $\|\cdot\|: V \to \mathbb{R}$ that satisfies the following conditions for all $x, y \in V$ and $\alpha \in \mathbb{K}$*

(a) $\|x\| = 0 \;\Rightarrow x = 0$

(b) $\|\alpha x\| = |\alpha|\,\|x\|$

(c) $\|x + y\| \le \|x\| + \|y\|$

is called (vector) norm on V and the pair $(V, \|\cdot\|)$ is said to be a ***normed (vector/linear) space****.* //.

Note that for $\alpha = 0$ in (b) we follows that $x = 0$ implies $\|x\| = 0$. Furthermore, letting $y = -x$ in (c) we obtain from (a)-(c) that $\|x\| > 0$ for all $x \ne 0$. Frequently used norms in $\mathbb{R}^n$ are the p-norms (also: l_p-norm)

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \quad 1 \le p < \infty. \tag{2.2}$$

Important special cases are the **Euclidean Norm**

$$\|x\|_2 = \sqrt{x_1^2 + \ldots + x_n^2} = \langle x, x\rangle^{\frac{1}{2}},$$

the l_1-norm (also: **Manhatten norm**)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

and the **maximum norm** (also: l_∞-norm)

$$\|x\|_\infty = \max_{1 \le i \le n} |x_i|.$$

Note that when passing to the limit $p \to \infty$ in Equation 2.2 we obtain the maximum norm

$$\|x\|_\infty = \lim_{p\to\infty} \|x\|_p.$$

Assuming $\|x\|_\infty = 1$ the latter follows from

$$1 \le \|x\|_p = \left(|x_1|^p + \ldots + |x_n|^p\right)^{\frac{1}{p}} \le n^{\frac{1}{p}}$$

letting $p \to \infty$. Figure 2.1 shows the respective l_p unit spheres for $p = 1$ (red), $p = 2$ (orange), $p = 4$ (green), $p = 7$ (yellow) and $p = \infty$ (blue).

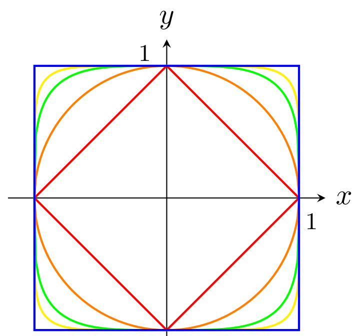


Fig. 2.1.: l_p unit spheres

Lemma 2.12. *All norms of $\mathbb{R}^n$ are equivalent, i.e. if $\|\cdot\|_*$ and $\|\cdot\|_{**}$ are two norms in $\mathbb{R}^n$, then there exist $\alpha, \beta > 0$ independently of x, so that for all $x \in \mathbb{R}^n$*

$$\alpha\|x\|_* \leq \|x\|_{**} \leq \beta\|x\|_*.$$

Norms are commonly used for measuring distances. In this context, we are also interested in a distance measure or a suitable norm for matrices, which allows us to put the value $\|Ax\|$ of a linear mapping A in relation to the value $\|x\|$. The matrix norm $\|A\|$ to be defined will make it possible to estimate the lengthening or shortening of the vector x after conversion to Ax as follows

$$\|Ax\| \leq \|A\|\|x\|. \tag{2.3}$$

For this, the matrix norm $\|A\|$ needs to be compatible with the vector space norm $\|\cdot\|$ in a way to be specified. This is the content of the following definitions.

Definition 2.13 (matrix norm). *A map $\|\cdot\|\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ is called* ***matrix norm*** *if the following conditions are satisfied for all $A, B \in \mathbb{R}^{n\times n}$ and $\alpha \in \mathbb{R}$:*

(a) $\|A\| = 0 \;\Rightarrow A = 0$

(b) $\|\alpha A\| = |\alpha|\,\|A\|$

(c) $\|A + B\| \leq \|A\| + \|B\|$

(d) $\|A \cdot B\| \leq \|A\| \cdot \|B\|$. //.

As for vector norms it follows that $\|A\| \geq 0$, and $A = 0$ implies $\|A\| = 0$.

Definition 2.14 (induced matrix norm). *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. Then*

$$\|A\| := \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

is called the ***induced matrix norm*** *(also: operator norm) induced by $\|\cdot\|$.* //.

The operator norm is a norm in the sense of Definition 2.13. A matrix norm $\|A\|$ is said to be **compatible** with a vector norm $\|x\|$, if Equation 2.3 holds for all $A \in \mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$. The induced matrix norm is compatible with the underlying vector norm. All (not only the induced) matrix norms on $\mathbb{R}^n$ are equivalent.

Example 2.15. For the maximum norm $\|\cdot\|_\infty$ we obtain the so-called **(maximum absolute) row sum norm**

$$\|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}|.$$

This can be seen, e.g., as follows: By definition

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

For arbitrary $x \in \mathbb{R}^n$ with $\|x\|_\infty = 1$ it then follows

$$\begin{aligned}\|Ax\|_\infty &= \max_{i=1,\dots,n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \\ &\leq \max_{i=1,\dots,n} \sum_{j=1}^{n} |a_{ij}| \underbrace{|x_j|}_{\leq 1}\end{aligned}$$

$$\leq \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|.$$

Conversely, let i_0 be an index such that

$$\sum_{j=1}^{n} |a_{i_0 j}| = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|.$$

By a suitable choice of $x_j = \pm 1$ for $j = 1, \ldots, n$, a vector x can be chosen such that

$$\max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| = \max_{i=1,\ldots,n} \sum_{j=1}^{n} a_{ij} x_j.$$

Thus

$$\|A\|_\infty = \max_{\|x\|_\infty = 1} \|Ax\|_\infty \geq \sum_{j=1}^{n} |a_{i_0 j}| = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|$$

and the assertion follows. //.

Example 2.16. For the Euclidean norm $\|\cdot\|_2$ one obtains the **spectral norm**

$$\|A\|_2 = \max_{\|x\|_1 = 1} \langle Ax, Ax \rangle^{\frac{1}{2}} = \max_{\|x\|_1 = 1} (x^T A^T A x)^{\frac{1}{2}} = \sqrt{\rho(A^T A)}$$

as the induced matrix norm, where for any matrix $B \in \mathbb{R}^{n \times n}$

$$\rho(B) := \max_{\lambda \in \sigma(B)} |\lambda|$$

is called the **spectral radius** of B. Any $\lambda \in \sigma(B)$ with $|\lambda| = \rho(B)$ is called a largest absolute eigenvalue (alternatively: dominant eigenvalue) of B. Note that all eigenvalues of $A^T A$ are real, since $A^T A$ is symmetric. Details in the exercises. //.

In many applications it is very expensive to calculate the spectral norm of a (large) matrix. One is then often satisfied with the Frobenius norm, which for $A \in \mathbb{R}^{n \times n}$ is defined as follows:

$$\|A\|_F := \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}}.$$

The Frobenius norm is a matrix norm which is not induced but which is compatible with the vector norm $\|\cdot\|_2$, i.e.

$$\|Ax\|_2 \leq \|A\|_F \|x\|_2.$$

Hence the Forbenius norm is an upper bound for the spectral norm, i.e.

$$\|A\|_2 \leq \|A\|_F$$

for all $A \in \mathbb{R}^{n \times n}$. The Frobenius norm, which is much easier to calculate, can therefore be used to estimate the spectral norm.

2.3. LU decomposition

First we consider the Gauss algorithm as it is performed by hand using Example 2.1. The system of equations to be solved there reads in compact notation

$$Ax = b \quad \Leftrightarrow \quad A \mid b \quad \Leftrightarrow \quad \begin{array}{ccc|c} 24 & 48 & 72 & 76800 \\ 1 & 1 & 1 & 1700 \\ 1 & 4 & 2 & 2850 \end{array}$$

The matrix shall now be transformed into a triangular shape. This can be done as follows:

$$\begin{array}{ccc|c} 24 & 48 & 72 & 76800 \\ 1 & 1 & 1 & 1700 \\ 1 & 4 & 2 & 2850 \end{array} \quad \xrightarrow[\mathrm{III}-\frac{1}{24}\mathrm{I}]{\mathrm{II}-\frac{1}{24}\mathrm{I}} \quad \begin{array}{ccc|c} 24 & 48 & 72 & 76800 \\ 0 & -1 & -2 & -1500 \\ 0 & 2 & -1 & -350 \end{array} \tag{2.4}$$

and further

$$\begin{array}{ccc|c} 24 & 48 & 72 & 76800 \\ 0 & -1 & -2 & -1500 \\ 0 & 2 & -1 & -350 \end{array} \quad \xrightarrow{\mathrm{III}+2\mathrm{I}} \quad \begin{array}{ccc|c} 24 & 48 & 72 & 76800 \\ 0 & -1 & -2 & -1500 \\ 0 & 0 & -5 & -3350 \end{array} \tag{2.5}$$

Then one can quickly obtain the solution by **backward substitution**: $x_3 = 670$ immediately follows from III. From this one gets x_2 by substitution of x_3 in II, and finally one gets x_1 by substitution of x_2 and x_3 in I. The solution is

$$x = \begin{pmatrix} 870 \\ 160 \\ 670 \end{pmatrix}.$$

If you now have different inhomogeneities, as in our example, for instance different inventory, it makes sense to remember the transformation steps up to the triangular form in order to obtain the solution immediately for all possible right-hand sides by backward substitution. Therefore, the steps of the elimination algorithm should be recorded mathematically. This leads to a decomposition of the original matrix, the so-called **LU decomposition** (also: LU factorization):

Let us look at the first step in Equation 2.4 with the first two elementary row transformations. These are recorded exactly by the Matrix L_1 (check this!): It is

$$L_1 A = \begin{pmatrix} 1 & 0 & 0 \\ -1/24 & 1 & 0 \\ -1/24 & 0 & 1 \end{pmatrix} \begin{pmatrix} 24 & 48 & 72 \\ 1 & 1 & 1 \\ 1 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 24 & 48 & 72 \\ 0 & -1 & -2 \\ 0 & 2 & -1 \end{pmatrix}.$$

Here, the matrix

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ -1/24 & 1 & 0 \\ -1/24 & 0 & 1 \end{pmatrix}$$

is called **Frobenius matrix**. It describes, when muliplied from the left, exactly the two elementary row transformations of the first step. Similarly, there is a matrix L_2 for the second step in Equation 2.5, which carries out the row transformation III + 2II when multiplied from the left. With

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix}$$

it follows that

$$L_2 L_1 A = U := \begin{pmatrix} 24 & 48 & 72 \\ 0 & -1 & -2 \\ 0 & 0 & -5 \end{pmatrix}.$$

Multiplying the last equation with $(L_2L_1)^{-1}$ from the left then results in a decompostition of A into a lower triangular matrix L and a upper triangular matrix U:

$$A = \underbrace{(L_2L_1)^{-1}}_{=:L} U.$$

In the general case of an $(n \times n)$-matrix A with

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix},$$

the matrix L_1 for the first step of the LU decomposition (elimination of the 1st column) takes on the following form:

$$L_1 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -l_{21} & 1 & \ddots & & \vdots \\ -l_{31} & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -l_{n1} & 0 & \cdots & 0 & 1 \end{pmatrix},$$

where

$$l_{i1} := \frac{a_{i1}}{a_{11}}.$$

With this matrix L_1 it follows

$$A^{(1)} := L_1 A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}.$$

Here, the superscript 1 in brackets indicates the values changed by the first step.
This procedure can now be continued if all diagonal elements $a_{kk}^{(k-1)}$ are non-zero. The latter is assumed below. In the kth step, the kth column below the diagonal is then to be eliminated after the first $k-1$ columns below the diagonal have already been eliminated in the first $k-1$ steps. To do so, the kth step is given by

$$A^{(k)} := L_k A^{(k-1)}$$

with

$$L_k = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ & & -l_{n,k} & & & 1 \end{pmatrix} \tag{2.6}$$

and

$$l_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \quad \text{for } i > k.$$

The matrix L_k differs from the identity matrix only in the kth column. All in all, one has reached the goal after $(n-1)$ steps, since the matrix $A^{(n-1)}$ has an upper triangular form - all $(n-1)$ columns to be cleaned up have been eliminated. But now we have

$$U = A^{(n-1)} = L_{n-1} \cdot \ldots \cdot L_1 A, \tag{2.7}$$

where U is the upper (right) triangular matrix

$$U = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nn} \end{pmatrix}.$$

If one now wants to solve the system of equations $Ax = b$, the corresponding transformations must also be applied to the inhomogeneity, since the corresponding row transformations are also carried out with b:

$$y = b^{(n-1)} = L_{n-1} \cdot \ldots \cdot L_1 b.$$

Analogous to the example above, the system

$$Ux = b$$

can then be quickly solved by backward substitution.

We will see, however, that the matrices L_k are benign in a certain sense and one can avoid the transformation of the inhomogeneity at b, since one can immediately state a decomposition of A, from which the solution can be determined just as quickly. The representation

$$A = (L_{n-1} \cdot \ldots \cdot L_1)^{-1} U = (L_1^{-1} \cdot \ldots \cdot L_{n-1}^{-1}) U \tag{2.8}$$

follows immediately from equation (2.7). Furthermore, one knows L_k^{-1} immediately without calculation, because one calculates that

$$L_k^{-1} = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & l_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ & & l_{n,k} & & & 1 \end{pmatrix},$$

i.e. the inverse matrix of L_k is obtained simply by changing the sign in the kth column below 1, compare with (2.6). Multiplying out the L_k^{-1} on the right-hand side in (2.8) is also very easy, since you just have to add the entries in the matrices below the diagonals, so you can simply copy the l_{ik}:

$$L := L_1^{-1} \cdot \ldots \cdot L_{n-1}^{-1} = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \ldots & l_{n,n-1} & 1 \end{pmatrix}.$$

In summary, the following Proposition holds.

Proposition 2.17. *Let $A \in \mathbb{R}^{n \times n}$. Are a_{11} and the Diagonal elements $a_{kk}^{(k-1)}$ arising from column elimination non-zero, Gauss elimination produces a LU decomposition*

$$A = LU$$

of A, where

$$L = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & \ddots & \ddots & \\ l_{n1} & \ldots & l_{n,n-1} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{12} & \cdots & u_{1n} \\ & & \ddots & \vdots \\ & & & u_{nn} \end{pmatrix},$$

*i.e. L is a lower triangular matrix with ones on the diagonal (=**unipotent lower triangular matrix**) and U is an upper triangular matrix. The system of equations*

$$Ax = LUx = b$$

can then be quickly solved using the following intermediate steps

(a) *Solve $Ly = b$ by **forward substitution**: here one obtains y_1 immediately and then in turn by substitution*
$$y_2 \to y_3 \to \ldots \to y_n.$$

(b) *Solve $Ux = y$ by backwards substitution: here one obtains x_n immediately and then one after the other by substitution*
$$x_{n-1} \to x_{n-2} \to \ldots \to x_1.$$

The vector $x = (x_1, \ldots, x_n)^T$ is then the solution we are looking for, because it is

$$Ax = LUx = Ly = b.$$

The (asymptotic) numerical effort, i.e. the number of essential floating-point operations (i.e. multiplications/divisions) of the LU decomposition is $\sim \frac{1}{3}n^3$ and that of the forward and backward substitution $\sim n^2$.

Proof. Only the estimate for the effort has to be shown. The cost of the LU decomposition is

$$\sum_{j=2}^{n} j(j-1) = \frac{n^3 - n}{3} \sim \frac{n^3}{3}$$

and the cost of forward and backward substitution is

$$2\sum_{j=1}^{n} j = 2\cdot\frac{n(n+1)}{2} \sim n^2. \qquad \blacksquare$$

With the LU decomposition, the determinant of the matrix A can be calculated immediately:

$$\det(A) = \det(LU) = 1\cdot\det(U) = \prod_{j=1}^{n} u_{ij}.$$

Example 2.18. Determine the LU decomposition of

$$A = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 8 & 16 \\ 2 & 4 & 2 \end{pmatrix}.$$

//.

Step 1: Eliminating the 1st column: we write A directly as a decomposition (i.e. we directly use L_1^{-1}):

$$A = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 8 & 16 \\ 2 & 4 & 2 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}}_{=L_1^{-1}} \underbrace{\begin{pmatrix} 2 & 2 & 2 \\ 0 & 4 & 12 \\ 0 & 2 & 0 \end{pmatrix}}_{=:A^{(1)}}.$$

Step 2: Eliminating the 2nd column:

$$A = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & \frac{1}{2} & 1 \end{pmatrix}}_{=L_1^{-1}} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}}_{=L_2^{-1}} \underbrace{\begin{pmatrix} 2 & 2 & 2 \\ 0 & 4 & 12 \\ 0 & 0 & -6 \end{pmatrix}}_{=:A^{(2)}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & \frac{1}{2} & 1 \end{pmatrix}}_{=L_1^{-1}L_2^{-1}} \underbrace{\begin{pmatrix} 2 & 2 & 2 \\ 0 & 4 & 12 \\ 0 & 0 & -6 \end{pmatrix}}_{=:A^{(2)}} =: LU.$$

The elimination steps can be combined into a short sheme:

$$A = \begin{pmatrix} 2 & 2 & 2 \\ 4 & 8 & 16 \\ 2 & 4 & 2 \end{pmatrix} = \left(\begin{array}{c|cc} 2 & 2 & 2 \\ \hline 2 & 4 & 12 \\ 1 & 2 & 0 \end{array}\right) = \left(\begin{array}{cc|c} 2 & 2 & 2 \\ 2 & 4 & 12 \\ \hline 1 & \frac{1}{2} & -6 \end{array}\right).$$

2.4. LU decomposition with permutations

In the last section, an essential requirement for the LU decomposition was the condition that all diagonal elements

$$a_{kk}^{(k-1)}$$

in the algorithm do not vanish. From this property it follows together with $A = LU$ that $\det(A) \neq 0$; i.e. A must be regular for the LU decomposition to work. However, there are also regular matrices for which the LU decomposition does not work without additional procedures:

Example 2.19. Show that the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

is regular but cannot be written in the form $A = LU$. //.

In this example, however, the decomposition is obvious after swapping rows, since then the identity matrix is created.

Lemma 2.20. *Let A be regular and after the $(k-1)$th step of the LU decomposition*

$$A^{(k-1)} = L_{k-1} \cdot \ldots \cdot L_1 \cdot A \tag{2.9}$$

assume that $a_{kk}^{(k-1)} = 0$ holds, i.e. a further step would not be possible without swapping rows. Then in the kth column below the diagonal there is always at least one element $a_{ik}^{(k-1)} \neq 0$ $(i > k)$, i.e. a successful row swapping is always possible.

Proof. Suppose the statement is false, i.e. there exists a number k such that $a_{ik}^{(k-1)} = 0$ for all $i \geq k$, which means

$$A^{(k-1)} = \begin{pmatrix} a_{11}^{(k-1)} & \cdots & a_{1,k-1}^{(k-1)} & a_{1,k}^{(k-1)} & a_{1,k+1}^{(k-1)} & \cdots & a_{1,n}^{(k-1)} \\ 0 & \ddots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & a_{n,k+1}^{(k-1)} & \cdots & a_{nn}^{(k-1)} \end{pmatrix}.$$

But in this case the kth column could be represented by a linear combination of the first $(k-1)$ columns. This would imply the singularity of $A^{(k-1)}$ and thus, according to (2.9), since all L_j are regular, entail the singularity of A. A contradiction to our assumption. ■

So in general, one will have to implement row permutations in the LU algorithm. A row swap in a matrix is done by left multiplication by a permutation matrix P of the following form:

$$P = P_{ij} = \begin{array}{cc} \begin{pmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & 0 & & & & 1 & & & \\ & & & & 1 & & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & 1 & & & & 0 & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix} & \begin{matrix} \\ \\ \\ \leftarrow i \\ \\ \\ \\ \leftarrow j \\ \\ \\ \\ \end{matrix} \\ \begin{matrix} \qquad\qquad\qquad \uparrow & \qquad\qquad\qquad \uparrow \\ \qquad\qquad\qquad i & \qquad\qquad\qquad j \end{matrix} & \end{array}$$

with $P_{ii} \coloneqq E$. The matrix P results from the identity matrix by interchanging the ith and jth column or row. The following holds:

(a) Multiplication by P from the left swaps rows i and j of A,

(b) Multiplication by P from the right swaps columns i and j of A.

Clearly, Permutation matrices are involutory and symmetric: $P = P^{-1}$ and $P = P^T$. If one now introduces a permutation matrix P_k or $P_k = E$ (if no row swapping is necessary) in each step of the LU decomposition, one obtains

$$U = L_{n-1}P_{n-1} \cdot \ldots L_1 P_1 A$$

or

$$A = \underbrace{(L_{n-1}P_{n-1} \cdot \ldots L_1 P_1)^{-1}}_{=:\widetilde{L}} U. \tag{2.10}$$

In general, however, the resulting $\widetilde{L}$ is no longer a lower triangular matrix. However, the triangular shape can be regained with a trick: Left-multiplying Equation 2.10 by $P \coloneqq P_1 \cdot \ldots \cdot P_{n-1}$ yields

$$\begin{aligned} PA &= \overbrace{P_{n-1} \cdot \ldots \cdot P_2 \cdot P_1}^{=P} \underbrace{P_1 (P_1}_{=E} L_1^{-1} P_2 L_2^{-1} \cdot \ldots \cdot P_{n-1} L_{n-1}^{-1}) U \\ &= \underbrace{P_{n-1} \cdot \ldots \cdot P_3 P_2 L_1^{-1} P_2 L_2^{-1} P_3 L_3^{-1} \cdot \ldots \cdot P_{n-1} L_{n-1}^{-1}}_{=:L} U. \end{aligned} \tag{2.11}$$

We now show that L represents a unipotent lower triangular matrix. In what follows, a matrix of the form

$$\begin{array}{cc} \begin{matrix} \\ \\ \\ k+1 \rightarrow \\ \\ \\ \end{matrix} & \begin{pmatrix} 0 & & & & \\ * & \ddots & & & \\ \vdots & \ddots & \ddots & & \\ * & \cdots & * & \ddots & \\ \vdots & & \vdots & & \ddots \\ * & \cdots & * & & & 0 \end{pmatrix} =: N_k \\ & \begin{matrix} \qquad\quad \uparrow \\ \qquad\quad k \end{matrix} \end{array}$$

will be denoted by N_k, where each $*$ stands for an arbitrary real number. Thus we can uniquely write

$$L_1^{-1} = \begin{pmatrix} 1 & & & \\ l_{21} & 1 & & \\ \vdots & & \ddots & \\ l_{n1} & & & 1 \end{pmatrix} =: N_1 + E =: M_1. \tag{2.12}$$

Note, that for each P_k in (2.11) we have $P_k = P_{k,j_k}$ with $j_k \geq k$, i.e. P_k arises from E by swapping the rows k and $j_k \geq k$. For $1 \leq k \leq n-2$ let

$$M_{k+1} \coloneqq P_{k+1} M_k P_{k+1} L_{k+1}^{-1}.$$

We now show by induction that for $1 \leq k \leq n-1$ we have

$$M_k = N_k + E. \tag{2.13}$$

For $k = 1$ this follows from (2.12). Assume that Equation 2.13 holds true for some arbitrary but fixed k with $1 \leq k \leq n-2$. Then

$$\begin{aligned} M_{k+1} &= P_{k+1} M_k P_{k+1} L_{k+1}^{-1} \\ &\overset{(2.13)}{=} P_{k+1}(N_k + E) P_{k+1} L_{k+1}^{-1} \\ &= P_{k+1} \underbrace{N_k P_{k+1}}_{=N_k} L_{k+1}^{-1} + \underbrace{P_{k+1} E P_{k+1}}_{=E} L_{k+1}^{-1} \\ &= P_{k+1} \underbrace{N_k L_{k+1}^{-1}}_{=N_k} + L_{k+1}^{-1} \\ &= P_{k+1} N_k + L_{k+1}^{-1} \\ &= N_k' + L_{k+1}^{-1} \\ &= N_{k+1} + E \end{aligned} \tag{2.14}$$

completing the induction step. Letting $k = n-1$ in Equation 2.13, we see that M_{n-1} is a unipotent lower triangular matrix and (2.11) now reads

$$PA = M_{n-1} U = LU.$$

Proposition 2.21. *Let $A \in \mathbb{R}^{n \times n}$ be regular. Then there exist permutations P_k, $1 \leq k \leq n-1$, such that with $P \coloneqq P_1 \cdot \ldots \cdot P_{n-1}$ it holds*

$$PA = LU,$$

where L is a unipotent lower triangular matrix and U is an upper triangular matrix. Moreover, for fixed P the matrices L and U are uniquely determined.

Proof. Only the claim of uniqueness remains to be shown. In the next three auxiliary steps, we perform induction on the dimension n. For the sake of brevity we abbreviate:

- t: triangular
- r: regular
- u: upper
- l: lower

e.g. rut matrix means: regular upper triangular matrix. Let

$$U = \left(\begin{array}{c|c} \alpha & r \\ \hline & \widetilde{U} \end{array}\right) \in \mathbb{R}^{n\times n}, \quad U' = \left(\begin{array}{c|c} \alpha' & r' \\ \hline & \widetilde{U}' \end{array}\right) \in \mathbb{R}^{n\times n} \tag{2.15}$$

both be (unipotent) ut matrices, where $r, r' \in \mathbb{R}^{1\times(n-1)}$. Hence $\widetilde{U}$ and $\widetilde{U}'$ also have (unipotent) ut shape.

Step 1: The product

$$UU' = \left(\begin{array}{c|c} \alpha\alpha' & \alpha r' + r\widetilde{U}' \\ \hline & \widetilde{U}\widetilde{U}' \end{array}\right) \tag{2.16}$$

is also a (unipotent) ut matrix, since this holds for $\widetilde{U}\widetilde{U}'$ by induction hypothesis.

Step 2: Let $L, L' \in \mathbb{R}^{n\times n}$ both be (unipotent) lt matrices. From step 1 it follows that the Product

$$(L')^T L^T = U$$

is a (unipotent) ut matrix and hence LL' is a (unipotent) lt matrix by transposition.

Step 3: Now we show that for a (unipotent) rut matrix U, the inverse is also a (unipotent) rut matrix: Since $\det U = \alpha \det \widetilde{U} \neq 0$, it follows that $\alpha \neq 0$ and $\widetilde{U}$ is a (unipotent) rut matrix. Note that $\widetilde{U}^{-1} \in \mathbb{R}^{(n-1)\times(n-1)}$ is a (unipotent) rut matrix by induction hypothesis and hence

$$U' := \left(\begin{array}{c|c} \frac{1}{\alpha} & -\frac{1}{\alpha} r\widetilde{U}^{-1} \\ \hline & \widetilde{U}^{-1} \end{array}\right)$$

is also a (unipotent) rut matrix. Using $\alpha' = \frac{1}{\alpha}$, $r' = \frac{1}{\alpha} r\widetilde{U}^{-1}$, $\widetilde{U}' = \widetilde{U}^{-1}$ yields

$$UU' = \left(\begin{array}{c|c} \alpha\alpha' & \alpha r' + r\widetilde{U}' \\ \hline & \widetilde{U}\widetilde{U}' \end{array}\right) = E,$$

we see that $U' = U^{-1}$, completing the induction step.

Step 4: $L \in \mathbb{R}^{n\times n}$ is a (unipotent) lower triangular matrix $\Rightarrow L^T$ is a (unipotent) upper triangular matrix $\overset{\text{step 3}}{\Longrightarrow} (L^T)^{-1} = (L^{-1})^T$ is a (unipotent) upper triangular matrix $\Rightarrow L^{-1}$ is also a (unipotent) upper triangular matrix.

Step 5: Let $PA = LU = L'U'$ be two LU decompositions, where L, L' both are unipotent lt matrices and U, U' are rut matrices. This implies

$$L^{-1}L' = U\big(U'\big)^{-1}. \tag{2.17}$$

Since the left-hand side of Equation 2.17 is a unipotent lt matrix and the right-hand side is a rut matrix it follows that

$$L^{-1}L' = U\big(U'\big)^{-1} = E$$

and hence $L = L'$ and $U = U'$. ■

The practical implementation of the decomposition with row permutations now works as follows. It is

$$Ax = b \quad \Leftrightarrow \quad PAx = Pb \quad \Leftrightarrow \quad LUx = Pb.$$

As mentioned above, the permutation matrices P_k swap the kth row with the row $j_k \geq k$. Therefore all permutations can also be stored in a single vector

$$p = (j_1, \ldots, j_{n-1})^T.$$

The considerations in (2.11) and (2.14) lead to the following practical calculation steps, where the matrices L and U will be bulid up step by step.

Step 1: Starting from the equation

$$A = A,$$

one has to determine the first permutation matrix P_1, which leads to

$$P_1 A = P_1 A.$$

Store P_1 in the vector p. After determining L_1 one obtains

$$P_1 A = \underbrace{L_1^{-1}}_{=M_1} \underbrace{L_1 P_1 A}_{=A^{(1)}}.$$

Step $k+1$: Assume we have the decomposition $P_k \cdot \ldots \cdot P_1 A = M_k A^{(k)}$. Determine P_{k+1}, $P_{k+1} A^{(k)}$ and the next component of the permutation vector p. Compute $P_{k+1} M_k P_{k+1}$ by swapping rows $k+1$ and $j_{k+1} \geq k+1$ in N_k leading to

$$P_{k+1} \cdot \ldots \cdot P_1 A = P_{k+1} M_k P_{k+1} P_{k+1} A^{(k)}.$$

Determine L_{k+1} and replace the $(k+1)$th column of $P_{k+1} M_k P_{k+1}$ with the $(k+1)$th column of L_{k+1}^{-1} to get

$$P_{k+1} \cdot \ldots \cdot P_1 A = P_{k+1} M_k P_{k+1} L_{k+1}^{-1} L_{k+1} P_{k+1} A^{(k)}.$$

For some matrices the following is useful (column or partial pivoting): In the elimination step $A^{(k-1)} \to A^{(k)}$ choose a (e.g. the smallest) $j_k \in \{k, \ldots, n\}$ such that

$$|a_{j_k,k}^{(k-1)}| \geq |a_{j,k}^{(k-1)}|$$

for all $j \in \{k, \ldots, n\}$, i.e. swap rows k and j_k.

Example 2.22. Determine the LU decomposition (with column pivoting) of the matrix

$$A = \begin{pmatrix} 0 & -2 & 2 & 1 \\ -2 & -4 & 5 & -7 \\ 6 & 12 & -18 & 24 \\ 3 & 10 & -11 & 18 \end{pmatrix}.$$

Also give the corresponding permutations P and P_k as a vector p. Finally, solve the system of equations

$$Ax = b, \quad b = \begin{pmatrix} 8 \\ -2 \\ 6 \\ 7 \end{pmatrix}$$

using the decomposition.

(1) Swap rows one and three, i.e. $p = (3)$, so that

$$P_1 A = \begin{pmatrix} 6 & 12 & -18 & 24 \\ -2 & -4 & 5 & -7 \\ 0 & -2 & 2 & 1 \\ 3 & 10 & -11 & 18 \end{pmatrix}$$

(2) Determine L_1. We use L_1^{-1} again to be able to state the decomposition straight away:

$$P_1 A = \underbrace{\begin{pmatrix} 1 & & & \\ -1/3 & 1 & & \\ 0 & 0 & 1 & \\ 1/2 & 0 & 0 & 1 \end{pmatrix}}_{=L_1^{-1}} \underbrace{\begin{pmatrix} 6 & 12 & -18 & 24 \\ 0 & 0 & -1 & 1 \\ 0 & -2 & 2 & 1 \\ 0 & 4 & -2 & 6 \end{pmatrix}}_{L_1 P_1 A}$$

(3) Swap rows two and four, i.e. $p = (3, 4)$, so that

$$P_2 P_1 A = \underbrace{\begin{pmatrix} 1 & & & \\ 1/2 & 1 & & \\ 0 & 0 & 1 & \\ -1/3 & 0 & 0 & 1 \end{pmatrix}}_{=P_2 L_1^{-1} P_2} \underbrace{\begin{pmatrix} 6 & 12 & -18 & 24 \\ 0 & 4 & -2 & 6 \\ 0 & -2 & 2 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}}_{P_2 L_1 P_1 A}$$

(4) Determine L_2.

$$P_2 P_1 A = \underbrace{\begin{pmatrix} 1 & & & \\ 1/2 & 1 & & \\ 0 & -1/2 & 1 & \\ -1/3 & 0 & 0 & 1 \end{pmatrix}}_{=P_2 L_1^{-1} P_2 L_2^{-1}} \underbrace{\begin{pmatrix} 6 & 12 & -18 & 24 \\ 0 & 4 & -2 & 6 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & -1 & 1 \end{pmatrix}}_{L_2 P_2 L_1 P_1 A}$$

(5) "Swap rows three and three" (no swapping necessary), i.e. $p = (3, 4, 3)$, so that

$$P_3 P_2 P_1 A = \underbrace{\begin{pmatrix} 1 & & & \\ 1/2 & 1 & & \\ 0 & 0 & 1 & \\ -1/3 & 0 & 0 & 1 \end{pmatrix}}_{=P_3 P_2 L_1^{-1} P_2 L_2^{-1} P_3} \underbrace{\begin{pmatrix} 6 & 12 & -18 & 24 \\ 0 & 4 & -2 & 6 \\ 0 & -2 & 2 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix}}_{P_3 L_2 P_2 L_1 P_1 A}$$

(6) Determine L_3.

$$P_3 P_2 P_1 A = \underbrace{\begin{pmatrix} 1 & & & \\ 1/2 & 1 & & \\ 0 & -1/2 & 1 & \\ -1/3 & 0 & -1 & 1 \end{pmatrix}}_{=P_3 P_2 L_1^{-1} P_2 L_2^{-1} P_3 L_3^{-1}} \underbrace{\begin{pmatrix} 6 & 12 & -18 & 24 \\ 0 & 4 & -2 & 6 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 5 \end{pmatrix}}_{L_3 P_3 L_2 P_2 L_1 P_1 A} = LU.$$

This determines the LU decomposition. If one now wants to solve the system $Ax = b$, one first has to apply the Permutation P on b:

$$PAx = LUx = Pb.$$

The permutation P is stored in the vector $p = (3, 4, 3)$, so in our case the components of b are swapped as follows:

$$b = \begin{pmatrix} 8 \\ -2 \\ 6 \\ 7 \end{pmatrix} \xrightarrow[\text{I}\leftrightarrow\text{III}]{P_1} \begin{pmatrix} 6 \\ -2 \\ 8 \\ 7 \end{pmatrix} \xrightarrow[\text{II}\leftrightarrow\text{IV}]{P_2} \begin{pmatrix} 6 \\ 7 \\ 8 \\ -2 \end{pmatrix} \xrightarrow[\text{III}\leftrightarrow\text{III}]{P_3=E} \begin{pmatrix} 6 \\ 7 \\ 8 \\ -2 \end{pmatrix} = Pb.$$

The system $LUx = Pb$ is then solved as before by successive forward and backward substitution. //.

2.5. Error analysis

As already indicated, matrix norms play an important role in determining the stability and error estimation for the numerical solution of a system of linear equations. In the following we want to briefly treat the input error for the right-hand side b and its propagation. The starting point is therefore the system

$$Ax = b,$$

where we assume A to be regular. The exact solution of this system is

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

If b now has an input error $b + \Delta b$, the new solution is

$$x + \Delta x = A^{-1}(b + \Delta b) = A^{-1}b + A^{-1}\Delta b,$$

i.e. the calculated solution $x + \Delta x$ contains the propagated error $\Delta x = A^{-1}\Delta b$. If the matrix norm $\|\cdot\|_M$ is compatible with the vector norm $\|\cdot\|$, then it follows

$$\begin{aligned} \frac{\|\Delta x\|}{\|x\|} &= \frac{\|A^{-1}\Delta b\|}{\|x\|} \leq \|A^{-1}\|_M \frac{\|\Delta b\|}{\|b\|} \frac{\|Ax\|}{\|x\|} \\ &\leq \|A^{-1}\|_M \|A\|_M \frac{\|\Delta b\|}{\|b\|}. \end{aligned} \tag{2.18}$$

The amplification of the relative error can thus be estimated by $\|A^{-1}\|_M\|A\|_M$.

Definition 2.23. *Let $A \in \mathbb{R}^{n\times n}$ be invertible. The number*

$$\operatorname{cond}_M(A) = \|A^{-1}\|_M\|A\|_M$$

is called the ***condition number*** *of the matrix A with respect to the norm $\|\cdot\|_M$.* //.

The estimate (2.18) makes it clear that the condition number describes the worst possible propagation of the input error when solving a system of linear equations. If $\|\cdot\|$ is induced by a vector norm, one can construct examples for b for which equality holds in (2.18).

One of the most frequently used norms is the Euclid norm $\|\cdot\|_2$, so that for a matrix A the value $\operatorname{cond}_2(A)$ is of particular importance. As we have already seen, for the determination of the spectral norm $\|A\|_2$ (second factor in $\operatorname{cond}_2(A)$) the largest eigenvalue of A^TA has to be calculated. However, the calculation of eigenvalues is very time-consuming and the inequality

$$\|A\|_2 \leq \|A\|_F$$

is often used instead; i.e. one determines only the much easier to calculate Frobenius norm as an upper bound for each of the two factors in $\operatorname{cond}_2(A)$.

For a symmetric matrix the calculation of $\operatorname{cond}_2(A)$ is a bit simpler: If $A \in \mathbb{R}^{n\times n}$ is regular and symmetric, then one can calculate the value of $\operatorname{cond}_2(A)$ directly from the eigenvalues of A.

Example 2.24. An example for an error amplification can already be found in the following simple system of equations: Consider $Ax = b$ with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0.001 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

under perturbation of the right hand side with

$$\Delta b = \begin{pmatrix} 0 \\ \varepsilon \end{pmatrix}.$$

The exact solution without perturbation is obviously $x = (1, 0)^T$ and the one with perturbation is $x + \Delta x = (1, 1000\varepsilon)^T$. It is easy to check that in this case ($\|\cdot\| = \|\cdot\|_2$) we have

$$\frac{\|\Delta x\|}{\|x\|} = 1000 \frac{\|\Delta b\|}{\|b\|}.$$

In fact, here we have $\mathrm{cond}_2(A) = 1000$, as we shall see in the example below. //.

Lemma 2.25. *Let $A \in \mathbb{R}^{n \times n}$ be regular and symmetric and let the real eigenvalues of A be in ascending order, i.e.*

$$0 < |\lambda_1| \leq |\lambda_2| \leq \ldots \leq |\lambda_n|.$$

Then

$$\mathrm{cond}_2(A) = \frac{|\lambda_n|}{|\lambda_1|}.$$

Proof. Exercise. ∎

Example 2.26. Determine cond_2 for the following matrices and compare the result with cond_F:

a) $E = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ b) $A = \begin{pmatrix} 1 & 0 \\ 0 & 0.001 \end{pmatrix}$ c) $A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$

In case a) it follows that

$$\mathrm{cond}_2(E) = \|E\|_2 \|E\|_2 = 1$$

and

$$\mathrm{cond}_F(E) = \|E\|_F \|E\|_F = \sqrt{3}\sqrt{3} = 3.$$

In case b) we have by Lemma 2.25

$$\mathrm{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 = 1 \cdot \frac{1}{0.001} = 1000$$

and

$$\begin{aligned} \mathrm{cond}_F(A) &= \|A\|_F \|A^{-1}\|_F \\ &= \sqrt{1 + 10^{-6}}\sqrt{1 + 10^6} \\ &\approx 1 \cdot 1000 \\ &= 1000. \end{aligned}$$

For c) we first calculate the eigenvalues of A according to Lemma 2.25. These are the zeros of

$$\chi_A(\lambda) = (1 - \lambda)^2 - 4,$$

i.e. $\lambda_1 = 3$ and $\lambda_2 = -1$. Thus

$$\mathrm{cond}_2(A) = \frac{3}{1} = 3.$$

For the Frobenius condition number, using $A^{-1} = \frac{1}{3}\begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}$, it follows that

$$\mathrm{cond}_F(A) = \|A\|_F \|A^{-1}\|_F = \sqrt{10}\frac{1}{3}\sqrt{10} = \frac{10}{3}.$$

//.

2.6. Cholesky decomposition

So far we have only considered arbitrary regular matrices $A \in \mathbb{R}^{n\times n}$. Frequently, however, one has even stronger assumptions than the regularity of a matrix - in many applications, for example, the matrices that occur are symmetric and positive definite. Is there for this class a faster way to decompose the matrix? We first consider the Gauss elimination for a symmetric and positive definite matrix

$$A = \left(\begin{array}{c|c} a_{11} & a_{21} \cdots\cdots a_{n1} \\ \hline a_{21} & \\ \vdots & B \\ a_{n1} & \end{array}\right)$$

for some symmetric B. Since A is positive definite, it follows that $a_{11} > 0$, and no row swapping is necessary. In the first step, the first column is eliminated by multiplication with a Forbenius matrix from the left:

$$L_1 A = \left(\begin{array}{c|c} a_{11} & a_{21} \cdots\cdots a_{n1} \\ \hline 0 & \\ \vdots & \widetilde{B} \\ 0 & \end{array}\right)$$

If one also multiplies by L_1^T from the right, it follows

$$\begin{aligned} A^{(1)} &:= L_1 A L_1^T = \left(L_1 (L_1 A)^T\right)^T \\ &= \left(L_1 \left(\begin{array}{c|c} a_{11} & 0 \cdots\cdots 0 \\ \hline a_{21} & \\ \vdots & \widetilde{B}^T \\ a_{n1} & \end{array}\right)\right)^T \\ &= \left(\begin{array}{c|c} a_{11} & 0 \cdots\cdots 0 \\ \hline 0 & \\ \vdots & \widetilde{\widetilde{B}}^T \\ 0 & \end{array}\right)^T \end{aligned}$$

$$
= \left(\begin{array}{c|c}
a_{11} & 0 \cdots\cdots 0 \\
\hline
0 & \\
\vdots & \\
\vdots & \widetilde{\widetilde{B}} \\
\vdots & \\
0 &
\end{array}\right)
$$

So, because of the symmetry, the first row and the first column have been eliminated in the first step. The matrix $A^{(1)}$ is now again symmetric

$$\big(A^{(1)}\big)^T = \big(L_1 A L_1^T\big)^T = L_1 A^T L_1^T = L_1 A L_1^T = A^{(1)}$$

and also positive definite due to the following consideration:

$$\langle x, A^{(1)}x\rangle = x^T L_1 A L_1^T x = \big(L_1^T x\big)^T A \big(L_1^T x\big) = y^T A y = \langle y, Ay\rangle.$$

By regularity of L_1^T, $x \neq 0$ also implies $y := L_1^T x \neq 0$. Finally, since A is symmetric and positive definite, we get

$$\langle x, A^{(1)}x\rangle = \langle y, Ay\rangle > 0 \quad \forall x \neq 0,$$

thus $A^{(1)}$ is positive definite.

This means that $a_{22}^{(1)} > 0$ and the second row and column can be eliminated analogously with a Frobenius matrix L_2:

$$
A^{(2)} = L_2 A^{(1)} L_2^T = L_2 L_1 A L_1^T L_2^T = \left(\begin{array}{cc|c}
a_{11} & 0 & 0 \cdots\cdots 0 \\
0 & a_{22}^{(1)} & 0 \cdots\cdots 0 \\
\hline
0 & 0 & \\
\vdots & \vdots & \\
\vdots & \vdots & \widetilde{\widetilde{B}} \\
\vdots & \vdots & \\
0 & 0 &
\end{array}\right).
$$

$A^{(2)}$ is then, analogous to the first step, symmetric and positive definite. After $(n-1)$ steps one has arrived at a diagonal matrix

$$
A^{(n-1)} = L_{n-1} \cdot \ldots \cdot L_1 A L_1^T \cdot \ldots \cdot L_{n-1}^T = \begin{pmatrix}
a_{11} & & & \\
& a_{22}^{(1)} & & \\
& & \ddots & \\
& & & a_{nn}^{(n-1)}
\end{pmatrix} =: D, \tag{2.19}
$$

where all diagonal elements of D are positive. Altogether it follows from Equation 2.19 that

$$A = L_1^{-1} \cdot \ldots \cdot L_{n-1}^{-1} D \big(L_{n-1}^T\big)^{-1} \cdot \ldots \cdot \big(L_1^T\big)^{-1}. \tag{2.20}$$

For every matrix $C \in \mathbb{R}^{n\times n}$ we now have $\big(C^T\big)^{-1} = \big(C^{-1}\big)^T$, since

$$\big(C^{-1}\big)^T C^T = \big(CC^{-1}\big)^T = E^T = E.$$

This implies for the matrix product on the right-hand side of D in Equation 2.20:

$$\big(L_{n-1}^T\big)^{-1} \cdot \ldots \cdot \big(L_1^T\big)^{-1} = \big(L_{n-1}^{-1}\big)^T \cdot \ldots \cdot \big(L_1^{-1}\big)^T$$

$$
\begin{aligned}
&= \left(L_1^{-1} \cdot \ldots \cdot L_{n-1}^{-1}\right)^T \\
&= L^T.
\end{aligned}
$$

Finally, this procedure produces a decomposition

$$A = LDL^T$$

with a diagonal matrix D with positive diagonal elements and a unipotent lower triangular matrix L.

Proposition 2.27. *Let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. Then A can be uniquely decomposed as follows:*

*(a) $A = LDL^T$ for a diagonal matrix D with positive diagonal elements and a unipotent lower triangular matrix L (**rational Cholesky decomposition**).*

*(b) $A = GG^T$ for a lower triangular matrix G (**Cholesky decomosition**). Once the decomposition is done, the solution of the system $Ax = b$ is then again obtained with forward and backward substitution, i.e. first find y with $Gy = b$ by forward substitution, then find x with $G^T x = y$ by backward substitution. The vector x is then the solution of $Ax = b$, because*

$$Ax = GG^T x = Gy = b.$$

Proof. The second assertion follows from the first if one lets

$$G := L\sqrt{D},$$

where for a diagonal Matrix

$$D = \begin{pmatrix} d_{11} & & & \\ & d_{22} & & \\ & & \ddots & \\ & & & d_{nn} \end{pmatrix}$$

with positive $d_{jj} > 0$ we set

$$\sqrt{D} := \begin{pmatrix} \sqrt{d_{11}} & & & \\ & \sqrt{d_{22}} & & \\ & & \ddots & \\ & & & \sqrt{d_{nn}} \end{pmatrix}.$$

Then, with $\sqrt{D}\sqrt{D} = D$, point (b) of the Proposition follows immediately.

Uniqueness in (a): Since

$$A = LDL^T = LU,$$

where $U := DL^T$ is a regular upper triangular matrix, it follows that L, U and hence D is uniquely determined by uniqueness of the LU decomposition.

Uniqueness in (b) follows from the direct computation of G, see below. ∎

For the practical implementation on the computer, the above derivation using LU decomposition is usually not used, but the elements of G are calculated directly instead. For this, let $A \in \mathbb{R}^{n\times n}$ be symmetric and positive definite. We are looking for G with

$$A = GG^T.$$

Writing

$$G = \begin{pmatrix} g_{11} & & & \\ g_{21} & g_{22} & & \\ \vdots & \vdots & \ddots & \\ g_{n1} & g_{n2} & \cdots & g_{nn} \end{pmatrix} = \left(\begin{array}{c|c} g_{11} & \\ \hline g^{(2)} & G^{(2)} \end{array}\right),$$

where

$$g^{(2)} = \begin{pmatrix} g_{21} \\ \vdots \\ g_{n1} \end{pmatrix} \in \mathbb{R}^{n-1}, \quad G^{(2)} = \begin{pmatrix} g_{22} & & \\ \vdots & \ddots & \\ g_{n2} & \cdots & g_{nn} \end{pmatrix} \in \mathbb{R}^{(n-1)\times(n-1)},$$

we obtain

$$A = \left(\begin{array}{c|ccc} a_{11} & a_{21} & \cdots & a_{n1} \\ \hline a_{21} & & & \\ \vdots & & B^{(2)} & \\ a_{n1} & & & \end{array}\right) \overset{!}{=} GG^T = \left(\begin{array}{c|ccc} g_{11} & 0 & \cdots & 0 \\ \hline g_{21} & & & \\ \vdots & & G^{(2)} & \\ g_{n1} & & & \end{array}\right) \left(\begin{array}{c|ccc} g_{11} & g_{21} & \cdots & g_{n1} \\ \hline 0 & & & \\ \vdots & & \left(G^{(2)}\right)^T & \\ 0 & & & \end{array}\right)$$

$$= \left(\begin{array}{c|ccc} g_{11}^2 & g_{11}g_{21} & \cdots & g_{11}g_{n1} \\ \hline g_{11}g_{21} & & & \\ \vdots & & G^{(2)}\left(G^{(2)}\right)^T + g^{(2)}\left(g^{(2)}\right)^T & \\ g_{11}g_{n1} & & & \end{array}\right).$$

A comparison of coefficients now yields

$$g_{11}^2 = a_{11} \quad \Rightarrow \quad g_{11} = \sqrt{a_{11}}$$

and

$$g_{11}g_{i1} = a_{i1} \quad \Rightarrow \quad g_{i1} = \frac{a_{i1}}{\sqrt{a_{11}}}, \qquad i = 2, \ldots, n.$$

This means that the first column of the decomposition matrix G we are looking for is known, in particular $g^{(2)}$.

Next, the elements of the submatrix $G^{(2)}$ have to be determined. First, one calculates again with coefficient comparison

$$G^{(2)}\left(G^{(2)}\right)^T = B^{(2)} - g^{(2)}\left(g^{(2)}\right)^T =: A^{(2)} =: (a_{ij})_{2\leq i,j\leq n}.$$

First, the coefficients of the auxiliary matrix

$$A^{(2)} := B^{(2)} - g^{(2)}\left(g^{(2)}\right)^T = G^{(2)}\left(G^{(2)}\right)^T,$$

where

$$a_{ij}^{(2)} = a_{ij} - g_{i1}g_{j1}, \quad i, j \geq 2$$

are calculated. Analogously to the first step, the first column of $G^{(2)}$ can be determined from

$$G^{(2)}\left(G^{(2)}\right)^T = A^{(2)},$$

for $A^{(2)}$ is symmetric and, due to the regularity of $G^{(2)}$, also positive definite. This procedure is repeated until all columns of G have been calculated. The kth step of the Cholesky factorization thus consists of the following sub-steps

(a) Determine the kth column of G using

$$
\begin{aligned}
g_{kk} &= \sqrt{a_{kk}^{(k)}} \\
g_{ik} &= \frac{a_{ik}^{(k)}}{g_{kk}}, \quad i = k+1, \ldots, n
\end{aligned}
$$

(b) Determine $A^{(k+1)} = \big(a_{ij}^{(k+1)}\big)_{k+1 \le i,j \le n}$ using

$$
a_{ij}^{(k+1)} = a_{ij}^{(k)} - g_{ik} g_{jk}, \quad i, j \ge k+1.
$$

For $k = 1$ one sets $A^{(1)} := A$. The numerical effort (multiplications and divisions) of this procedure amounts to

$$
\sum_{k=1}^{n-1} (n-k) = \sum_{k=1}^{n-1} k = \frac{n(n-1)}{2} \sim \frac{n^2}{2}
$$

for step (a), and for step (b) we have

$$
\begin{aligned}
\sum_{k=1}^{n-1} (1 + \ldots + (n-k)) &= \sum_{k=1}^{n-1} \frac{(n-k)(n-k+1)}{2} \\
&= \sum_{k=1}^{n-1} \frac{k(k+1)}{2} \\
&\sim \frac{n^3}{6}.
\end{aligned}
$$

The latter follows from

$$
n^3 = \sum_{k=0}^{n-1} \big[(k+1)^3 - k^3\big] = \sum_{k=0}^{n-1} \big[3k^2 + 3k + 1\big].
$$

The effort compared to conventional LU decomposition is reduced by a factor of two due to the symmetry.

Example 2.28. Determine the Cholesky decomposition of A and then the solution of the system $Ax = b$ with

$$
A = \big(4\big)24; 2105; 4521, \quad b = \big(1\big)0; 14; 13.
$$

Following the procedure discussed, the first column of G is immediately obtained as follows:

$$
\begin{aligned}
g_{11} &= \sqrt{a_{11}} = 2, \\
g_{21} &= \frac{a_{21}}{g_{11}} = \frac{2}{2}, \\
g_{31} &= \frac{a_{31}}{g_{11}} = \frac{4}{2} = 2.
\end{aligned}
$$

The auxiliary matrix $A^{(2)} \in \mathbb{R}^{2\times 2}$ is now determined by (due to the symmetry we only need to calculate the values on one side of the diagonal):

$$
\begin{aligned}
(a_{ij}^{(2)})_{i,j\ge 2} &= \big(a\big)_{22}*; a_{32}a_{33} - \big(g\big)_{21}^{2}*; g_{31}g_{21}g_{31}^{2} \\
&= \big(1\big)0 - 1\cdot 1*; 5 - 2\cdot 121 - 2\cdot 2 \\
&= \big(9\big)*; 317.
\end{aligned}
$$

In the second step, the second column of G is determined. One calculates

$$g_{22} = \sqrt{a_{22}^{(2)}} = 3,$$
$$g_{32} = \frac{a_{32}^{(2)}}{g_{22}} = \frac{3}{3} = 1.$$

Now, the auxiliary matrix $A^{(3)} \in \mathbb{R}^{1\times 1}$ is just a number:

$$a_{33}^{(3)} = a_{33}^{(2)} - g_{32}^2 = 17 - 1\cdot 1 = 16$$

and it finally follows in the third step

$$g_{33} = \sqrt{a_{33}^{(3)}} = 4,$$

i.e.

$$G = \big(2\big); 13; 214.$$

The solution of the system of equations with the right-hand side given in the example then results from forward and backward substitution as

$$x = \begin{pmatrix} 2 \\ 1 \\ 0 \end{pmatrix}. \qquad //.$$

2.7. Numerical calculation of a dominant eigenvalue: the Power method

In the previous sections we saw that largest absolute eigenvalues (= dominant eigenvalues) or smallest absolute eigenvalues of a matrix $A \in \mathbb{R}^{n\times n}$ play a decisive role in the calculation of the spectral norm or the quantity $\mathrm{cond}_2(A)$, i.e. the condition number of A with respect to the spectral norm, which estimates the error amplification with respect to the Euclid norm. To determine the dominant eigenvalue, one would like a simple procedure that avoids the problem of calculating all eigenvalues as far as possible. The central idea of the **power method** is the continued application of the matrix A to an (arbitrary) start vector $v^{(0)}$, i.e. one carries out the iteration rule

$$v^{(k)} = Av^{(k-1)} = A^k v^{(0)}.$$

One can now suppose that $v^{(k)}$ for large k essentially looks like an eigenvector for the dominant eigenvalue. This is based on the following consideration:

Let for simplicity $A \in \mathbb{R}^{n\times n}$ be a diagonalizable Matrix, i.e. we can choose a Basis $\{v_1, \ldots, v_n\}$ of eigenvectors of $\mathbb{R}^n$ with $Av_i = \lambda_i v_i$. Let

$$v^{(0)} = \sum_{j=1}^{n} \mu_j v_j$$

be an arbitrary start vector with $\mu_1 \neq 0$ and $\|v^{(0)}\| = 1$, where $\|\cdot\|$ is the Euclidean norm. Assuming

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|,$$

it follows that

$$v^{(k)} = A^k v^{(0)}$$

$$
\begin{aligned}
&= A^k\Big(\sum_{j=1}^{n}\mu_j v_j\Big) \\
&= \sum_{j=1}^{n}\mu_j A^k v_j \\
&= \sum_{j=1}^{n}\mu_j \lambda_j^k v_j && (2.21) \\
&= \underbrace{\lambda_1^k\mu_1}_{=:\alpha_k\neq 0}\Big(v_1 + \underbrace{\sum_{j=2}^{n}\frac{\mu_j}{\mu_1}\Big(\frac{\lambda_j}{\lambda_1}\Big)^k v_j}_{=:r^{(k)}\overset{k\to\infty}{\longrightarrow}0}\Big) \\
&= \alpha_k(v_1 + r^{(k)}) && (2.22)
\end{aligned}
$$

So, for large k, the vector $v^{(k)}$ essentially looks like the eigenvector v_1 for the largest absolute eigenvalue λ_1.

Approximation of λ_1: From Equation 2.22 it follows using $\alpha_{k+1} = \lambda_1\alpha_k$:

$$
\begin{aligned}
\frac{\langle v^{(k)}, Av^{(k)}\rangle}{\|v^{(k)}\|^2} &= \frac{\langle \alpha_k(v_1 + r^{(k)}), \alpha_{k+1}(v_1 + r^{(k+1)})\rangle}{\|\alpha_k(v_1 + r^{(k)})\|^2} && (2.23) \\
&= \lambda_1\frac{\langle v_1 + r^{(k)}, v_1 + r^{(k+1)}\rangle}{\|v_1 + r^{(k)}\|^2} \\
&\to \lambda_1\frac{\|v_1\|^2}{\|v_1\|^2} = \lambda_1.
\end{aligned}
$$

Problem when calculating $v^{(k)}$: If $|\lambda_1| \neq 1$ it follows that $|\lambda_1|^k \to \infty$ or $|\lambda_1|^k \to 0$ and hence either $\|v^{(k)}\| \to \infty$ or $\|v^{(k)}\| \to 0$, which leads to an arithmetic over-/underflow. One way out is to normalize in each step:

For this, we first remark that if the first component of an arbitrary vector $x = \sum_{j=1}^{n} x_j v_j$ is nonzero, then this also applies to $Ax = x_1\lambda_1 v_1 + \sum_{j=2}^{n} x_j\lambda_j v_j \neq 0$ and thus also to $Ax/\|Ax\|$. Therefore, we may define

$$
n^{(k+1)} := \frac{An^{(k)}}{\|An^{(k)}\|},
$$

where $n^{(0)} := v^{(0)}$.
Now we show by induction that for all $k \geq 0$ it holds

$$
n^{(k)} = \frac{v^{(k)}}{\|v^{(k)}\|}.
$$

(Note that the first component $\mu_1\lambda_1^k$ of $v^{(k)}$ is nonzero in (2.21), which implies that $\|v^{(k)}\| > 0$.) For $k = 0$ this is true since $v^{(0)}$ is normalized. The induction step follows from

$$
\begin{aligned}
n^{(k+1)} &= \frac{An^{(k)}}{\|An^{(k)}\|} \\
&= \frac{A\Big(\frac{v^{(k)}}{\|v^{(k)}\|}\Big)}{\Big\|A\Big(\frac{v^{(k)}}{\|v^{(k)}\|}\Big)\Big\|} \\
&= \frac{Av^{(k)}}{\|Av^{(k)}\|}
\end{aligned}
$$

$$= \frac{v^{(k+1)}}{\|v^{(k+1)}\|}.$$

Finally, (2.23) shows that

$$\lambda_1^{(k)} := \langle n^{(k)}, An^{(k)} \rangle \to \lambda_1.$$

Moreover, from (2.22) it follows that

$$n^{(k)} = \frac{v^{(k)}}{\|v^{(k)}\|} = \frac{\alpha_k(v_1 + r^{(k)})}{\|\alpha_k(v_1 + r^{(k)})\|} = \operatorname{sign}(\lambda_1^k \mu_1) \underbrace{\frac{v_1 + r^{(k)}}{\|v_1 + r^{(k)}\|}}_{\to \frac{v_1}{\|v_1\|}}.$$

The numerical procedure thus takes the following form:

Power Method (Von Mises Iteration)

Initialization: Let a matrix $A \in \mathbb{R}^{n \times n}$ be given. Choose an arbitrary $n^{(0)} \in \mathbb{R}^n$.

for $k = 0, 1, 2, \ldots$ **do**

$$h^{(k+1)} = An^{(k)}$$

$$n^{(k+1)} = \frac{h^{(k+1)}}{\|h^{(k+1)}\|}$$

$$\lambda_1^{(k+1)} = \langle n^{(k+1)}, An^{(k+1)} \rangle$$

until stop

In order to calculate the smallest absolute eigenvalue of a regular matrix A, one notices that the spectrum of A^{-1} is given by

$$\sigma(A^{-1}) = \left\{ \frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n} \right\}$$

and thus the smallest absolute eigenvalue of A is exactly the reciprocal value of the dominant eigenvalue of A^{-1}.

To determine the smallest absolute eigenvalue, the power method must be applied to A^{-1}. You can even avoid the calculation of A^{-1} because

$$v^{(k+1)} = A^{-1} v^{(k)} \quad \Leftrightarrow \quad Av^{(k+1)} = v^{(k)}.$$

This means that in each iteration step one has to solve a system of linear equations for which e.g. the LU decomposition of A is sufficient.

3. Systems of linear Equations - Iterative Methods

In the last chapter we got to know decomposition results for matrices that are generated by elimination methods. The complexity of these decomposition algorithms increases with the third power of the matrix dimension. For very large matrices, which are often sparse (a sparse matrix is a matrix in which most of the elements are zero), it can make sense to use other methods for solving linear systems: iterative methods. A central result for all iterative methods is Banach's fixed point theorem.

3.1. The Banach Fixed Point Theorem

Theorem 3.1 (Banach fixed point theorem). *Let $\Phi\colon \mathcal{K} \to \mathcal{K}$ a contracting self-map (a contraction) of a closed subset $\mathcal{K} \subset \mathbb{R}^n$ with contraction factor q, i.e.*

$$\|\Phi(x) - \Phi(y)\| \le q\|x - y\| \qquad \text{for some } q < 1 \text{ and all } x, y \in \mathcal{K} \tag{3.1}$$

Then the fixed point equation

$$x = \Phi(x) \tag{3.2}$$

has exactly one solution $\hat{x} \in \mathcal{K}$ ($\hat{x}$ is called the fixed point of Φ), and the fixed point iteration

$$x^{(k+1)} := \Phi(x^{(k)}), \quad k = 0, 1, 2, \ldots \tag{3.3}$$

converges for each initial vector $x^{(0)} \in \mathcal{K}$ to $\hat{x}$ when $k \to \infty$. Furthermore, for $k \ge 1$, we have

$$\|x^{(k)} - \hat{x}\| \le \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\| \tag{a priori estimate}$$

and

$$\|x^{(k)} - \hat{x}\| \le \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\|. \tag{a posteriori estimate}$$

as well as

$$\|x^{(k+1)} - \hat{x}\| \le q\|x^{(k)} - \hat{x}\|. \tag{error reduction formula}$$

Proof. Choose an arbitrary $x^{(0)} \in \mathcal{K}$ an let $(x^{(k)})$ be defined as in (3.2).

Step 1: $x^{(k)}$ is a Cauchy sequence:

Due to the contraction property (3.1), for any $k \in \mathbb{N}$ it holds

$$\|x^{(k+1)} - x^{(k)}\| = \|\Phi(x^{(k)}) - \Phi(x^{(k-1)})\| \le q\|x^{(k)} - x^{(k-1)}\|.$$

Therefore, we obtain by induction (over $p \in \mathbb{N}_0$) that for all $k \ge 1$ and $p \ge 0$

$$\|x^{(k+p)} - x^{(k+p-1)}\| \le q^p\|x^{(k)} - x^{(k-1)}\|. \tag{3.4}$$

From this one can deduce (using the triangle inequality for the telescoping sum)

$$\begin{aligned}
\|x^{(k+p)} - x^{(k)}\| &\le \|x^{(k+p)} - x^{(k+p-1)}\| + \ldots + \|x^{(k+1)} - x^{(k)}\| \\
&\overset{(3.4)}{\le} (q^p + q^{p-1} + \ldots + q)\|x^{(k)} - x^{(k-1)}\| \\
&\le \frac{q}{1-q}\|x^{(k)} - x^{(k-1)}\|
\end{aligned} \tag{3.5}$$

$$\leq \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\|. \tag{3.6}$$

Since $0 < q < 1$ we can conclude that for all $k \leq m := k + p$ large enough it is

$$\|x^{(m)} - x^{(k)}\| \leq \frac{q^k}{1-q}\|x^{(1)} - x^{(0)}\| < \varepsilon,$$

meaning that $(x^{(k)})$ is a Cauchy sequence. Letting $\hat{x}$ denote the unique limit (it may depend on $x^{(0)}$), the closedness of $\mathcal{K}$ then implies that $\hat{x} \in \mathcal{K}$.

Step 2: $\hat{x}$ is a unique fixed point of Φ:

First, Φ is lipschitz continuous (and therefore continous) on $\mathcal{K}$. Hence we can pass to the limit in the iteration equation (3.3) to show that $\hat{x}$ is a fixed point of Φ:

$$\hat{x} = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} \Phi(x^{(k)}) = \Phi(\hat{x}),$$

where the last equality follows from continuity of Φ.
Uniqueness now follows again with inequality (3.1): Suppose there are two fixed points $\hat{x}$ and $\tilde{x}$. Then

$$\|\hat{x} - \tilde{x}\| = \|\Phi(\hat{x}) - \Phi(\tilde{x})\| \leq q\|\hat{x} - \tilde{x}\|$$

follows and because of $q < 1$ this can only hold if $\hat{x} = \tilde{x}$.

Step 3: Proof of the a priori/a posteriori estimates:

Letting $p \to \infty$ in (3.6) and (3.5) the a priori and a posteriori estimates follow. For the error reduction formula we finally estimate

$$\|x^{(k+1)} - \hat{x}\| = \|\Phi(x^{(k)}) - \Phi(\hat{x})\| \leq q\|x^{(k)} - \hat{x}\|. \qquad \blacksquare$$

Example 3.2. The Banach fixed point theorem should be illustrated using several examples.

(a) The Newton method for determining a zero of a differentiable function $f\colon \mathbb{R} \to \mathbb{R}$ is given by the iteration formula

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

with given initial guess $x^{(0)} \in \mathbb{R}$.

For the function $f(x) = x^2 - 2$, this results in the iteration formula

$$x_{k+1} = \frac{1}{2}\Big(x_k + \frac{2}{x_k}\Big).$$

This defines the fixed point iteration

$$x_{k+1} = \Phi(x_k)$$

for the approximation of $\sqrt{2}$, where

$$\Phi(x) := \frac{1}{2}\Big(x + \frac{2}{x}\Big).$$

For $I := [1; 3]$ we have that $\Phi\colon I \to I$ is a contraction with contraction factor $q = \frac{1}{2}$ and thus the Banach fixed point theorem applies. That is, there is exactly one fixed point and Newton's iteration converges to this.

(b) One can also use alternative fixed point iterations to determine a zero of a function. For example, if you look for the zeros of the function

$$f(x) := x^4 - x + \frac{1}{4}$$

in the intervall $D = [0; ½]$, one can rewrite the problem in fixed point form as follows:

$$x = x^4 + \frac{1}{4}.$$

The associated operator

$$\Phi(x) = x^4 + \frac{1}{4}$$

then describes a contraction on D so that the Banach fixed point theorem applies here as well:

- Φ is contracting: For $x, y \in D$ we have

$$|\Phi(x) - \Phi(y)| \leq \max_{\xi \in D} |\Phi'(\xi)||x - y|,$$

  and therefore Φ is contracting with contraction factor $\frac{1}{2}$.
- Φ is a self-map: For $x \in D$ it follows that

$$0 \leq \Phi(x) \leq \frac{1}{16} + \frac{1}{4} = \frac{5}{16} \leq \frac{1}{2}$$

  and hence, using the iteration

$$x_{k+1} = \Phi(x_k),$$

  one can find the unique zero of f in D for all initial data $x^{(0)} \in D$.

(c) The Banach fixed point theorem also reveals an interesting property of maps/city plans: We denote the urban area of Rosenheim with X and we spread out a city map directly on the ground at some point (not at the edge), e.g. in the city center. Then there is exactly one point of X that lies exactly under its picture on the city map: i.e. if we pierce the ground with a needle through the city map, there is exactly one point on the city map where the puncture point (i.e. a point on the map) exactly matches the point in reality directly below on the ground.

This can be thought of as follows. We define a mapping $\Phi\colon X \to X$ as follows: We first assign an element x from X to the point to which it corresponds on the city map. Then we stick a needle through that point on the ground and hit an element y out of X that is directly under the city map. This point y should then be the image point, i.e. $\Phi(x) := y$. Since the image of X lies directly under the city map, one obtains that the mapping Φ describes a contraction. Namely, one obtains

$$|\Phi(x) - \Phi(y)| \leq \frac{1}{M}|x - y|,$$

if the map scale of the city is $1\colon M$. Obviously, Φ is also a self map and our statement above follows from the Banach fixed point theorem. //.

For given $A \in \mathrm{Gl}(n)$ and $b \in \mathbb{R}^n$, the Banach fixed point theorem can be used to construct convergent iteration methods for the numerical solution of regular systems of linear equations $Ax = b$. Choose a decomposition of A of the form $A = M - N$, where M should be invertible. The equation $Ax = b$ can then be brought into fixed point form as follows:

$$Ax = b \quad \Leftrightarrow \quad Mx = Nx + b \quad \Leftrightarrow \quad x = Tx + c,$$

where $T = M^{-1}N$ and $c = M^{-1}b$. In this case, the fixed point operator is the affine function

$$\Phi(x) := Tx + c.$$

For this approach it is of course necessary that the matrix M in particular is simple enough (no inverse calculation necessary) and that the fixed point iteration converges. The following theorems provide conditions for the convergence of such methods.

Proposition 3.3. *Let $\|\cdot\|_*$ be a Norm in $\mathbb{R}^{n\times n}$ that is compatible with a vector norm $\|\cdot\|$. Suppose $A = M - N$ with an invertible matrix M and further let $T = M^{-1}N$ with*

$$\|T\|_* = \|M^{-1}N\|_* < 1.$$

Then the fixed point iteration with fixed point operator $\Phi(x) = Tx + c$ and $c = M^{-1}b$ converges according to the iteration

$$x^{(k+1)} = \Phi(x^{(k)}) = Tx^{(k)} + c \tag{3.7}$$

for each initial guess $x^{(0)}$ to the solution $\hat{x}$ of $Ax = b$, i.e. $y = A^{-1}b$.

Proof. We first show that Φ is a contracting self map. Φ obviously satisfies $\Phi\colon \mathbb{R}^n \to \mathbb{R}^n$, and by setting $\mathcal{K} := \mathbb{R}^n$ (in Theorem 3.1) the property of being a self map is clear.
Furthermore, because of the compatibility of the norms we have

$$\|\Phi(x) - \Phi(y)\| = \|T(x-y)\| \le \|T\|_*\|x-y\|$$

and because

$$\|T\|_* = \|M^{-1}N\|_* < 1$$

Φ is a contraction. According to Theorem 3.1, the sequence defined in (3.7) converges for each initial datum to the uniquely determined fixed point $\hat{x}$ with

$$\hat{x} = T\hat{x} + c.$$

According to item 3.2, this is equivalent to $A\hat{x} = b$. Thus the fixed point is the solution of the system of linear equations $Ax = b$ according to the construction. Conversely, since every solution with a variation of the right-hand side b always corresponds to a fixed point of Φ, its uniqueness means that A is invertible. ■

There is another convergence criterion for iterative methods of the form discussed:

Proposition 3.4. *Let A be invertible and $T = M^{-1}N$ be as above. Then the method (3.7) converges for every $x^{(0)}$ to $\hat{x} = A^{-1}b$ if and only if for the spectral radius of T it holds*

$$\rho(T) < 1.$$

Furthermore, the following identity holds true:

$$\max_{x^{(0)}} \limsup_{k\to\infty} \|x^{(k)} - \hat{x}\|^{\frac{1}{k}} = \rho(T).$$

In the case of convergence, i.e. $\rho(T) < 1$, for every $\varepsilon > 0$ and every initial guess $x^{(0)}$ there exists an index $K \in \mathbb{N}$, such that

$$\|x^{(k)} - \hat{x}\| \le \big(\rho(T) + \varepsilon\big)^k \quad \forall k \ge K.$$

That is, the asymptotic convergence speed is $\rho(T)^k$.

Proof. Without proof. ■

The following example shows that this speed of convergence really only applies asymptotically:

Example 3.5. Consider the invertible Matrix

$$A = \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ \vdots & & & \ddots & \ddots & -1 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}.$$

Letting $M := E$, we obtain from the decomosition $A = M - N$ that

$$T = M^{-1}N = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ \vdots & & & \ddots & 1 \\ 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}.$$

The matrix T is called a shift matrix, because if $x = (x_1, \ldots, x_n)^T$ the vector $Tx = (x_2, \ldots, x_n, 0)^T$ contains the entries of x shifted up by one. From the matrix T one can immediately see that $\rho(T) = 0$, i.e. one would expect a very fast convergence. Letting $x^{(0)} = \hat{x} + e_n$, where $\hat{x}$ is the solution of $Ax = b$ (and thus the unique fixed point of the iteration (3.7) at the same time), one obtains for $1 \leq k \leq n-1$

$$\begin{aligned} \|x^{(k)} - \hat{x}\| &= \|\Phi(x^{(k-1)}) - \Phi(\hat{x})\| \\ &= \|T(x^{(k-1)} - \hat{x})\| \\ &= \|T^k(x^{(0)} - \hat{x})\| \\ &= \|T^k e_n\| \\ &= \|e_{n-k}\| \\ &= 1 \\ &= \|x^{(0)} - \hat{x}\|. \end{aligned}$$

This means that in the first $n-1$ steps no error reduction occurs at all. The meaning of the convergence rate in Proposition 3.4 is therefore only asymptotic in nature. //.

3.2. The Jacobi and Gauss-Seidel method

The simplest method for the iterative solution of a system of linear equations $Ax = b$ with $A = (a_{ij})_{i,j=1,\ldots,n}$ is the Jacobi method. In this method, all diagonal entries of A must be non-zero. One gets an iterative algorithm by solving the i-th equation for the i-th unknown:

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\big(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)}\big) \tag{3.8}$$

Here k denotes the iteration index. This leads to the following algorithm

JACOBI METHOD

Initialization: Given the system of linear equations $Ax = b$ with $a_{ii} \neq 0$ for $i = 1, \ldots, n$. Choose any $x^{(0)} \in \mathbb{R}^n$.

for $k = 0, 1, 2, \ldots$ **do**

  for $i = 1, \ldots, n$ **do**

    $x_i^{(k+1)} = \frac{1}{a_{ii}}\big(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)}\big)$

  end

until stop

For each iteration, the computational complexity of this procedure includes as many multiplications/divisions as the number of nonzero elements of the matrix A, i.e. a maximum of n^2 (per iteration!).

We want to reduce the question of convergence of the Jacobi method to the Banach fixed point theorem and method (3.7). For this we decompose

$$A = D - L - U \tag{3.9}$$

where D is a diagonal matrix with the diagonal elements of A, L is a strict lower triangular matrix, and U is a strict upper triangular matrix. Then the equations of the Jacobi method (3.8) can also be written as

$$x^{(k+1)} = D^{-1}\big(b + (L + U)x^{(k)}\big). \tag{3.10}$$

So this corresponds to the fixed point iteration (3.7) with $T = M^{-1}N$ and $M = D$, as well as $N = L + U$. The iteration matrix $T = \mathcal{J} = D^{-1}(L + U)$ is also called the Jacobi iteration matrix.

Example 3.6. Carry out the Jacobi iteration method for the first four iterations of the system $Ax = b$ with

$$A = \begin{pmatrix} 5 & 1 & 1; 1 & 5 & 0; 1 & 0 & 5 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}.$$

How is the Jacobi iteration matrix defined? Does the method converge? Choose $x^{(0)} = 0$ as initial guess.

First, the iteration process is defined by the equations

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{5}(1 - x_2^{(k)} - x_3^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{5}(2 - x_1^{(k)}) \\ x_3^{(k+1)} &= \frac{1}{5}(0 - x_1^{(k)}) \end{aligned}$$

For the first four iterations we get

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.0000	0.0000	0.0000
1	0.2000	0.4000	0.0000
2	0.1200	0.3600	−0.0400
3	0.1360	0.3760	−0.0240
4	0.1296	0.3728	−0.0272

The exact solution rounded to 4 decimal places is $x = (0.1304; 0.3739; -0.0261)$ and the error in the Euclid norm after four iterations is $1.8 \cdot 10^{-3}$.

The Jacobi iteration matrix is calculated from the matrices D, L, U to $\mathcal{J} = D^{-1}(L+U)$, where in this case the following holds:

$$D = \begin{pmatrix} 5 & 0 & 0; 0 & 5 & 0; 0 & 0 & 5 \end{pmatrix}, \quad L = \begin{pmatrix} 0 & 0 & 0; -1 & 0 & 0; -1 & 0 & 0 \end{pmatrix}, \quad U = \begin{pmatrix} 0 & -1 & -1; 0 & 0 & 0; 0 & 0 & 0 \end{pmatrix}.$$

Hence, we have

$$\mathcal{J} = D^{-1}(L+U) = \frac{1}{5}\begin{pmatrix} 0 & -1 & -1; -1 & 0 & 0; -1 & 0 & 0 \end{pmatrix}.$$

One easily verifies that

$$\sigma(\mathcal{J}) = \left\{0, \frac{1}{5}\sqrt{2}, -\frac{1}{5}\sqrt{2}\right\}.$$

Hence the convergence of the algorithm follows with Proposition 3.4. //.

In the very similar Gauss-Seidel method (also: method of successive displacement), all components of $x^{(k+1)}$ that have already been calculated are inserted into the right-hand side of (3.8). The iteration rule is therefore

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\Big(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{k}\Big), \quad i = 1, \ldots, n. \tag{3.11}$$

GAUSS-SEIDEL METHOD

Initialization: Given the system of linear equations $Ax = b$ with $a_{ii} \neq 0$ for $i = 1, \ldots, n$. Choose any $x^{(0)} \in \mathbb{R}^n$.

```
for k = 0,1,2,... do
  for i = 1,...,n do
    x_i^(k+1) = 1/a_ii (b_i - sum_{j<i} a_ij x_j^(k+1) - sum_{j>i} a_ij x_j^k)
  end
until stop
```

If one writes Equation 3.11 in matrix form, one obtains with the same decomposition $A = D - L - U$ as in (3.9):

$$(D - L)x^{(k+1)} = b + Ux^{(k)}.$$

So we have again a fixed point iteration as in Proposition 3.3 or Proposition 3.4 with

$$M = D - L \quad \text{und} \quad N = U.$$

The operator

$$\mathcal{L} = M^{-1}N = (D - L)^{-1}U$$

is the Gauss-Seidel matrix.

Example 3.7. We calculate the Gauss-Seidel matrix for the previous example, i.e. we want to solve $Ax = b$ with the Gauss-Seidel method, where

$$A = \begin{pmatrix} 5 & 1 & 1; 1 & 5 & 0; 1 & 0 & 5 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 0 \end{pmatrix}$$

and it follows that

$$D - L = \begin{pmatrix} 5 & 0 & 0; 1 & 5 & 0; 1 & 0 & 5 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} 0 & -1 & -1; 0 & 0 & 0; 0 & 0 & 0 \end{pmatrix}.$$

Furthermore, we have

$$(D-L)^{-1} = \frac{1}{5}\big(1 \quad 0 \quad 0; -\tfrac{1}{5} \quad 1 \quad 0; -\tfrac{1}{5} \quad 0 \quad 1\big)$$

and we obtain for the Gauss-Seidel matrix

$$\mathcal{L} = (D-L)^{-1}U = \frac{1}{5}\big(0 \quad -1 \quad -1; 0 \quad \tfrac{1}{5} \quad \tfrac{1}{5}; 0 \quad \tfrac{1}{5} \quad \tfrac{1}{5}\big).$$

Finally, one easily verifies that

$$\sigma(\mathcal{L}) = \big\{0, \frac{2}{25}\big\} = \sigma(\mathcal{J})^2. \tag{3.12}$$

According to Proposition 3.4, the Gauss-Seidel method should converge faster than the Jacobi method. However, the identity (3.12) is not accidental (we shall later see that this is true for all matrices A of a certain form.) We calculate the first iterates of the Gauss-Seidel method with initial datum $x^{(0)} = (0,0,0)^T$:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$
0	0.0000	0.0000	0.0000
1	0.2000	0.3600	−0.0400
2	0.1360	0.3728	−0.0272
3	0.1309	0.3738	−0.0262
4	0.1305	0.3739	−0.0261

The error with respect to the exact solution is $3.7\cdot 10^{-5}$ after four iterations and is thus a good deal (almost two decimal places) better than the error of the Jacobi method after four iterations. //.

We can now apply Proposition 3.3 to obtain a convergence criterion for the Jacobi and Gauss-Seidel method. We first need the following definition.

Definition 3.8. *A matrix $A = (a_{ij})_{i,j=1,\dots,n} \in \mathbb{R}^{n\times n}$ is called strictly diagonally dominant if the inequality*

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|$$

holds for all $i = 1,\dots,n$. This means that the absolute value of the diagonal elements a_{ii} are each greater than the sum of the absolute values of the remaining respective row entries. //.

Proposition 3.9. *If $A \in \mathbb{R}^{n\times n}$ is strictly diagonally dominant, then the Jacobi and Gauss-Seidel method both converge for each initial guess $x^{(0)} \in \mathbb{R}^n$ to the unique solution of $Ax = b$.*

Proof. Due to the strict diagonal dominance, all diagonal entries of A are nonzero and the two iteration methods are well defined. For the proof of convergence we want to apply Proposition 3.3. First, let's look at the Jacobi method. From the strict diagonal dominance it follows immediately that

$$\|\mathcal{J}\|_\infty = \|D^{-1}(L+U)\|_\infty = \max_{i=1,\dots,n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} =: q < 1 \tag{3.13}$$

and the assumption of Proposition 3.3 is satisfied for the row sum norm (matrix norm induced by the maximum vector space norm).

For the Gauss-Seidel method, the proof is a bit more difficult. Again using the row sum norm, we want to prove that

$$\|\mathcal{L}\|_\infty := \max_{\|x\|_\infty=1} \|\mathcal{L}x\|_\infty < 1.$$

So let $\|x\|_\infty = 1$ and q be defined as in Equation 3.13. The individual components y_i of $y = \mathcal{L}x$ result from the definition of the Gauss-Seidel method (3.11) with

$$b = 0, \quad x^{(k)} = x \quad \text{and} \quad y = x^{(k+1)}$$

to

$$y_i = \frac{1}{a_{ii}}\Big(- \sum_{j<i} a_{ij} y_j - \sum_{j>i} a_{ij} x_j \Big).$$

We now show inductively that $|y_i| \le q < 1$ holds for all $i = 1, \dots, n$.

(I) Initial case $i = 1$: We have

$$\begin{aligned}
|y_1| &= \Big| - \frac{1}{a_{11}} \sum_{j=2}^{n} a_{1j} x_j \Big| \\
&\le \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{1j}| \|x\|_\infty \\
&= \frac{1}{|a_{11}|} \sum_{j=2}^{n} |a_{1j}| \\
&\le q.
\end{aligned}$$

(II) Induction step $i-1 \to i$: Assume $|y_k| \le q$ holds true for $0 \le k \le i-1$. For y_i we now conclude by induction hypothesis

$$\begin{aligned}
|y_i| &\le \frac{1}{|a_{ii}|}\Big(\sum_{j<i} |a_{ij}||y_j| + \sum_{j>i} |a_{ij}||x_j| \Big) \\
&\le \frac{1}{|a_{ii}|}\Big(\sum_{j<i} |a_{ij}| q + \sum_{j>i} |a_{ij}| \|x\|_\infty \Big) \\
&\le \frac{1}{|a_{ii}|}\Big(\sum_{j<i} |a_{ij}| + \sum_{j>i} |a_{ij}| \Big) \\
&\le q.
\end{aligned}$$

This implies $\|y\|_\infty \le q$ and hence we have $\|\mathcal{L}\|_\infty \le q < 1$. ■

Example 3.10. Give the system of linear equations $Ax = b$, where

$$A = \begin{pmatrix} 2 & 0 & 1; 1 & -4 & 1; 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 1; & 4; -1 \end{pmatrix},$$

with the exact solution $x = (1, -1, -1)^T$. A is then strictly diagonally dominant and the proof of the last Proposition yields the contraction factor $q = \frac{1}{2}$ regarding the row sum norm. The Jacobi and Gauss-Seidel method, both now converge and the error correction (regarding the maximum norm!) should be done with at least a factor of $q = \frac{1}{2}$. For the initial datum $x^{(0)} = (1, 1, 1)^T$ we first get $\|x^{(0)} - \hat{x}\|_\infty = 2$ and after one iteration we get for the

(a) Jacobi method:

$$x_{\mathcal{J}}^{(1)} = \begin{pmatrix} \frac{1}{2}(1-1) \\ -\frac{1}{4}(4-1-1) \\ \frac{1}{2}(-1+1) \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{1}{2} \\ 0 \end{pmatrix}$$

and it follows that $\|x_{\mathcal{J}}^{(1)} - \hat{x}\|_\infty = 1$.

(b) Gaus-Seidel method:

$$x_{\mathcal{L}}^{(1)} = \begin{pmatrix} \frac{1}{2}(1-1) \\ -\frac{1}{4}(4-0-1) \\ \frac{1}{2}(-1-\frac{3}{4}) \end{pmatrix} = \begin{pmatrix} 0 \\ -\frac{3}{4} \\ -\frac{7}{8} \end{pmatrix}$$

With regard to the maximum norm, the error is actually reduced by the factor q in both cases. However, based on the components, one can also see that the iterate of the Gauss-Seidel method is somewhat better (and thus also the error, e.g. with regard to the Euclid norm). //.

Although examples can be constructed for which the Jacobi method is superior (cf. exercise), the Gauss-Seidel method often converges faster than the Jacobi method. Such comparisons can be made more precise for special matrices A. In the following, such a comparison for matrices of the form

$$A = \begin{pmatrix} E & -B^T \\ -B & E \end{pmatrix} \in \mathbb{R}^{n\times n} \tag{3.14}$$

with $B \in \mathbb{R}^{p\times q},\ 0 < p,\ q < n,\ p+q = n$ will be shown as an example.

In the present case it is $D = E$ and

$$L = \begin{pmatrix} 0 & 0 \\ B & 0 \end{pmatrix}, \quad R = \begin{pmatrix} 0 & B^T \\ 0 & 0 \end{pmatrix}.$$

Hence, we obtain for the Jacobi iterative matrix and the Gauss-Seidel matrix

$$\mathcal{L} = \begin{pmatrix} E & 0 \\ -B & E \end{pmatrix}^{-1} \begin{pmatrix} 0 & B^T \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} E & 0 \\ B & E \end{pmatrix} \begin{pmatrix} 0 & B^T \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & B^T \\ 0 & BB^T \end{pmatrix} \tag{3.15}$$

or

$$\mathcal{J} = \begin{pmatrix} 0 & B^T \\ B & 0 \end{pmatrix}$$

respectively and therefore

$$\mathcal{J}^2 = \begin{pmatrix} B^TB & 0 \\ 0 & BB^T \end{pmatrix}. \tag{3.16}$$

The goal now is to show that $\rho(\mathcal{L}) = \rho(\mathcal{J}^2) = \rho(\mathcal{J})^2$ (we already saw this in Example 3.7). For this we need an auxiliary lemma.

Lemma 3.11. *Given $X \in \mathbb{R}^{p\times q},\ Y \in \mathbb{R}^{q\times p}$ and $Z \in \mathbb{R}^{n\times n}$ with $p, q, n \in \mathbb{N}$. Then*

(a) $\sigma(XY)\backslash\{0\} = \sigma(YX)\backslash\{0\}$

(b) $\sigma(Z^2) = \{\lambda^2 \mid \lambda \in \sigma(Z)\} =: \sigma(Z)^2$

Proof. **Ad (a):** If $\lambda \in \sigma(XY)\backslash\{0\}$ then there exists an eigenvector $u \neq 0$ with $XYu = \lambda u \neq 0$. Hence, we also have $v := Yu \neq 0$ and thus

$$YXv = Y(XYu) = Y(\lambda u) = \lambda Yu = \lambda v.$$

Consequently, we obtain $\lambda \in \sigma(YX)\backslash\{0\}$ and therefore $\sigma(YX)\backslash\{0\} \subset \sigma(XY)\backslash\{0\}$. With the same argument one also sees that $\sigma(YX)\backslash\{0\} \subset \sigma(XY)\backslash\{0\}$.

Ad (b): If $\lambda \in \sigma(Z)$ then there exists some $x \neq 0$ with $Zx = \lambda x$ and from this we deduce

$$Z^2x = Z(\lambda x) = \lambda Zx = \lambda^2 x.$$

Thus, it is $\lambda^2 \in \sigma(Z^2)$. Conversely, if $\mu \in \sigma(Z^2)$ and if $\pm\lambda$ are the (possibly complex) roots of μ, then it follows that

$$0 = \det(Z^2 - \mu E) = \det\big((Z - \lambda E)(Z + \lambda E)\big) = \det(Z - \lambda E)\det(Z + \lambda E).$$

From this we can conclude that λ or $-\lambda$ is an eigenvalue of Z and the assertion follows. ∎

Proposition 3.12. *If A has the form (3.14), then $\rho(\mathcal{L}) = \rho(\mathcal{J})^2$.*

Proof. From Equation 3.15 we obtain

$$\det(\mathcal{L} - \lambda E) = \det\begin{pmatrix} -\lambda E & B^T \\ 0 & BB^T - \lambda E \end{pmatrix} = (-\lambda)^q \det(BB^T - \lambda E).$$

So $\sigma(\mathcal{L}) = \{0\} \cup \sigma(BB^T)$ and thus $\rho(\mathcal{L}) = \rho(BB^T)$. From the representation

$$\mathcal{J}^2 = \begin{pmatrix} B^T B & 0 \\ 0 & BB^T \end{pmatrix}$$

we conclude

$$\begin{aligned} \sigma(\mathcal{J}^2) &= \{\lambda \mid \det(\mathcal{J}^2 - \lambda E) = 0\} \\ &= \{\lambda \mid \det(B^T B - \lambda E)\cdot\det(BB^T - \lambda E) = 0\} \\ &= \sigma(B^T B) \cup \sigma(BB^T) \end{aligned}$$

and further using Lemma 3.11 (a)

$$\begin{aligned} \sigma(\mathcal{J}^2) \cup \{0\} &= \big(\sigma(B^T B) \cup \{0\}\big) \cup \big(\sigma(BB^T) \cup \{0\}\big) \\ &= \sigma(BB^T) \cup \{0\} \end{aligned}$$

Employing Lemma 3.11 (b) we finally obtain

$$\rho(\mathcal{J})^2 = \rho(\mathcal{J}^2) = \rho(BB^T) = \rho(\mathcal{L}),$$

completing the proof. ∎

In summary we have shown: For matrices of the form (3.14) it either holds

(a) $\rho(\mathcal{J}) < 1 \;\Rightarrow\; \rho(\mathcal{L}) = \rho(\mathcal{J})^2 < 1$, i.e. both methods converge, or

(b) $\rho(\mathcal{J}) \geq 1 \;\Rightarrow\; \rho(\mathcal{L}) = \rho(\mathcal{J})^2 \geq 1$, i.e. both methods diverge.

In the convergent case, the Gauss-Seidel method (roughly estimated) needs only half as many iterations as the Jacobi-method.

4. Linear Least-Squares Problems

If the number of equations exceeds that of the unknowns, the system of equations is overdetermined and is usually not solvable. Overdetermined equation systems $Ax = b$ with $A \in \mathbb{R}^{m\times n}$ $(m > n)$ often occur in applications. It often makes sense to search the solution $\hat{x}$ of the so-called linear least-squares problem

$$\text{minimize} \quad \|Ax - b\|_2 \;\; \text{for } x \in \mathbb{R}^n.$$

Problems of this type often occur when fitting measurement data. Given a model function, which is supposed to depend linearly from certain parameters $x_1, \dots, x_n$. This must now be optimally adapted to a large amount of measurement data. We are therefore looking for the optimal fit of a model function to a set of given points. The number of points given is often much larger than the number of parameters to be fitted and this then leads to a problem of the above type. The simplest example is the determination of the optimal linear regression line for a set of points in the x-y plane.

4.1. The Method of Normal Equations

In the following we assume $A \in \mathbb{R}^{m\times n}$ for $m > n$ and $b \in \mathbb{R}^m$. In this section we will show that the solution to the least-squares problem can be reduced to the solution of a system of linear equations, the so-called normal equations. These are given in our notation by

$$A^T Ax = A^T b.$$

Proposition 4.1. *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The minimizers of*

$$x \mapsto \|Ax - b\|_2$$

are precisely the solutions of

$$A^T Ax = A^T b. \tag{4.1}$$

Before we prove this proposition, we need a lemma concerning the properties of the Matrix $A^T A$ in the normal equations.

Lemma 4.2. *Let $A \in \mathbb{R}^{m\times n}$ with $m > n$. Then the matrix $A^T A \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite. Further, $A^T A$ is positive definite if and only if one of the following conditions is satisfied:*

(a) $\ker(A) = \{0\}$

(b) $\operatorname{rank}(A) = n$.

Moreover, in any case it is

$$\ker(A^T A) = \ker(A) \quad \textit{and} \quad \operatorname{im}(A^T A) = \operatorname{im}(A^T) = \ker(A)^\perp.$$

Proof. Obviously, $A^T A$ is symmetric. Further, for all $x \in \mathbb{R}^n$ one has

$$\langle x, A^T Ax\rangle = \|Ax\|_2^2 \geq 0 \tag{4.2}$$

From Equation (4.2) it follows that $A^T A$ is positive definit if and only if $\ker(A) = \{0\}$. The last condition is equivalent to $\operatorname{rank}(A) = n$, because according to the dimension formula:

$$\dim\ker(A) + \operatorname{rank}(A) = n.$$

Equation (4.2) also shows that

$$\ker(A^T A) \subset \ker(A).$$

The opposite inclusion

$$\ker(A) \subset \ker(A^T A)$$

is obvious and so it follows $\ker(A) = \ker(A^T A)$. Furthermore, one immediately sees

$$\operatorname{im}(A^T A) \subset \operatorname{im}(A^T)$$

and using what we have already shown for $\ker(A^T A)$, we again obtain using the dimension formula

$$\begin{aligned}\dim \operatorname{im}(A^T A) &= n - \ker(A^T A)\\ &= n - \dim \ker(A)\\ &= \dim \operatorname{im}(A)\\ &= \operatorname{rank}(A)\\ &= \operatorname{rank}(A^T)\\ &= \dim \operatorname{im}(A^T).\end{aligned}$$

Thus we have proved that $\operatorname{im}(A^T A) = \operatorname{im}(A^T)$ and it remains to show that $\operatorname{im}(A^T)$ and $\ker(A)$ are orthogonal complementary subspaces of $\mathbb{R}^n$. For this, let $z \in \operatorname{im}(A^T)$ and $x \in \ker(A)$ be arbitrary chosen. We need to show that

$$\langle x, z \rangle = 0.$$

Since $z \in \operatorname{im}(A^T)$ there exists a $y \in \mathbb{R}^m$ with $z = A^T y$ and it follows

$$\langle x, z \rangle = \langle x, A^T y \rangle = \langle Ax, y \rangle = \langle 0, y \rangle = 0.$$

Hence $\operatorname{im}(A^T)$ and $\ker(A)$ are orthogonal and since

$$\dim \ker(A) + \operatorname{rank}(A^T) = n$$

we obtain that $\operatorname{im}(A^T) = \ker(A)^\perp$. ■

Now we turn to the proof of Proposition 4.1.

PROOF (PROPOSITION 4.1). For $x, y \in \mathbb{R}^n$ consider the symmetric bilinear form

$$\psi(x, y) := \langle Ax, Ay \rangle.$$

Since $\operatorname{im}(A^T) \subset \operatorname{im}(A^T A)$, we may consider a solution $\hat{x}$ of the normal equations $A^T A x = A^T b$. Using

$$\langle Ax, b \rangle = \langle x, A^T b \rangle = \langle x, A^T A \hat{x} \rangle = \psi(x, \hat{x})$$

we first obtain

$$\mathcal{E}(x) := \|Ax - b\|_2^2 = \psi(x, x) - 2\psi(x, \hat{x}) + \|b\|_2^2$$

and from this for all $x \in \mathbb{R}^n$

$$\begin{aligned}\mathcal{E}(x) - \mathcal{E}(\hat{x}) &= \psi(x, x) - 2\psi(x, \hat{x}) + \psi(\hat{x}, \hat{x})\\ &= \psi(x - \hat{x}, x - \hat{x})\\ &= \|A(x - \hat{x})\|_2^2\\ &\geq 0,\end{aligned} \tag{4.3}$$

i.e.

$$\mathcal{E}(x) \geq \mathcal{E}(\hat{x}). \tag{4.4}$$

Therefore, every solution of the normal equations is also a minimizer of $\mathcal{E}$ on $\mathbb{R}^n$. We proceed to show that the converse also holds true: Assuming $x = z$ is a minimizer of $\mathcal{E}$, it follows from (4.4) that $\mathcal{E}(z) = \mathcal{E}(\hat{x})$. But this implies that

$$z - \hat{x} =: y \overset{(4.3)}{\in} \ker(A) = \ker(A^T A)$$

and thus

$$z = \hat{x} + y$$

is also a solution of the normal equations. ■

It should be noted that in the case that $\operatorname{rank}(A) < n$, the matrix $A^T A$ is singular and thus the normal equations have infinitely many solutions. Namely, in such a case we have

$$\dim\ker(A^T A) = \dim\ker(A) = n - \operatorname{rank}(A) > 0.$$

One can then search among those solutions that have certain further optimality properties. The solution theory for singular $A^T A$ leads to the problem of the pseudoinverse and the singular value decomposition (but we don't want to go into that here).

In the following we always assume $\operatorname{rank}(A) = n$. After the proof of the last Proposition or Lemma, it follows that the global minimum of the initial problem is attained at exactly one point, namely at the unique solution of the normal equations.

Remark.

(a) The normal equations can be derived as follows: Let $x \in \mathbb{R}^n$ and $i = 1, \ldots, n$ be arbitrary and

$$x(t) := x + te_i.$$

First, note that by the chain rule for curves we have

$$\frac{d}{dt}\Big|_{t=0} \mathcal{E}(x(t)) = \nabla\mathcal{E}(x(0))^T \dot{x}(0) = \nabla\mathcal{E}(x)^T e_i = \frac{\partial\mathcal{E}}{\partial x_i}(x).$$

Using

$$\begin{aligned}\mathcal{E}(x(t)) &= \mathcal{E}(x + te_i)\\ &= \langle A(x + te_i) - b, A(x + te_i) - b\rangle\\ &= \langle tAe_i + Ax - b, tAe_i + Ax - b\rangle\\ &= t^2\langle Ae_i, Ae_i\rangle + 2t\langle Ae_i, Ax - b\rangle + \|Ax - b\|_2^2\end{aligned}$$

we obtain

$$\begin{aligned}\frac{\partial\mathcal{E}}{\partial x_i}(x) &= \frac{d}{dt}\Big|_{t=0}\mathcal{E}(x(t))\\ &= 2\langle Ae_i, Ax - b\rangle\\ &= 2\langle e_i, A^T Ax - A^T b\rangle\\ &= 2(A^T Ax - A^T b)_i.\end{aligned}$$

Now every minimizer $\hat{x}$ of $\mathcal{E}$ is also a critical point of $\mathcal{E}$, i.e.

$$0 = \nabla\mathcal{E}(\hat{x}) = 2(A^T A\hat{x} - A^T b).$$

This shows that every minimizer solves the normal equations.

(b) If $\operatorname{rank}(A) = n$ we have seen that the matrix $A^T A$ is symmetric and positive definite, hence the Cholesky decomposition is a suitable method to solve the normal equations. Note, however, that the matrix $A^T A$ generally has a worse condition than the matrix A! ♢

4.2. Examples

Example 4.3. Given the data points

x	0	3	4	7
y	1	2	6	4

find the line $y = \alpha + \beta x$ (defined by the coefficients α and β) that minimizes the sum of squared residuals

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

Such a line is called the linear regression line for the $n = 4$ points $(x_i, y_i)^T$ given above. This leads to the system of linear equations

$$A \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = b \quad \text{with} \quad A = \begin{pmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \\ 6 \\ 4 \end{pmatrix},$$

which has to be solved in the least-squares sense. Since

$$A^T A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 3 & 4 & 7 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 3 \\ 1 & 4 \\ 1 & 7 \end{pmatrix} = \begin{pmatrix} 4 & 14 \\ 14 & 74 \end{pmatrix}$$

and

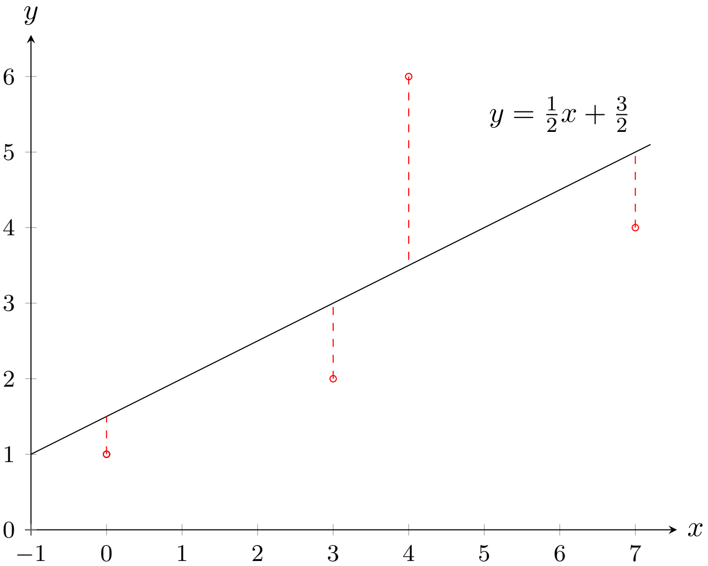
$$A^T b = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 3 & 4 & 7 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 6 \\ 4 \end{pmatrix} = \begin{pmatrix} 13 \\ 58 \end{pmatrix}$$

the corresponding normal equations are

$$\begin{pmatrix} 4 & 14 \\ 14 & 74 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 13 \\ 58 \end{pmatrix}.$$

This system has the unique solution (we already knew the uniqueness beforehand, because it is $\operatorname{rank}(A) = 2$)

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \frac{3}{2} \\ \frac{1}{2} \end{pmatrix}.$$



The corresponding regression line is shown in the figure above. The sum of squares of the lengths of all dashed lines is minimized by this straight line. //.

In what follows, we want to take a closer look at the general case of fitting a set of points by a polynomial of degree k. Let N points be given with the following coordinates:

x	x_1	x_2	$\cdots$	x_N
y	y_1	y_2	$\cdots$	y_N

The fit should be done with a polynomial of degree k, i.e.

$$p(x) = a_k x^k + a_{k-1}x^{k-1} + \ldots + a_1 x + a_0$$

is intended to minimize the distance to the given points in the least-squares sense. This leads to the problem:

$$\text{minimize} \quad \|Aa - b\|_2^2 \quad \text{for } a \in \mathbb{R}^{k+1}$$

with

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 \cdots & x_1^k \\ 1 & x_2 & x_2^2 \cdots & x_2^k \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \cdots & x_N^k \end{pmatrix} \in \mathbb{R}^{N\times(k+1)}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} \in \mathbb{R}^{k+1}, \quad b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \in \mathbb{R}^N. \tag{4.5}$$

In order for this example to fit into the framework presented earlier, we require $k \le N-1$ since we want to guarantee uniqueness of the fit. In fact, the fit is unique:

Proposition 4.4. *Given N data points in $\mathbb{R}^2$ having pairwise different first coordinates. Then for all $k \le N-1$ there is a unique polynomial of degree k, which minimizes the distance to the data points in the least-squares sense.*

Before we start with the proof of this proposition, we first determine the so-called Vandermonde determinant as a preliminary result.

Lemma 4.5. *Assume that $r_1, r_2, \ldots, r_n \in \mathbb{R}$ and $n \ge 2$. Then for*

$$A = \begin{pmatrix} 1 & r_1 & r_1^2 \cdots & r_1^{n-1} \\ 1 & r_2 & r_2^2 \cdots & r_2^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & r_n & r_n^2 \cdots & r_n^{n-1} \end{pmatrix}$$

it follows

$$\det A = \prod_{1\le i<j\le n} (r_j - r_i).$$

Proof. We prove this by induction on $n \ge 2$. For $n = 2$ we have

$$\det \begin{pmatrix} 1 & r_1 \\ 1 & r_2 \end{pmatrix} = r_2 - r_1 = \prod_{1\le i<j\le 2} (r_j - r_i).$$

Now let $n \ge 3$ and assume the theorem proved for $n-1$. Apply the following operation to columns $n-1, \dots, 1$ in order: multiply each column by r_1 and subtract it from the next column on the right. Doing this, we obtain from A the new matrix

$$A' = \begin{pmatrix} 1 & 0 & 0 & \cdots\cdots & 0 \\ 1 & r_2 - r_1 & r_2^2 - r_1 r_2 & \cdots\cdots & r_2^{n-1} - r_1 r_2^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & r_n - r_1 & r_n^2 - r_1 r_n & \cdots\cdots & r_n^{n-1} - r_1 r_n^{n-2} \end{pmatrix}$$

$$= \left(\begin{array}{c|cccc} 1 & 0 & 0 & \cdots\cdots & 0 \\ \hline 1 & r_2 - r_1 & (r_2 - r_1) r_2 & \cdots\cdots & (r_2 - r_1) r_2^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & r_n - r_1 & (r_n - r_1) r_n & \cdots\cdots & (r_n - r_1) r_n^{n-2} \end{array}\right).$$

Since elementary column transformations do not change the determinant we obtain

$$\det(A) = \det(A')$$

and with Laplace's expansion theorem when expanding by the first line it follows

$$\det(A) = \det(A') = \det \begin{pmatrix} r_2 - r_1 & (r_2 - r_1) r_2 & \cdots\cdots & (r_2 - r_1) r_2^{n-2} \\ \vdots & \vdots & & \vdots \\ r_n - r_1 & (r_n - r_1) r_n & \cdots\cdots & (r_n - r_1) r_n^{n-2} \end{pmatrix}.$$

The multilinearity of the determinant and the induction hypothesis now yield

$$\begin{aligned} \det(A) &= (r_2 - r_1) \cdot \ldots \cdot (r_n - r_1) \cdot \det \begin{pmatrix} 1 & r_2 & r_2^2 & \cdots & r_2^{n-2} \\ 1 & r_3 & r_3^2 & \cdots & r_3^{n-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & r_n & r_n^2 & \cdots & r_n^{n-2} \end{pmatrix} \\ &= (r_2 - r_1) \cdot \ldots \cdot (r_n - r_1) \cdot \prod_{2\le i<j\le n} (r_j - r_i) \\ &= \prod_{1\le i<j\le n} (r_j - r_i). \end{aligned}$$

■

Proof (of Proposition 4.4). From Lemma 4.5 it follows that the linear space of the first $k+1$ rows of A (cf. (4.5)) has dimension $k+1$, yielding that the matrix A has full rank. But since

$$\operatorname{rank}(A) = \operatorname{rank}(A^T A)$$

this also holds true for $A^T A$ and therefore the normal equations have a unique solution. ■

We calculate another concrete

Example 4.6. The optimal 2nd degree polynomial is sought for the following data points

x	1	2	3	4
y	2	1	0	1

So we are looking for the optimal polynomial

$$p(x) = a_0 + a_1 x + a_2 x^2.$$

The corresponding matrix A and the vector b are

$$A = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{pmatrix}, \quad b = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 0 \\ 1 \end{pmatrix}.$$

From this we deduce

$$A^T A = \begin{pmatrix} 4 & 10 & 30 \\ 10 & 30 & 100 \\ 30 & 100 & 354 \end{pmatrix}, \quad A^T b = \begin{pmatrix} 4 \\ 8 \\ 22 \end{pmatrix}$$

and the unique solution of the normal equations

$$A^T A a = A^T b$$

gives the optimal parameters of the polynomial p. Namely, we obtain

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 4.5 \\ -2.9 \\ 0.5 \end{pmatrix},$$

so that the optimal polynomial is given by

$$p(x) = \frac{1}{2}x^2 - \frac{29}{10}x + \frac{9}{2},$$

see Figure 4.1.

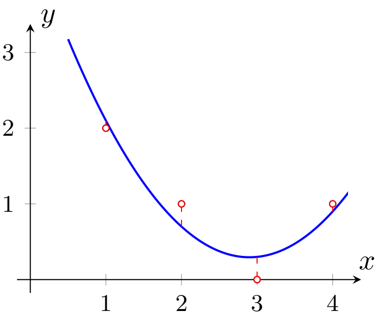


Fig. 4.1.: Graph of p with data points

//.

Example 4.7. In principle, the theory presented here works for every model function in which the parameters to be optimized are linear. So, for example, the function

$$f(x) = a_0 + a_1 x + a_2 e^x + a_3 e^{-x}$$

can be optimally adapted to many data points in the same way, see exercises. //.

5. Nonlinear Equations

After the linear systems of equations, the treatment of non-linear equations of one and more variables now follows. Nonlinear equations are usually formulated as a root-finding problem. That is, the zero of a mapping

$$F\colon \operatorname{dom}(F) \subset \mathbb{R}^n \to \mathbb{R}^n$$

or, somewhat weakend, the minimum

$$\min_{x \in \operatorname{dom}(F)} \|F(x)\|_2^2$$

is sought. Here and in the following, $\operatorname{dom}(F)$ denotes the domain of definition of F, which is assumed to be open and connected. Since non-linear equations cannot usually be solved in closed form, i.e. their exact solution (if it exists) cannot be calculated in a finite number of steps, iteration methods are used almost exclusively to approximate the solution.

5.1. Convergence and Rate of Convergence

Of course, one would like the iteration sequence $x^{(k)}$ to approach the limit $\hat{x}$ as quickly as possible. A measure for the speed of convergence of a sequence is the concept of **rate of convergence** (alternatively: **order of convergence**).

Definition 5.1. *A convergent sequence $(x^{(k)})_{k\in\mathbb{N}}$ in $\mathbb{R}^n$ with limit $\hat{x}$ has convergence rate $p \geq 1$, if it exists $k_0 \in \mathbb{N}$ such that*

$$\|x^{(k+1)} - \hat{x}\| \leq c\|x^{(k)} - \hat{x}\|^p$$

for all $k \geq k_0$, where

$$0 < c < 1 \quad \text{if} \;\; p = 1. \qquad /\!\!/.$$

The next example illustrates the large difference in speed between methods of order $p = 1$ (linear convergence) and methods of order $p = 2$ (quadratic convergence).

Example 5.2. Let $x^{(k)}$ be a convergent sequence in $\mathbb{R}^n$ with $\|x^{(0)} - \hat{x}\| = 0.2$. For $p = 1$ and $c = \frac{1}{2}$ it then follows

k	1	2	3	4	5	6
$\|x^{(k)} - \hat{x}\| \leq$	0.1	0.05	0.025	0.0125	$6.25\cdot10^{-3}$	$3.125\cdot10^{-3}$

and for $p = 2$ and e.g. $c = 3$ we obtain

k	1	2	3	4	5	6
$\|x^{(k)} - \hat{x}\| \leq$	0.12	0.0432	0.0056	$9.4\cdot10^{-5}$	$3\cdot10^{-8}$	$2\cdot10^{-15}$

$/\!\!/.$

The choice of the norm is irrelevant for the order of convergence for $p > 1$: If the sequence has an order of convergence $p > 1$ wrt a vector norm $\|\cdot\|$, then it has this order of convergence (with a different constant c) wrt every vector space norm for all norms in $\mathbb{R}^n$ are equivalent.

We turn again to iterative methods of the form

$$x^{(k+1)} = \Phi(x^{(k)})$$

but now with generally nonlinear Φ (in contrast to affine linear $\Phi(x) = Tx + c$ in connection with SLE). We call an iterative procedure for determining $\hat{x}$ **locally convergent** with convergence rate $p \geq 1$, if there is a neighbourhood U of $\hat{x}$, such that for all initial values $x^{(0)} \in U$ it holds that the sequence induced by Φ converges to $\hat{x}$ with convergence rate p. Furthermore, the method is said to be globally convergent if the method converges for all initial values $x^{(0)} \in \mathrm{dom}(\Phi)$.

In `095_fixed_point_method.ggb` with $\Phi(x) = x^2 - x$ one sees that the iterative procedure for determining $\hat{x} = 2$ is not locally convergent, but for $\hat{x} = 0$ it is.

We first consider the general case :

$$\Phi \colon \mathbb{R}^n \to \mathbb{R}^n$$

and examine the iteration

$$x^{(k+1)} = \Phi(x^{(k)})$$

with initial value $x^{(0)}$ for convergence. The Banach fixed-point theorem now provides the following result:

Proposition 5.3. *Let the function* $\phi\colon \mathrm{dom}(\Phi) \subset \mathbb{R}^n \to \mathbb{R}^n$ *be continuously differentiable and have a fixed point* $\hat{x}$ *in* $\mathrm{dom}(\Phi)$. *Furthermore let* $\|\cdot\|_*$ *be a matrix norm on* $\mathbb{R}^{n\times n}$ *that is compatible with a vector norm* $\|\cdot\|$ *in* $\mathbb{R}^n$ *with*

$$\|J\Phi(\hat{x})\|_* < 1.$$

Then Φ *is a contraction in a neighbourhood* U *of* $\hat{x}$ *and the fixed point iteration*

$$x^{(k+1)} = \Phi(x^{(k)})$$

converges locally to $\hat{x}$. *That is, the method converges for all initial values* $x^{(0)} \in U$ *and the convergence rate is at least linear.*

Proof. Because of the continuity of $J\Phi$ wrt $\|\cdot\|_*$ it follows that there is a closed ball $B := \overline{B_r(\hat{x})} \subset \mathrm{dom}(\phi)$ around $\hat{x}$ with radius $r > 0$ such that

$$\|J\Phi(x)\|_* \leq q < 1 \quad \text{for all } x \in B.$$

Now, for arbitrary but fixed $x, y \in B$ we define the function

$$f\colon [0,1] \to \mathbb{R}^n, \quad f(t) = \Phi(x + t(y - x)).$$

Then, by the fundamental theorem of calculus and the chain rule, we can deduce

$$\Phi(y) - \Phi(x) = f(1) - f(0) = \int_0^1 f'(t)\,dt = \int_0^1 J\Phi\big(x + t(y-x)\big)(y-x)\,dt.$$

Thus, it follows that

$$\begin{aligned}
\|\Phi(y) - \Phi(x)\| &\overset{B.2}{=} \left\| \int_0^1 J\Phi\big(x + t(y-x)\big)(y-x)\,dt \right\| \\
&\leq \int_0^1 \|J\Phi\big(x + t(y-x)\big)(y-x)\|\,dt \\
&\leq \int_0^1 \|J\Phi\big(x + t(y-x)\big)\|_* \|(y-x)\|\,dt \\
&\leq q\|y - x\|
\end{aligned} \tag{5.1}$$

for all $x, y \in B$. Especially for $y = \hat{x}$ this yields

$$\|\Phi(x) - \hat{x}\| \leq q\|x - \hat{x}\| \leq qr < r.$$

That is, Φ is a contracting self-map of B. The rest of the claim now follows with Banach's fixed point theorem. ∎

Remark. The triangle inequality (for integrals) also holds for curves $f\colon [a,b] \to \mathbb{R}^n$, i.e. the norm may be drawn under the integral sign in Equation 5.1, because for every continuous vector-valued function

$$f\colon [0,1] \to \mathbb{R}^n$$

it follows when representing the following integral with the help of a Riemann sum using the continuity of norms in $\mathbb{R}^n$:

$$\begin{aligned}\Big\| \int_0^1 f(t)\,dt \Big\| &= \Big\| \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} f\Big(\frac{j}{N}\Big) \Big\| \\ &= \lim_{N\to\infty} \frac{1}{N} \Big\| \sum_{j=1}^{N} f\Big(\frac{j}{N}\Big) \Big\| \\ &= \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \Big\| f\Big(\frac{j}{N}\Big) \Big\| \\ &= \int_0^1 \|f(t)\|\,dt. \end{aligned}$$ ◊

We first treat the case of a nonlinear scalar equation.

5.2. Calculation of roots of a nonlinear scalar equation

In this case one has an equation of the form

$$f(x) = 0$$

with a function $f\colon \operatorname{dom}(f) \subset \mathbb{R} \to \mathbb{R}$. Possible iteration methods then always have the form

$$x^{(k+1)} = \Phi(x^{(k)})$$

with $\Phi\colon \operatorname{dom}(\Phi) \subset \mathbb{R} \to \mathbb{R}$. If one is now looking for a fast procedure for determining a zero $\hat{x}$, the last proposition or Banach's fixed point theorem shows that the convergence should be faster the smaller

$$\|J\Phi(\hat{x})\| < 1,$$

or, in our one-dimensional case,

$$|\Phi'(\hat{x})| < 1$$

is. In the case $\Phi'(\hat{x}) = 0$ even the following theorem holds:

Proposition 5.4. *Let the iteration function $\Phi\colon \operatorname{dom}(\Phi) \subset \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable and assume there is a fixed point $\hat{x} \in \operatorname{dom}(\Phi)$. Furthermore, assume that $\Phi'(\hat{x}) = 0$. Then the iteration procedure*

$$x^{(k+1)} = \Phi(x^{(k)})$$

converges locally quadratically to $\hat{x}$.

Proof. Since $\Phi(\hat{x}) = 0$, local (i.e. for all initial values in a neighbourhood B of $\hat{x}$) convergence follows with Proposition 5.3. The Taylor expansion for Φ now results in the Lagrange's form of the remainder

$$x^{(k+1)} - \hat{x} = \Phi(x^{(k)}) - \Phi(\hat{x}) = \underbrace{\Phi'(\hat{x})}_{=0}(x^{(k)} - \hat{x}) + \frac{1}{2}\Phi''(\xi)(x^{(k)} - \hat{x})^2$$

with ξ between x_k and $\hat{x}$. Letting

$$K := \max_{\xi \in B} \Phi''(\xi)$$

it follows that

$$|x^{(k+1)} - \hat{x}| \leq \frac{1}{2} K |x^{(k)} - \hat{x}|^2 \quad \text{for all } k \in \mathbb{N},$$

if an initial guess was chosen in B ($\Rightarrow x_k \in B$ for all k). Thus, quadratic convergence has been proven. ■

The most important method with local quadratic convergence is Newton's method, which we will discuss in the following.

5.2.1. Newton's Method

We are looking for a zero $\hat{x}$ of a function f. We further assume that $x^{(k)}$ is an already known approximation for this zero. For a small neighbourhood around $x^{(k)}$, the tangent $T_{f;x^{(k)}}(x)$ is a good approximation of the function f. The tangent is the Taylor polynomial of the first order, i.e. the straight line

$$T(x) := T_{f;x^{(k)}}(x) = f(x^{(k)}) + f'(x^{(k)})(x - x^{(k)}).$$

So if $x^{(k)}$ is a good approximation for the zero $\hat{x}$, we expect the zero of $T(x)$ to be a good approximation for the zero $\hat{x}$ of f. The zero of the tangent exists if $f'(x^{(k)}) \neq 0$ and we take this as the next iterate $x^{(k+1)}$ of our method. This then calculates to

$$T(x^{(k+1)}) = 0 \quad \Leftrightarrow \quad x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})},$$

and this (iteration) rule is just Newton's method.

Proposition 5.5. *Let $f\colon \operatorname{dom}(f) \to \mathbb{R}$ be three times continuously differentiable and assume that for $\hat{x} \in \operatorname{dom}(f)$*

$$f(\hat{x}) = 0 \quad \textit{and} \quad f'(\hat{x}) \neq 0.$$

Then Newton's method converges locally quadratically to $\hat{x}$. (cf. section C.1, namely Proposition C.1.)

PROOF. We apply Proposition 5.4. The iteration function in the Newton method is given by

$$\Phi(x) = x - \frac{f(x)}{f'(x)}$$

so that

$$\Phi'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}. \tag{5.2}$$

Thus $\Phi(\hat{x}) = \hat{x}$, i.e. the zero of f is a fixed point of Φ and inserting $\hat{x}$ in (5.2) together with $f'(\hat{x}) \neq 0$ yields $\Phi'(\hat{x}) = 0$. Proposition 5.4 now proves the assertion. ■

Example 5.6.

(a) What happens when you press the square root key on the calculator? The newton method is used here. More precisely, a zero is sought for the function

$$f(x) = x^2 - a \quad \text{for } a > 0.$$

Newton's method in this case is

$$x^{(k+1)} = \frac{1}{2}\Big(x^{(k)} + \frac{a}{x^{(k)}}\Big).$$

To determine $\sqrt{2}$ we calculate for $a = 2$ the first 10 iterates to the (bad) initial guess $x^{(0)} = 100$:

k	$x^{(k)}$	$\sqrt{2} - x^{(k)}$
0	100.0	$-9.86e+01$
1	50.01	$-4.86e+01$
2	25.02	$-2.36e+01$
3	12.55	$-1.11e+01$
4	6.356	$-4.94e+00$
5	3.335	$-1.92e+00$
6	1.967	$-5.53e-01$
7	1.492	$-7.78e-02$
8	1.416	$-2.03e-03$
9	1.414	$-1.45e-06$
10	1.414	$-7.45e-13$

The second column was rounded to 3 decimals after the decimal point from the 4th iteration onwards. The error in the third column, however, was determined exactly in each case. You can see the halving of the error up to about $k = 6$. Then the quadratic convergence sets in, whereby the number of digits is approximately doubled per step. Although up to now in connection with Newton's method there has always been talk of local convergence in connection with Newton's method, here we find that Newton's method here for each initial value $x^{(0)} > 0$ converges to $\sqrt{2}$. To prove this, we let $x^{(0)} > 0$ be arbitrary. Then from

$$x^{(k+1)} - \sqrt{a} = \frac{1}{2}\left(x^{(k)} + \frac{a}{x^{(k)}}\right) - \sqrt{a} = \frac{1}{2x^{(k)}}\left(x^{(k)} - \sqrt{a}\right)^2 \geq 0 \quad \text{if } x^{(k)} > 0$$

it follows that $x^{(k)} \geq \sqrt{a}$ holds for all $k \geq 1$. From this we obtain

$$0 \leq x^{(k+1)} - \sqrt{a} = \frac{1}{2}\frac{x^{(k)} - \sqrt{a}}{x^{(k)}}(x^{(k)} - \sqrt{a}) \leq \frac{1}{2}(x^{(k)} - \sqrt{a}) \quad \text{for } k \geq 1.$$

So the error is at least halved in each step, which we can also confirm for the first iterations starting from the bad initial value $x^{(0)} = 100$ in the table above.

(b) (general Heron's Method:) To determine the pth root of the number a, we apply Newton's method to the function

$$f(x) = x^p - a.$$

The iteration formula now reads

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

(c) Here, we want to determine the root $\hat{x} \in [0, 2]$ of the function

$$f(x) = x^6 - x - 1.$$

Newton's method is

$$x^{(k+1)} = x^{(k)} - \frac{\left(x^{(k)}\right)^6 - x^{(k)} - 1}{6\left(x^{(k)}\right)^5 - 1}$$

here. By the intermediate value theorem, the function f has a zero in the interval $[0, 2]$, since $f(0) = -1 < 0$ and $f(2) = 61 > 0$. The following table shows the iterates for the starting values $x^{(0)} = 0.5$ and $x^{(0)} = 2$. Since for Newton's method one usually has only local convergence, the starting value $x^{(0)}$ must be sufficiently good to reach the desired root. Here, for example, the initial value $x^{(0)} = 0.5$ is not good enough; in this case Newton's method converges to another root of f outside the interval.

k	$x^{(k)}$ for $x^{(0)} = 0.5$	$x^{(k)}$ for $x^{(0)} = 2.0$
0	0.50000	2.00000
1	-1.32692	1.68063
2	-1.10165	1.43074
3	-0.92568	1.25497
4	-0.81642	1.16154
5	-0.78099	1.13635
6	-0.77811	1.13473
7	-0.77809	1.13472

//.

For Newton's method, the constant for the quadratic rate of convergence, i.e. c in the inequality

$$|x^{(k+1)} - \hat{x}| \leq c|x^{(k)} - \hat{x}|^2,$$

can be given more precisely:

Proposition 5.7. *Let f be twice continuously differentiable in a neighbourhood $U = (a;b)$ of $\hat{x}$ and let*

$$f(\hat{x}) = 0 \quad \textit{and} \quad f'(\hat{x}) \neq 0.$$

Further let $x^{(k)} \in U$ and assume that the next iteration of Newton's method

$$x^{(k+1)} = x^{(k)} - \frac{f(x^{(k)})}{f'(x^{(k)})}$$

is also in U (this condition is certainly fulfilled for U small enough). Then there exists a $\xi^{(k)} \in U$, such that

$$x^{(k+1)} - \hat{x} = \frac{f''(\xi^{(k)})}{f'(x^{(k)})}(x^{(k)} - \hat{x})^2.$$

So if

$$K := \frac{\sup_{x\in U}|f''(x)|}{\inf_{x\in U}|f'(x)|} > 0$$

holds true, which can be achieved by reducing U if necessary, then

$$|x^{(k+1)} - \hat{x}| \leq \frac{1}{2}K|x^{(k)} - \hat{x}|^2$$

holds true for all k large enough.

Proof. We use the Taylor expansion of f around $x^{(k)}$ with the remainder term in Lagrangian form

$$f(x) = f(x^{(k)}) + (x - x^{(k)})f'(x^{(k)}) + \frac{1}{2}(x - x^{(k)})^2 f''(\xi^{(k)}),$$

where $\xi^{(k)}$ is between x and $x^{(k)}$, i.e. $\xi^{(k)} \in U$. Substitution of $x = \hat{x}$ yields

$$0 = f(\hat{x}) = f(x^{(k)}) + (\hat{x} - x^{(k)})f'(x^{(k)}) + \frac{1}{2}(\hat{x} - x^{(k)})^2 f''(\xi^{(k)}) \quad \text{with } \xi^{(k)} \in U,$$

and therefore

$$-\frac{f(x^{(k)})}{f'(x^k)} + x^{(k)} - \hat{x} = \frac{1}{2}(\hat{x} - x^{(k)})^2 \frac{f''(\xi^{(k)})}{f'(x^{(k)})}.$$

Because of the definition of $x^{(k+1)}$ according to Newton's method we further obtain

$$x^{(k+1)} - \hat{x} = \frac{1}{2}(\hat{x} - x^{(k)})^2 \frac{f''(\xi^{(k)})}{f'(x^{(k)})}$$

and hence the claim follows. ■

Remark. If f has an root $\hat{x}$ of multiplicity m, i.e.

$$f'(\hat{x}) = \ldots = f^{(m-1)}(\hat{x}) = 0, \quad f^{(m)} \neq 0,$$

then the conditions of Proposition 5.5 or 5.7 are not fulfilled and one can no longer expect local quadratic convergence. The Newton method then converges only locally linear - but the quadratic convergence can be recovered by the following modification of Newton's method:

$$x^{(k+1)} = x^{(k)} - m\frac{f(x^k)}{f'(x^{(k)})}. \qquad \text{(modified Newton Method)}$$

Details about this in the exercises. $\Diamond$

5.3. Newton's Method for Systems

In many applications one is not only dealing with a scalar equation. Often one has to deal with the case of n nonlinear equations for n unknowns. That is, one is looking for a zero of a vector-valued function

$$F\colon \operatorname{dom}(F) \subset \mathbb{R}^n \to \mathbb{R}^n$$

with component functions $F_j\colon \operatorname{dom}(F) \to \mathbb{R}, \quad (j = 1, \ldots, n)$. Here we assume that F is twice continuously differentiable. What is the iteration rule induced by Newton's method in this case? First we have to extend the one-dimensional geometric approach of the tangent and its zero to the general case.

The Taylor expansion up to the first order (neglecting the remainder) of the component F_j near $x^{(k)}$ gives for $x \in \mathbb{R}^n$

$$\begin{aligned} F_j(x) &\approx F_j(x^{(k)}) + \nabla F_j(x^{(k)})(x - x^{(k)}) \\ &= F_j(x^{(k)}) + \sum_{i=1}^{n} \partial_i F_j(x^{(k)})(x_i - x_i^{(k)}). \end{aligned}$$

With the usual notation

$$JF(x) = [\partial_1 F(x), \ldots, \partial_n F(x)] = \begin{pmatrix} \partial_1 F_1(x) & \cdots & \partial_n F_1(x) \\ \vdots & & \vdots \\ \partial_1 F_n(x) & \cdots & \partial_n F_n(x) \end{pmatrix}$$

for the Jacobian matrix, where

$$\partial_i F := \begin{pmatrix} \partial_i F_1 \\ \vdots \\ \partial_i F_n \end{pmatrix},$$

we obtain

$$F(x) \approx F(x^{(k)}) + JF(x^{(k)})(x - x^{(k)}) + \text{higher order terms}.$$

We again assume that this first-order Taylor polynomial is a good approximation for F in a neighbourhood of $x^{(k)}$ and take as the next iterate $x^{(k+1)}$ again its zero, i.e. the solution of the following equation:

$$0 = F(x^{(k)}) + JF(x^{(k)})(x^{(k+1)} - x^{(k)})$$

If $JF(x^{(k)})$ is nonsingular, we get from the last equation

$$x^{(k+1)} = x^{(k)} - \big(JF(x^{(k)})\big)^{-1} F\big(x^{(k)}\big).$$

This is obviously the vector-valued analogue of the scalar Newton iteration. In the numerical implementation of Newton's method for systems, the calculation of the inverse of $JF(x^{(k)})$ must be avoided, since this is more costly than the solution of the system of equations! The correction

$$s^{(k)} = -\big(JF(x^{(k)})\big)^{-1} F\big(x^{(k)}\big)$$

is exactly the solution of the system of linear equations

$$JF(x^{(k)}) s^{(k)} = -F(x^{(k)}).$$

Therefore, the algorithm for Newton's method for several variables reads as follows:

NEWTON METHOD IN HIGHER DIMENSIONS
Initialization: Given: Initial guess $x^{(0)}$ and a function $F\colon \operatorname{dom}(F) \subset \mathbb{R}^n \to \mathbb{R}^n$.
`for` $k = 0, 1, 2, \ldots$ `do`
`(i) compute` $f(x^{(k)}),\ JF(x^{(k)})$
`(ii) solve System of linear equations in` $s^{(k)}$`:` $$JF(x^{(k)}) s^{(k)} = -F(x^{(k)})$$
`(iii) define (Newton correction):` $$x^{(k+1)} = x^{(k)} + s^{(k)}$$
`until stop`

To solve the system of linear equations (ii) one can use, for example, the LU decomposition. The general Newton method can also be understood as a fixed point iteration

$$x^{(k+1)} = \Phi(x^{(k)})$$

with the corresponding iteration function

$$\Phi(x) = x - \big(JF(x)\big)^{-1} F(x).$$

As in the scalar case, one expects quadratic convergence under certain additional conditions, but these are relatively complicated and in practice hardly or mostly not verifiable. Let us first consider an example.

Example 5.8. The system to solve is:

$$F(x_1, x_2) = \begin{pmatrix} F_1(x_1, x_2) \\ F_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} 2x_1 x_2 - 4 \\ x_1^2 - x_2^2 - 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{5.3}$$

Determine the first two iterates of Newton's method at the initial value $x^{(0)} = (x_1^{(0)}, x_2^{(0)}) = (1, 1)$ and sketch them in a single sketch together with the two sets

$$M_1 = \{(x_1, x_2) \in \mathbb{R}^2 \mid F_1(x_1, x_2) = 0\} \quad \text{and} \quad M_2 = \{(x_1, x_2) \in \mathbb{R}^2 \mid F_2(x_1, x_2) = 0\}.$$

Solution: First we determine the Jacobian matrix

$$JF(x) = \begin{pmatrix} 2x_2 & 2x_1 \\ 2x_1 & -2x_2 \end{pmatrix}.$$

We have

$$JF(x^{(0)}) = \begin{pmatrix} 2 & 2 \\ 2 & -2 \end{pmatrix} \quad \text{and} \quad F(x^{(0)}) = \begin{pmatrix} -2 \\ -3 \end{pmatrix}.$$

To calculate $x^{(1)}$, we first have to solve the system of linear equations

$$JF(x^{(0)})s^{(0)} = -F(x^{(0)}).$$

In our case this reads

$$\left(\begin{array}{cc|c} 2 & 2 & 2 \\ 2 & -2 & 3 \end{array}\right)$$

with corresponding solution

$$s^{(0)} = \frac{1}{4}\begin{pmatrix} 5 \\ -1 \end{pmatrix}.$$

Thus we obtain

$$x^{(1)} = x^{(0)} + s^{(0)} = \frac{1}{4}\begin{pmatrix} 9 \\ 3 \end{pmatrix}.$$

For the calculation of the next iterate we need $JF(x^{(1)})$ and $F(x^{(1)})$. One obtains

$$JF(x^{(1)}) = \frac{1}{4}\begin{pmatrix} 6 & 18 \\ 18 & -6 \end{pmatrix} \quad \text{and} \quad F(x^{(1)}) = \frac{1}{8}\begin{pmatrix} -5 \\ 12 \end{pmatrix}.$$

We now want to solve

$$JF(x^{(1)})s^{(1)} = -F(x^{(1)}),$$

which brings us to the linear system

$$\left(\begin{array}{cc|c} 6 & 18 & 2.5 \\ 18 & -6 & -6 \end{array}\right).$$

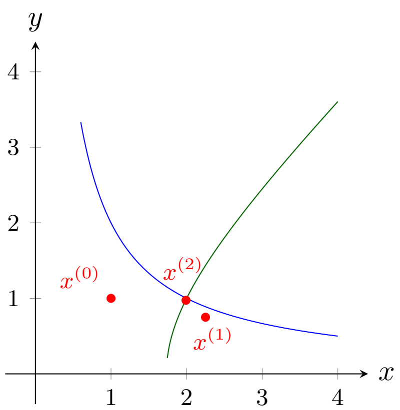
Its solution is

$$s^{(1)} = \frac{1}{120}\begin{pmatrix} -31 \\ 27 \end{pmatrix}$$

so that

$$x^{(2)} = x^{(1)} + s^{(1)} = \frac{1}{4}\begin{pmatrix} 9 \\ 3 \end{pmatrix} + \frac{1}{120}\begin{pmatrix} -31 \\ 27 \end{pmatrix} = \frac{1}{120}\begin{pmatrix} 239 \\ 117 \end{pmatrix}.$$

The sketch of the sets M_1 and M_2 in the first quadrant as well as the iterates now looks as follows:



//.

Remarks on the practical implementation of Newton's method:

(a) In real applications, the most complex part of the calculation is often the evaluation of the Jacobian matrix $JF(x^{(k)})$. Therefore, it is often not recalculated for each iteration, but left constant for 3-5 iterations and then recalculated again. This simplification is justified as long as F or JF does not change too much in a neighbourhood of the iterates. This generally reduces the speed of convergence somewhat, but increases the numerical efficiency.

(b) In many cases, the Jacobi matrix cannot be calculated in closed form, or only with great effort. Instead of the derivatives

$$\partial_j F_i(x^{(k)}),$$

one therefore uses numerical differentiation by replacing the derivatives with difference quotients

$$\partial_j F_i(x) \approx \frac{F_i(x + he_j) - F_i(x)}{h},$$

where e_j is the jth unit vector. The size of h depends on the specific task. In general, a too large h will impair the accuracy of the approximation of $JF(x^{(k)})$ and thus also the convergence of Newton's method. A too small h, on the other hand, carries the danger of cancellation.

(c) There is another variant of Newton's method that is often used, the so-called damped Newton method. In the conventional Newton method, the correction term $s^{(k)}$ provides a direction in which the function decreases. Often, however, it is better to make only a part of the step in this direction, i.e. one sets

$$x^{(k+1)} = x^{(k)} + \lambda s^{(k)}$$

for a suitable $0 < \lambda \leq 1$ (cf. sketch).

This procedure makes sense especially when the iterates are still far from a zero and/or the next iterate does not bring any improvement, i.e.

$$\|F(x^{(k+1)})\| > \|F(x^{(k)})\|.$$

To determine a suitable damping, one often introduces an improvement condition. To do this, one chooses a suitable vector space norm $\|\cdot\|$ and then chooses λ so that

$$\|F(x^{(k+1)})\| = \|F(x^{(k)} + \lambda s^{(k))}\| \overset{!}{<} \|F(x^{(k)})\|. \tag{5.4}$$

Under suitable conditions - which we will not go into here - a $0 < \lambda \leq 1$ and a norm can always be found, so that the descent condition (5.4) is fulfilled for every step of the Newton iteration.

Appendix

Appendix A.

List of Symbols

Appendix B.

Supplements to calculus and linear algebra.

Definition B.1. *A vector valued function $f\colon [a,b] \to \mathbb{R}^n$ with component functions $f_i\colon [a,b] \to \mathbb{R}$ is called* ***(Riemann) integrable*** *over $[a,b]$ if the component functions f_i are (Riemann) integrable on $[a,b]$ for $i = 1, \ldots, n$. For integrable f we define*

$$\int_a^b f(t)\,dt := \begin{pmatrix} \int_a^b f_1(t)\,dt \\ \vdots \\ \int_a^b f_n(t)\,dt \end{pmatrix}.$$

//.

Proposition B.2. *Let $f\colon [a,b] \to \mathbb{R}^n$ be continously differentiable. Then*

$$f(b) - f(a) = \int_a^b f'(t)\,dt.$$

Proof.

$$f(b) - f(a) = \begin{pmatrix} f_1(b) - f_1(a) \\ \vdots \\ f_n(b) - f_n(a) \end{pmatrix} = \begin{pmatrix} \int_a^b f_1'(t)\,dt \\ \vdots \\ \int_a^b f_n'(t)\,dt \end{pmatrix} = \int_a^b f'(t)\,dt.$$

∎

Appendix C.

Additional topics

C.1. Newton's method

As in the scalar case, one expects that Newton's method converges (locally) quadratically. However, in order to guarantee this convergence strictly mathematically, one has to make a number of assumptions, among others on the initial value $x^{(0)}$, which are unfortunately usually difficult or impossible to verify in practice. To illustrate the type of such assumptions and the convergence behavior of Newton's method, we formulate a simple variant of such a Proposition.

Proposition C.1. *Let $\Omega \subset \mathbb{R}^n$ be open and convex and $F \in C^1(\Omega, \mathbb{R}^n)$ such that the Jacobian matrix $JF(x)$ is invertible for all $x \in \Omega$. Let $\|\cdot\|_*$ be any matrix norm that is compatible with a vector norm $\|\cdot\|$. Let $\beta > 0$ be chosen such that*

$$\|(JF(x))^{-1}\| \leq c \quad \textit{for all } x \in \Omega.$$

Let also JF be Lipschitz continuous on Ω with constant L, i.e.

$$\|JF(x) - JF(y)\| \leq L\|x - y\|, \quad \forall x, y \in \Omega. \tag{C.1}$$

Furthermore, assume that there exists a solution $\hat{x}$ of $F(x) = 0$ in Ω. Moreover, let the initial value $x^{(0)}$ satisfy $x^{(0)} \in B_r(\hat{x}) := \{x \in \mathbb{R}^n \mid \|\hat{x} - x\| < r\}$, where r is supposed to be sufficiently small so that $B_r(\hat{x}) \subset \Omega$ and

$$r \leq \frac{2}{cL}$$

Then the Newton sequence $x^{(k)}$ remains within the ball $B_r(\hat{x})$ and converges quadratically to x:

$$\|x^{(k+1)} - \hat{x}\| \leq \frac{cL}{2}\|x^{(k)} - \hat{x}\|^2, \quad k = 0, 1, 2, \ldots.$$

Remark. Under these assumions one can moreover show that $\hat{x}$ is the only zero of F in $B_r(\hat{x})$. ◊

Example C.2. Let $f\colon \mathbb{R} \to \mathbb{R},\ f(x) = e^{-x} - x$ and $\Omega := (0, \infty)$. Clearly $f \in C^1(\Omega)$. Because

$$f'(x) = -e^{-x} - 1 < -1$$

for all $x \in \Omega$ we infer that $f'(x)$ is invertibel for all $x \in \Omega$. Furthermore, we have that

$$|(f'(x))^{-1}| = \left|\frac{1}{e^{-x} + 1}\right| \leq 1 =: c.$$

In addition, since

$$|f''(x)| = e^{-x} \leq 1$$

we conclude that f' is Lipschitz continuous with Lischitz constant $L = 1$ (by the mean value theorem). Noting that

$$f\Big(\frac{1}{2}\Big) = e^{-\frac{1}{2}} - \frac{1}{2} > 0 \quad \Leftrightarrow \quad 4 > e$$

and

$$f(1) = e^{-1} - 1 < 0$$

we see that there exists a zero $\hat{x} \in (\frac{1}{2}, 1) \subset \Omega$. Letting $r \coloneqq \frac{1}{2}$ we verify that

$$B_{\frac{1}{2}}(\hat{x}) \subset \Omega \quad \text{and} \quad r \leq \frac{2}{cL} = 2.$$

Thus we have shown that for any $x^{(0)} \in (\frac{1}{2}, 1)$ the Newton sequence converges quadratically to $\hat{x}$. //.