Numerical Methods and Optimization

Bachelor
Applied Artificial Intelligence
(AAI-B3)
DRAFT

Florian Link
October 11, 2022

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*

$$Av = \lambda v \qquad /\!/.$$

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 i-th 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

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

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

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

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

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) $\|x + y\| \leq \|x\| + \|y\|$

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 $x \neq 0$. Frequently used norms in $\mathbb{R}^n$ are the p-norms (also: l_p-norm)

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

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)

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

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

$$1 \leq \|x\|_p = \big(|x_1|^p + \ldots + |x_n|^p \big)^{\frac{1}{p}} \leq n^{\frac{1}{p}}$$

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).

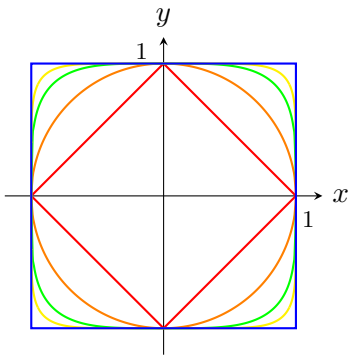

Fig. 2.1.: l_p unit spheres

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) $\|A + B\| \le \|A\| + \|B\|$

(d) $\|A \cdot B\| \le \|A\| \cdot \|B\|$. *//.*

As for vector norms it follows that $\|A\| \ge 0$, 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

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

Conversely, let i_0 be an index such that

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

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

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

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 $\rho(A^T A)$ denotes the largest absolute eigenvalue of $A^T A$. 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

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

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

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

and further

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

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 $(L_2 L_1)^{-1}$ from the left then results in a decompostition of A into a lower triangular matrix L and a upper triangular matrix U:

$$A = \underbrace{(L_2 L_1)^{-1}}_{=:L} U.$$

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 k-th step, the k-th 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 k-th 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 k-th 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

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

i.e. the inverse matrix of L_k is obtained simply by changing the sign in the k-th 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}:

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

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

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

■

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:

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

The elimination steps can be combined into a short sheme:

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

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 $A = LR$ 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: