Computational Mathematics 1

Lecture notes, University (of Technology) Graz
based on the lecture by Tobias Breiten

Lukas Prokop

November 2, 2018

Contents

Course

↓ *This lecture took place on 2018/10/02.*

- Homepage at uni-graz.at
- Contents:
  1. Linear equation system
  2. Numerical error analysis
  3. Curve fitting (dt. "Lineare Ausgleichsrechnung")
  4. Non-linear equations
  5. Interpolation
  6. Numerical integration
- Literature:
  1. Deuflhard, Hohmann: Numerische Mathematik 1
  2. Schwarz, Köckler: Numerische Mathematik

1 What is Computational Mathematics?

Computational Mathematics as branch of mathematics considers the construction and analysis of algorithms for continuous mathematical problems. More specifically, based on a model and input data, we construct specific output data of interest. Formally, we associate a model using function $f : X \to Y$ where X is the set of input data and Y the set of output data. The problem is to find output $f(x) \in Y$ for given $x \in X$.

Examples:

- Addition of two numbers: $X = \mathbb{R}^2, Y = \mathbb{R}, f : (x_1, x_2) \mapsto x_1 + x_2$
- Solution of a linear equation system: $Ax = b$

$$X = \mathbb{R}^{n \times n} \times \mathbb{R}^n, Y = \mathbb{R}^n, f : (A, b) \mapsto A^{-1}b$$

An algorithm is defined as unambiguous sequence of steps to solve a given problem. A problem is characterized as

- the input data required for computation
- the output data that represent the solution of the algorithm
- the description of the steps to be done, including auxiliary values

We require, the algorithm is

executable by a machine

terminating after a finite number of steps

deterministic as the same input always leads to the same output

Furthermore we analyze the algorithm in terms of

stability small errors in the input lead to small errors in the output

precision output data should solve the problem with maximum accuracy

efficiency large problems can be solved in practical time

Consider the linear equation system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b \\ \vdots &= \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b \end{aligned}$$

The same can be specified in compact notation like $Ax = b$ with $A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$.

Theorem 1.0.1 (Cramer's Rule). *Let $A \in \mathbb{R}^{n\times n}$ with $\det(A) \neq 0$ and $b \in \mathbb{R}^n$. Then there exists exactly one $x \in \mathbb{R}^n$ such that $Ax = b$. We can determine x using "Cramer's Rule":*

$$x_i = \frac{\det(A_i)}{\det(A)} \text{ where } A_i = \begin{bmatrix} a_{11} & \dots & a_{1,i-1} & b_1 & a_{1,i+1} & \dots & a_{1n} \\ \vdots & & & & & & \vdots \\ a_{n1} & \dots & a_{n,i-1} & b_n & a_{n,i+1} & \dots & a_{nn} \end{bmatrix}$$

Remark (Problem). *Computation using $\det(A) = \sum_{\sigma\in S_n} \operatorname{sign}(\sigma) a_{1,\sigma(1)} \dots a_{n,\sigma(n)}$ requires $n \cdot n!$ operations.*

1.1 Solving linear eq. system by forward/backwards substitution

Consider the special case of a linear equation system of structure

$$\begin{array}{lllll} \gamma_{11}x_1 & +\gamma_{12}x_2 & +\dots & +\gamma_{1n}x_n & = z_1 \\ & \gamma_{22}x_2 & +\dots & +\gamma_{2n}x_n & = z_2 \\ & & \ddots & & =\vdots \\ & & & \gamma_{nn}x_n & = z_n \end{array}$$

and accordingly, $Rx = z$ with an upper triangular matrix, hence $\gamma_{ij} = 0$ for $i > j$. We retrieve x by recursive solution (by so-called "backwards substitution") starting with row n:

$$\begin{aligned} x_n &= \frac{z_n}{\gamma_{nn}} \text{ if } \gamma_{nn} \neq 0 \\ x_{n-1} &= \frac{z_{n-1} - \gamma_{n-1,n}x_n}{\gamma_{n-1,n-1}} && \text{if } \gamma_{n-1,n-1} \neq 0 \\ x_1 &= \frac{z_1 - \gamma_{12}x_2 - \dots - \gamma_{1n}x_n}{\gamma_{11}} && \text{if } \gamma_{11} \neq 0 \end{aligned}$$

Obviously, it holds that

$$\det(R) = \gamma_{11} \dots \gamma_{nn} \neq 0 \iff \gamma_{ii} \neq 0 \qquad \forall i = 1, \dots, n$$

The algorithm is applicable (just like Cramer's rule), if $\det(R) \neq 0$. Then the existence of a solution is guaranteed and this solution is unique.

Remark (Computing time).

1. *for the i-th row, we need $(n-i)$ additions, $(n-i)$ multiplications and one division*
2. *in total for the rows n to 1,*

$$\sum_{i=1}^{n}(i-1) = \frac{n(n-1)}{2} \doteq \frac{n^2}{2}$$

Multiplication and the same amount of additions. The notation $\doteq$ is used to denote equivalence up to terms of same degree. Analogously, the equation system $Lx = z$ can be solved with a lower triangular matrix L by forward substitution starting with the first row.

1.2 Gaussian elimination process

Basic idea:

Beginning with a linear equation system of structure,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots &= \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{1}$$

we apply equivalence transformations to end up with an upper triangular matrix.

As a first step, we eliminate variable x_1 of rows 2 to n.

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a'_{22}x_2 + \cdots + a'_{2n}x_n &= b_1 \\ \vdots &= \vdots \\ a'_{12}x_2 + \cdots + a'_{nn}x_n &= b'_n \end{aligned} \tag{2}$$

So we apply this elimination step ($r_i := \frac{r_i}{a_{i1}} - \frac{r_1}{a_{11}}$ where r_i denotes the i-th row) to the last $n-1$ rows and determine the triangular matrix recursively this way.

Consider the transformation of (1) to (2). Let $a_{11} \neq 0$ and we recognize $\text{row}_i := \text{row}_i - l_{i_1} \cdot \text{row}_1$. Formally,

$$\underbrace{(a_{i1} - l_{i1}a_{11})x_1}_{=0} + \underbrace{(a_{i2} - l_{i1}a_{12})}_{a'_{i2}} x_2 + \cdots + \underbrace{(a_{in} - l_{i1}a_{1n})}_{a'_{in}} x_n = \underbrace{b_i - l_{in}b_1}_{b'_i}$$

$$\implies l_{in} = \frac{a_{i1}}{a_{11}}$$

hence, the first elimination step is applicable if $a_{11} \neq 0$ is given. If we apply this step, we retrieve a sequence of matrices:

$$A = A^{(1)} \to A^{(2)} \to \ldots \to A^{(n)} = R$$

where

$$A^{(k)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & & & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & & & a_{2n}^{(2)} \\ & & \ddots & & & \vdots \\ & & & a_{kk}^{(k)} & \ldots & a_{kn}^{(k)} \\ & & & \vdots & & \vdots \\ & & & a_{nk}^{(k)} & \ldots & a_{nn}^{(k)} \end{bmatrix}$$

with a remaining matrix of dimension $(n-k+1) \times (n-k+1)$. For every remaining matrix we can apply the elimination step

$$\begin{aligned} l_{ik} &= \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} && \text{for } i = k+1, \ldots, n \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - l_{ik}a_{kj}^{(k)} && \text{for } i, j = k+1, \ldots, n \\ b_i^{(k+1)} &= b_i^{(k)} - l_{ik}b_k^{(k)} && \text{for } i = k+1, \ldots, n \end{aligned}$$

under the assumption that $a_{kk}^{(k)} \neq 0$ (pivot element).

Remark. *Every elimination step is a linear operation of the rows of A. Hence it is representable by left multiplication with $L_k \in \mathbb{R}^{n\times n}$ according to $A^{k+1} = L_k A^{(k)}$ and $b^{(k+1)} = L_k b^{(k)}$.*

The following matrix is a Frobenius matrix, *which has an interesting property regarding its inverse:*

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

and

$$L = L_1^{-1} \quad \dots \quad L_{n-1}^{-1} = \begin{pmatrix} 1 & & & & 0 \\ l_{21} & \ddots & & & \\ l_{31} & l_{32} & \ddots & & \\ \vdots & \vdots & & & \\ l_{n1} & \dots & & l_{n,n-1} & 1 \end{pmatrix}$$

$$L^{-1} = L_{n-1} \dots L_1$$

This way, we retrieve an equivalent linear equation system to $Rx = z$, $R = L^{-1}A$, $z = L^{-1}b$. The representation $A = LR$ is called *LR decomposition of A*. If they exists, L and R are unambiguously determined.

$$A = L_1 R_1 = L_2 R_2 \implies R_1 = L_1^{-1} L_2 R_2 \implies R_1 R_2^{-1} = L_1^{-1} L_2$$

$$\implies \text{diagonal representation of } L_1^{-1} L_2 \dots, R_1 R_2^{-1}$$

Algorithm 1 (Gaussian elimination)**.**

1. *$A = LR$ with R as upper triangular matrix, L as lower triangular matrix*
2. *Solve $Lz = b$ by forward substitution*
3. *Solve $Rx = z$ by backwards substitution*

The entries $l_{ik} \neq 1$ can be stored in the remaining null-entries of the matrix $A^{(k)}$. Thus the total memory requirements are bounded by $n(n+1)$.

Remark (Computing time)**.**

- $\sum_{k=1}^{n-1} k^2 \doteq \frac{n^3}{3}$ *for step 1*
- $\sum_{k=1}^{n-1} k \doteq \frac{n^2}{2}$ *for step 2 and 3*

1.3 Pivot strategies

Problem: The algorithm above does not work for the simple example:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \det(A) = -1 \neq 0$$

We *need* row exchanges. We can evaluate, that no LU-decomposition exists.

But exchange of the rows gives

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \rightsquigarrow \hat{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I = LU$$

with $L = U = I$.

Example 1.1. *Consider the equation system*

$$\begin{aligned} 1 \cdot 10^{-4} x_1 + 1 \cdot x_2 &= 1 \\ 1 \cdot x_1 + 1 \cdot x_2 &= 2 \end{aligned}$$

Assumption: *Compute up to three decimal points. Then for the "accurate" solution, we get* $x_1 = 1$ *and* $x_2 = 1$.

Using the Gaussian elimination process, we get

$$l_{21} = \frac{a_{21}}{a_{11}} = 10^4$$

$$\implies 0 \cdot x_1 + (1 - 10^4 \cdot 1) x_2 = 2 \cdot 10^4 \cdot 1$$

Therefore we get the triangular system

$$10^{-4} x_1 + 1 \cdot x_2 = 1$$

$$-10^4 x_2 = -10^4$$

By this, we get an approximation $x_2 = 1$ and $x_1 = 0$.

In contrast, if we exchange the two rows,

$$\begin{aligned} 1 \cdot x_1 + 1 \cdot x_2 &= 2 \\ 10^{-4} x_1 + 1 \cdot x_2 &= 1 \end{aligned}$$

then $\tilde{l}_2 1 = \frac{10^4}{1}$ follows and therefore the triangular system

$$\begin{aligned} 1 \cdot x_1 + 1 \cdot x_2 &= 2 \\ 1 \cdot x_2 &= 1 \end{aligned}$$

with the "correct" solution is $x_2 = 1$ and $x_1 = 1$.

Our exchange of the rows has given us $\left|\tilde{l}_{21}\right| < 1$ and $|\tilde{a}_{11}| \geq |\tilde{a}_{21}|$.

↓ *This lecture took place on 2018/10/04.*

The new pivot element $\tilde{a}_{11}$ is the largest absolute value in the first column. Thus, we apply the column pivot strategy: In every step, choose the row with the largest, absolute value in the pivot column.

Algorithm 2 (Gaussian elimination with column pivot strategy)**.**

1. *In step $A^{(k)} \to A^{(k+1)}$, choose some $p \in \{k, \dots, n\}$ such that $\left|a_{p_k}^{(k)}\right| \geq \left|a_{jk}^{(k)}\right|$ for $j = k, \dots, n$*
2. *Exchange rows p and k*

$$A^{(k)} \rightsquigarrow \tilde{A}^{(k)} \text{ with } \tilde{a}_{ij}^{(k)} = \begin{cases} a_{kj}^{(k)} & i = p \\ a_{pj}^{(k)} & i = k \\ a_{ij}^{(k)} & \text{else} \end{cases}$$

3. *Apply the elimination step $\tilde{A}^{(k)} \to \tilde{A}^{(k+1)}$*

Remark 1.2. *We can apply the row pivot strategy with column exchange instead of the column pivot strategy with row exchange.*

We use permutation matrices $P \in \mathbb{R}^{n \times n}$ to analyze the column pivot search. By permutation $\pi \in S_n$, we associate the *permutation matrix*

$$P_\pi = [e_{\pi(1)}, e_{\pi(2)}, \dots, e_{\pi(n)}]$$

where e_j denotes the j-th unit vector in $\mathbb{R}^n$. Row and column exchanges of a matrix A can followingly be represented as

$$\pi_z : A \mapsto P_\pi A, \qquad \pi_S : A \mapsto A P_\pi$$

Remark. *It holds that* $\det(P_\pi) = \operatorname{sign}(\pi) \in \{\pm 1\}$ *and* $P_\pi^{-1} = P_\pi^T$. *[Consider* $P_\pi = e_{\pi(1)} \cdot e_1^T + e_{\pi(2)} e_2^T + \dots + e_{\pi(n)} e_n^T$*.]*

$$P_\pi^T = e_1 e_{\pi(1)}^T + e_2 e_{\pi(2)}^T + \dots + e_n e_{\pi(n)}^T$$
$$P_\pi P_\pi^T = e_{\pi(1)} e_{\pi(1)}^T + e_{\pi(2)} e_{\pi(2)}^T + \dots + e_{\pi(n)} e_{\pi(n)}^T$$

Theorem 1.2.1. *Let $A \in \mathbb{R}^{n \times n}$ and* $\det(A) \neq 0$. *Then there exists some permutation matrix $P \in \mathbb{R}^{n \times n}$ such that*

$$PA = LU$$

with lower and upper triangular matrices $L, U \in \mathbb{R}^{n \times n}$. Furthermore P can be chosen such that $|L| := \max_{i,j} |l_{ij}| \leq 1$.

Proof. We use Algorithm 2. Because $\det(A) \neq 0$, $\max_i |a_{i1}| > 0$ and therefore there exists some transposition $\tau_1 \in S_n$ such that for $A^{(1)} = P_{\tau_1} A$,

$$\max_i \left|a_{i1}^{(i)}\right| = \left|a_{11}^{(1)}\right| > 0$$

We can apply the elimination step and retrieve a matrix of structure

$$A^{(2)} = L_1 A^{(1)} = L_1 P_{\tau_1} A = \left[\begin{array}{ccccc} a_{11}^{(1)} & x & x & \dots & x \\ \hline 0 & & & & \\ \vdots & & & B^{(2)} & \\ 0 & & & & \end{array}\right]$$

Especially, it holds that $\max_{i,j} \left| l_{i,j}^{(1)} \right| \leq 1$ and $\det(L_1) = 1$. And

$$0 \neq \operatorname{sign}(\tau_1) \cdot \det(A) = \det(A^{(2)}) = a_{11}^{(1)} \det(B^{(2)})$$

and thus $\det(B^{(2)}) \neq 0$. Inductively, we get $U = A^{(n)} = L_{n-1} P_{\tau_{n-1}} \dots L_1 P_{\tau_1} A$ with $|L_k| \leq 1$ and transposition of two numbers $\geq k$. If $\pi \in S_n$ exchanges only numbers $\geq k+1$, then

$$L_k = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & -l_{k+1,k} & 1 & & \\ 0 & \vdots & & \ddots & \\ & l_{n,k} & & 0 & 1 \end{bmatrix}, \hat{L}_k = P_\pi L_k P_\pi^{-1} = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & -l_{\pi(k+1),k} & 1 & & \\ 0 & \vdots & & \ddots & \\ & l_{\pi(n),k} & & 0 & 1 \end{bmatrix}$$

By insertion of the identities $P_{\tau_k}^{-1} P_{\tau_k}$ we get,

$$\begin{aligned} &= L_{n-1} P_{\tau_{n-1}} L_{n-2} P_{\tau_{n-1}}^{-1} P_{\tau_{n-1}} P_{\tau_{n-2}} L_{n-3} P_{\tau_{n-3}} \dots L_1 P_{\tau_1} A \\ &= \underbrace{(L_{n-1})}_{\hat{L}_{n-1}} (P_{\tau_{n-1}} L_{n-2} P_{\tau_{n-1}}^{-1})(P_{\tau_{n-1}} P_{\tau_{n-2}} L_{n-3} P_{\tau_{n-2}}^{-1} P_{\pi_{n-1}}^{-1} P_{\tau_{n-1}} P_{\tau_{n-2}} P_{\tau_{n-3}} \dots L_1 P_{\tau_n} A) \\ &= \hat{L}_{n-1} \hat{L}_{n-2} \hat{L}_{n-3} \dots \hat{L}_1 P_{\pi_0} A \end{aligned}$$

with $\hat{L}_k = P_{\pi_k} L_k P_{\pi_k}^{-1}$, $\pi_{n-1} = \mathrm{id}$, $\pi_k = \tau_{n-1} \dots \tau_{k+1}$, $k = 0, \dots, n-2$. Be aware that π_k only exchanges numbers $\geq k+1$ and the matrices $\hat{L}_k$ are therefore of structure previously mentioned. As a consequence, we created the decomposition $P_{\pi_0} A = LU$ with

$$L := \hat{L}_1^{-1} \dots \hat{L}_{n-1}^{-1}, \quad L = \begin{bmatrix} 1 & & & & \\ l_{\pi_1(2),1} & 1 & & & \\ l_{\pi_2(3),1} & l_{\pi_2(3)} & \ddots & & \\ \vdots & \vdots & & \ddots & \\ l_{\pi_1(n),1} & \dots & & l_{\pi_{n-1}(n),n-1} & 1 \end{bmatrix}$$

Especially, $|L| = 1$. □

Remark 1.3. *Recognize that the method mentioned in the proof can be used to compute the determinant of A.*

$$\det(A) = \det(P) \cdot \det(LU) = \operatorname{sign}(\pi_0) \cdot \gamma_{11} \dots \gamma_{nn}$$

1.4 Cholesky process for symmetric positive definite matrices

↓ *This lecture took place on 2018/10/09.*

In the following, let $A = A^T > 0$ be symmetric positive definite.

Remark. *By symmetric positive definite property,*

1. $\langle x, Ax \rangle > 0 \forall x \neq 0$
2. *there exists an orthonormal basis* $q_1, \dots, q_n$ *of eigenvectors such that* $Q^T Q = I$

Theorem 1.3.1. *For every symmetric positive definite matrix* $A \in \mathbb{R}^{n \times n}$ *it holds that*

1. A *is invertible*
2. $a_{ii} > 0$ *for* $i = 1, \dots, n$
3. $\max_{i,j=1,\dots,m} |a_{ij}| = \max_{i=1,\dots,n} a_{ii}$
4. *Gaussian elimination without pivot search gives a symmetric positive definite remainder matrix in every step*

Proof.

1. follows by $\langle x, Ax \rangle > 0 \forall x \neq 0$. Assume A is not invertible, then $\exists x \in \text{kernel}(A) \implies \langle x, Ax \rangle = \langle x, 0 \rangle = 0$ which contradicts with A being symmetric positive definite.
2. also follows by $\langle x, Ax \rangle > 0$ with choice $x = e_i$ for $i = 1, \dots, n$ because $\langle e_i, Ae_i \rangle = a_{ii}$
3. Left as an exercise
4. We denote $A = A^{(1)}$ according to

$$A^{(1)} = \begin{bmatrix} a_{11} & z^T \\ z & B^{(1)} \end{bmatrix}$$

with $z = [a_{21}, \dots, a_{11}]^T$. After the first eliminiation step,

$$A^{(2)} = L_1 A^{(1)} = \begin{bmatrix} a_{11}^{(1)} & z^T \\ 0 & \\ \vdots & B^{(2)} \\ 0 & \end{bmatrix}$$

$$L_1 = \begin{bmatrix} 1 & & & \\ -l_{21} & \ddots & & \\ \vdots & 0 & \ddots & \\ -l_{n1} & & & 1 \end{bmatrix}$$

Right-side multiplication with L_1^T gives

$$L_1 A^{(1)} L_1^T = \begin{bmatrix} a_{11}^{(1)} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & B^{(2)} & \\ 0 & & & \end{bmatrix}$$

Furthermore with $A^{(1)} = A > 0$ it also holds that $L_1 A^{(1)} L_1^T > 0$. By this, it is especially true that $B^{(2)} > 0$.

□

Theorem 1.3.2. *For every symmetric positive definite matrix there exists a unique decomposition $A = LDL^T$ where L is an upper triangular matrix (with $l_{ii} = 1$) and D is a positive diagonal matrix.*

Proof. This follows by the construction made in proof 1.4 for $k = 2, \dots, n-1$. Thus we get $L = L_1^{-1} L_2^{-1} \dots L_{n-1}^{-1}$ and D as diagonal matrix of the pivot elements. □

Because all diagonal elements d_{ii} are positive definite, there exists

$$D^{\frac{1}{2}} := \text{diag}\left(\sqrt{d_{11}}, \dots, \sqrt{d_{nn}}\right)$$

Corollary 1.4. *There exists some $\overline{L} := LD^{\frac{1}{2}}$ (lower triangular matrix) such that $A = \overline{L}\overline{L}^T$.*

Algorithm 3 (Cholesky decomposition)**.**

- *For* $k = 1, \dots, n$
  - $l_{kk} := (a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2)^{\frac{1}{2}}$
  - *For* $i = k+1, \dots, n$
    - $l_{ik} := \frac{a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj}}{l_{kk}}$

2 Errors and matrix conditioning

So far, we use input data (A, b) to retrieve the result $A^{-1}b$.

Consider an abstract problem characterized by (f, x) with given map f and input data x. Thus in theory, we have the input, apply the algorithm and retrieve the result. But in practice, we have input with some error, an error in the algorithm and an error in the result.

In the following, we investigate input errors, which we cannot avoid in the general case. In the best case we can change the problem task. This refers to *conditioning of the problem*.

The algorithm triggers errors, which we can reduce or avoid by adapting the algorithm. This refers to *stability of the algorithm*.

2.1 Number representation and rounding errors

Definition 2.1. *Let x be a real number. $\tilde{x} \in \mathbb{R}$ is an approximating value for x. Then $\tilde{x} - x$ is called* absolute error *of $\tilde{x}$ and, assuming $x \neq 0$, $\frac{x-\tilde{x}}{x}$ is the* relative error *of x.*

Even if we know the input know exactly, the representation of continuous numbers yields rounding errors.

A number $x \in M$ is represented as $x = \text{sign}(x) \cdot a \cdot E^{e-k}$. The number system M is defined by the following four integer parameters:

basis $E \in \mathbb{N}$ with $E > 1$; mostly $E = 2$

precision $k \in \mathbb{N}$

exponent e in domain $e_{\text{min}} \leq e \leq e_{\text{max}}$ where $e_{\text{min}}, e_{\text{max}} \in \mathbb{Z}$

mantissa $a \in \mathbb{N}_0$ is defined as

$$0 \leq a = a_1 E^{k-1} + a_2 \cdot E^{k-2} + \cdots + a_{k-1} E^1 + a_k E^0 \leq E^k - 1$$

The mantissa length k and a_i are numbers of the number system, so 0 or 1 in case $E = 2$. If $x \neq 0$, then we require, that the first number is non-zero, then

$$E^{k-1} \leq q < E^k \qquad \text{if } x \neq 0$$

We call x *k-digit normalized floating point number with basis E.*

Calculating with these numbers is called *calculation with k significant digits.* This way, we get the domain of normalized floating points $x \neq 0$:

$$E^{e_{\text{min}}-1} \leq |x| \leq E^{e_{\text{max}}}(1 - E^{-k})$$

Example 2.2. *Let $k = 3$, $E = 2$, $e_{\text{min}} = -1$ and $e_{\text{max}} = 3$. Let $x > 0$. Then the smallest number is 0.25 and the largest number is 7. Now look at the distribution of numbers. There are 3 numbers between 1 and 2.*

Remark. *The numbers are not equidistantly distributed.*

After every arithmetic operation, the result x is rounded to a uniquely defined value $\text{rd}(x)$. We define the *machine precision* $\varepsilon := \frac{E}{2} E^{-k}$.

Lemma 2.3. *If $x \neq 0$ is within the domain of normalized floating point values and $\text{rd}(x) \in M$, then*

$$|\text{rd}(x) - x| \leq \frac{E^{e-k}}{2}$$

is the maximum absolute error *with respect to rounding.*

$$\frac{|\text{rd}(x) - x|}{|x|} \leq \frac{1}{2} E^{1-k} = \varepsilon$$

is the maximum relative error *with respect to rounding.*

Proof. Without loss of generality, let $x > 0$. Then we can denote,

$$x = \mu E^{e-k} \qquad E^{k-1} \leq \mu \leq E^k - 1$$

So, x lies in between the neighboring floating point numbers $x_1 = \lfloor \mu \rfloor E^{e-k}$ and $x_2 = \lceil \mu \rceil E^{e-k}$. By this, we either get $\text{rd}(x) = x_1$ or $\text{rd}(x) = x_2$.

Rounding gives

$$|\mathrm{rd}(x) - x| \leq \frac{x_2 - x_1}{2} \leq \frac{E^{e-k}}{2}$$

Thus, it follows that

$$\frac{|\mathrm{rd}(x) - x|}{|x|} \leq \frac{\frac{E^{e-k}}{2}}{\mu E^{e-k}} \leq \frac{1}{2} E^{1-k} = \varepsilon$$

□

precision	k	e_{min}	e_{max}	ε
single	24	-125	128	$2^{-24} \approx 6 \cdot 10^{-8}$
double	53	-1021	1024	$2^{-53} \approx 1 \cdot 10^{-16}$
extended	64	-16381	16384	$2^{-64} \approx 5 \cdot 10^{-20}$

Table 1: Examples for common number systems. Assuming basis $E = 2$

Lemma 2.4.

1. *It holds that* $\mathrm{rd}(x) = x(1+\delta)$ *with* $|\delta| \leq \varepsilon$
2. *Let* $\circ$ *be one of the elementary operations* $\{+, -, \cdot, /\}$. *Then* $\mathrm{rd}(x \circ y) = (x \circ y)(1+\delta)$ *with* $|\delta| \leq \varepsilon$.
3. *The machine precision* ε *is the smallest positive number* g *for which* $\mathrm{rd}(1+g) > 1$ *is true.*

Definition 2.5.

1. *The significant digits of a number are mantissa with normalized floating point representation.*
2. *When calculating with* k *significant digits, all input values must be rounded to* k *significant digits. Followingly the results of every elementary operation will be rounded to* k *significant digits before continuing. Thus the result is also affected.*
3. *Elimination is the cancellation of leading mantissa in the subtraction of two numbers of same sign.*

Example 2.6. *Depending on the chosen expressions, we can get different results. One example is:*

$$99 - 70\sqrt{2} = \sqrt{9801} - \sqrt{9800} = \frac{1}{\sqrt{9801} + \sqrt{9800}} \approx 0.05050633883346584\ldots$$

2.2 Condition of a problem

Question: What effects do deviations in input data have independent of the chosen algorithm?

	$99 - 70\sqrt{2}$	$\sqrt{9801} - \sqrt{9800}$	$\frac{1}{\sqrt{9801}-\sqrt{9800}}$
$k = 2$	1	0	0.0050
$k = 4$	0.02000	0.02000	0.005051
$k = 10$	0.005050660000	0.005050660000	0.005050633884

Table 2: Difference in result depending on the expression. Apparently elimination occurs for subtraction $99 - 70\sqrt{2}$ yielding ill-conditioned results

We already observed, that no difference occurs between input x and all inputs $\tilde{x}$ with an absolute error smaller than machine precision. Therefore, consider the input set E, which contains all disrupted inputs $\tilde{x}$

$$E = \{\tilde{x} \in \mathbb{R} \mid |\tilde{x} - x| < \varepsilon\varepsilon \cdot |x|\}$$

The map f (description of the problem) maps an input set E in a resulting set $U = f(E)$.

Example (Intersection of two lines)**.** *If we compute the intersection of lines with angles close to $0°$, the corresponding LES is ill-conditioned. Computing intersection for almost orthogonal lines is well-conditioned.*

↓ *This lecture took place on 2018/10/11.*

We consider the problem (f, X) given by the map $f : U \subset \mathbb{R}^n \to \mathbb{R}^m$ with $U \subset \mathbb{R}^n$ open. Let $x \in U$ and δ as (relative or absolute) precision of the input data be given.

Distinction between:

- $\|\tilde{x} - x\| \leq \delta$ (absolute) or $\|\tilde{x} - x\| \leq \delta \|x\|$ (relative) for some norm $\|\cdot\|$ on $\mathbb{R}^n$
- $|\tilde{x}_i - x_i| \subseteq \delta$ (absolute) or $|\tilde{x}_i - x_i| \leq \delta |x_i|$ (relative), hence componentwise for $i = 1, \dots, n$.

2.3 Normwise condition analysis

Assumption: The input error δ is sufficiently small, so we use *linearized error theory* for the asymptotic behavior of $\delta \to 0$.

Notation: Two functions $g, h : \mathbb{R}^n \to \mathbb{R}^m$ are equal in the *first approximation* or *leading approximation* for $x \to x_0$.

$$g(x) \doteq h(x) \text{ for } x \to x_0$$

if $g(x) = h(x) + \mathcal{O}(\|h(x)\|)$ for $x \to x_0$. The Landau symbol $o(\|h(x)\|)$ for $x \to x_0$ denotes a function φ such that

$$\lim_{x \to x_0} \frac{\|\varphi(x)\|}{\|h(x)\|} = 0$$

If f is differentiable in x, we have $f(\tilde{x}) - f(x) \doteq f'(x)(\tilde{x} - x)$ for $\tilde{x} \to x$. Analogously, "$g(x) \dot{\leq} h(x)$ for $x \to x_0$" (componentwise).

Definition 2.7. *The* absolute normwise condition *of problem (f, X) is the smallest number $\kappa_{\text{abs}} \geq 0$, such that*

$$\|f(\tilde{x}) - f(x)\| \dot{\leq} \kappa \|\tilde{x} - x\| \text{ for } \tilde{x} \to x$$

The problem (f, X) is ill-posed, *if there is no such number (formally $\kappa_{\text{abs}} = \infty$).*

The relative normwise condition of (f, X) is the smallest number $\kappa_{\text{rel}} \geq 0$ such that

$$\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \dot{\leq} \kappa_{\text{rel}} \frac{\|\tilde{x} - x\|}{\|x\|} \text{ for } \tilde{x} \to x$$

Remark. *κ_{abs} give the amplification of the absolute error, κ_{rel} is the relative error.*

If f is differentiable in x, then

$$\kappa_{\text{abs}} = \|f'(x)\|, \kappa_{\text{rel}} = \frac{\|x\|}{\|f(x)\|} \|f'(x)\|$$

where $\|f'(x)\|$ is the number of the Jacobian matrix $f'(x) \in \mathbb{R}^{m \times n}$ in regards of the norm

$$\|A\| := \sup_{x \neq 0} \frac{\|ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\| \text{ for } A \in \mathbb{R}^{m \times n}$$

Example 2.8.

$$\textit{Addition:} \quad f : \mathbb{R}^2 \to \mathbb{R} \qquad \begin{pmatrix} a \\ b \end{pmatrix} \mapsto f(a, b) = a + b$$

with the derivative $f'(a, b) = (1, 1) \in \mathbb{R}^{1 \times 2}$. For the 1-norm on $\mathbb{R}^2$ it holds that

$$\left\| \begin{pmatrix} a \\ b \end{pmatrix} \right\| = |a| + |b|$$

The norm (associated operator norm) is

$$\|f'(a, b)\| = \sup_{\|x\|_1 = 1} \left| \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \right| = 1$$

We retrieve the condition numbers of addition as

$$\kappa_{\text{abs}} = 1 \text{ and } \kappa_{\text{rel}} = \frac{|a| + |b|}{|a + b|}$$

Consequence:

- *$\kappa_{\text{rel}} = 1$ for addition of two numbers with same sign*
- *subtraction of two almost equal numbers given $|a + b| \ll |a| + |b| \implies \kappa_{\text{rel}} \gg 1$.*

Example 2.9 (Condition of a linear equation system $Ax = b$)**.** *If we only consider $b \in \mathbb{R}^n$ as input data, the problem is specified by the* linear map $f : \mathbb{R}^n \to \mathbb{R}^n$ *and $b \mapsto f(b) = A^{-1}b$. The derivative is $f'(b) = A^{-1}$ and we retrieve the condition numbers*

$$\kappa_{\text{abs}} = \left\|A^{-1}\right\| \text{ and } \kappa_{\text{rel}} = \frac{\|b\|}{\|A^{-1}b\|}\left\|A^{-1}\right\| = \frac{\|Ax\|}{\|x\|}\left\|A^{-1}\right\|$$

What about the deviations in A? For this, consider A as input data

$$f : \mathbb{R}^{n\times n} \supset \mathrm{GL}(n) \to \mathbb{R}^n, A \mapsto f(A) = A^{-1}b$$

for some fixed $b \in \mathbb{R}^n$. The map f is non-linear, but differentiable.

Lemma 2.10. *The map $g : \mathbb{R}^{n\times n} \supset \mathrm{GL}(n) \to \mathrm{GL}(n)$ with $g(A) = A^{-1}$ is differentiable and $g'(A) \cdot C = -A^{-1}CA^{-1}$ for all $C \in \mathbb{R}^{n\times n}$.*

Proof. Consider $I = (A + tC)(A + tC)^{-1}$ for sufficiently small t and derive by t:

$$0 = C(A + tC)^{-1} + (A + tC)\frac{d}{dt}(A + tC)^{-1}$$

For $t = 0$, it especially follows that $g'(A)C = \frac{d}{dt}(A + tC)^{-1}\big|_{t=0} = -A^{-1}CA^{-1}$. □

By Lemma 2.10, by the derivative of the solution $f(A) = A^{-1}b$ to A,

$$f'(A)C = -A^{-1}CA^{-1}b = -A^{-1}Cx \text{ for } C \in \mathbb{R}^{n\times n}$$

We retrieve the condition numbers

$$\kappa_{\text{abs}} = \|f'(A)\| = \sup_{\|C\|=1}\left\|A^{-1}Cx\right\| \leq \left\|A^{-1}\right\|\|x\|$$

$$\kappa_{\text{rel}} = \frac{\|A\|}{\|x\|}\|f'(A)\| \leq \|A\|\left\|A^{-1}\right\|$$

By the sub-multiplicativity, $\|Ax\| \leq \|A\|\|x\|$, of the matrix norm, it also holds for the relative condition in regards of the input b, that $\kappa_{\text{rel}} \leq \|A\| \cdot \left\|A^{-1}\right\|$. The value $\kappa(A) := \|A\|\left\|A^{-1}\right\|$ is called *condition number* of matrix A.

↓ *This lecture took place on 2018/10/16.*

We can show that for invertible matrices, we have $\kappa(A) = \frac{\max_{\|x\|=1}\|Ax\|}{\min_{\max_{\|x\|=1}\|Ax\|}}$ depending on the choice of the norm. By this it follows that:

1. $\kappa(A) \geq 1$
2. $\kappa(\alpha A) = \kappa(A) \forall \alpha \in \mathbb{R}, \alpha \neq 0$
   hence the condition number is invariant under scalar transformations (in contrast to determinants).
   Remember that $\det(\alpha A) = \alpha^n \det(A)$.
3. $A \in \mathbb{R}^{n\times n}$ with $A \neq 0$ is singular iff $\kappa(A) = \infty$

2.4 Componentwise condition analysis

The normwise condition analysis is inappropriate in particular special cases.

Example 2.11. *Consider the equation system $Ax = b$ with a diagonal matrix*

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \varepsilon \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\varepsilon} \end{pmatrix}$$

The problem is completely decoupled and we expect a well-conditioned problem (at least for diagonal deviations). For normwise conditioning (in regards of $\|\cdot\|_\infty$), it holds that

$$\kappa_\infty(A) = \left\|A^{-1}\right\|_\infty \|A\|_\infty = \frac{1}{\varepsilon} \qquad \varepsilon \to 0$$

For $\varepsilon \to 0$ the condition number will be arbitrary large, because it permits arbitrary deviations in the matrix.

By componentwise condition analysis, we get $(f, x) \rightsquigarrow$ smallest number κ_{rel} such that

$$\max_i \frac{|f_i(\tilde{x}) - f_i(x)|}{|f_i(x)|} \leq \max_i \frac{|\tilde{x}_i - x_i|}{|x_i|}$$

2.5 Stability of an algorithm

We now consider error, that are created, if the map f is approximated by some map $\tilde{f}$ (algorithmic realization). It encompasses all rounding errors and approximation errors and return an approximation $\tilde{f}(x)$ instead of $f(x)$.

Question: Is $\tilde{f}(x)$ acceptable as an substitute of $f(x)$? Compare with Figure 1.

Definition 2.12 (Forward stability)**.** *Let $\tilde{f}$ be the floating point implementation of an algorithm to solve problem (f, x) of relative normwise condition κ_{rel}. The* stability indicator *of a normwise forward analysis is the smallest number $\sigma \geq 0$, such that for all $\tilde{x} \in E$ it holds that*

$$\frac{\left\|\tilde{f}(\tilde{x}) - f(\tilde{x})\right\|}{\|f(\tilde{x})\|} \leq \sigma \kappa_{\text{rel}} \varepsilon \qquad \text{for } \varepsilon \to 0$$

2.6 Stability of the componentwise forward analysis

$\overline{\kappa}_{\text{rel}}$ is the smallest number $\sigma \geq 0$ such that

$$\max_i \frac{\left|\tilde{f}_i(\tilde{x}) - f_i(\tilde{x})\right|}{|f_i(\tilde{x})|} \dot{\leq} \sigma \cdot \overline{\kappa}_{\text{rel}} \varepsilon \qquad \text{for } \varepsilon \to 0$$

where $\overline{\kappa}_{\text{rel}}$ is the componentwise relative conditioning of (f, x). We call algorithm $\tilde{f}$ *stable* in regards of forward analysis if σ is smaller than the number of consecutively executed elementary operations.

Lemma 2.13. *For the elementary operations $\{+, -, \cdot, /\}$ and the floating point implementations $\left\{\hat{+}, \hat{-}, \hat{\cdot}, \hat{/}\right\}$ it holds that $\sigma \kappa_{\text{rel}} \leq 1$.*

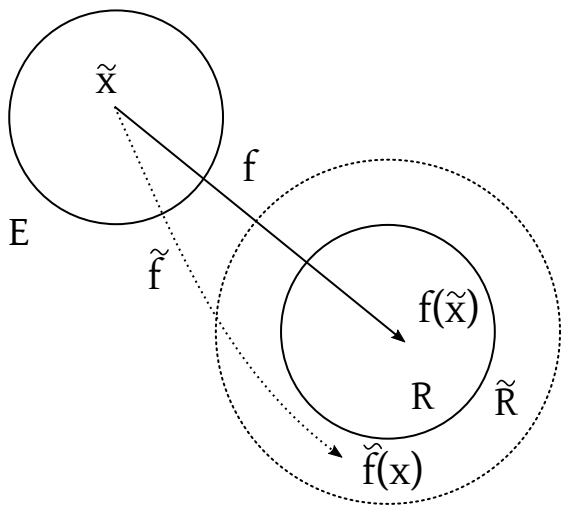

Figure 1: Map of the error with $\tilde{f}$ instead of f

Proof. For every floating point implementation it holds that

$$x \hat{*} y = (x * y)(1 + \delta) \qquad * \in \{+, -, \cdot, /\}$$

by Lemma 2.4 for some δ with $|\delta| \leq \varepsilon$. Thus,

$$\frac{|x \hat{*} y - x * y|}{x * y} = \frac{|(x * y)(1 + \delta) - x * y|}{|x * y|} = |\delta| \leq 1 \cdot \varepsilon$$

□

This approach is rather impractical (determination of the conditioning number of the problem is difficult).

Idea: interpret the error in the algorithm like an input error. The result $\tilde{y} = \tilde{f}(\tilde{x})$ is considered as the exact result $\tilde{y} = f(\hat{x})$ for distorted data $\hat{x}$. $E = \{\tilde{x} \mid \|\tilde{x} - x\| \leq \varepsilon \|x\|\}$. This is only possible if $\tilde{y}$ is a admissible solution of f at all. In the other case, we call the algorithm *unstable*.

Definition 2.14. *The* normwise backwards error *of the algorithm $\tilde{f}$ to solve the problem (f, x) is the smallest number $\eta \geq 0$, for which $\forall \tilde{x} \in E_\varepsilon$ some $\hat{x}$ exists such that*

$$\frac{\|\hat{x} - \tilde{x}\|}{\|\tilde{x}\|} \dot{\leq} \eta \qquad \textit{for } \varepsilon \to 0$$

The componentwise backwards error *is defined as*

$$\max_i \frac{|\hat{x}_i - \tilde{x}_i|}{|\tilde{x}_i|} \dot{\leq} \eta \qquad \textit{for } \varepsilon \to 0$$

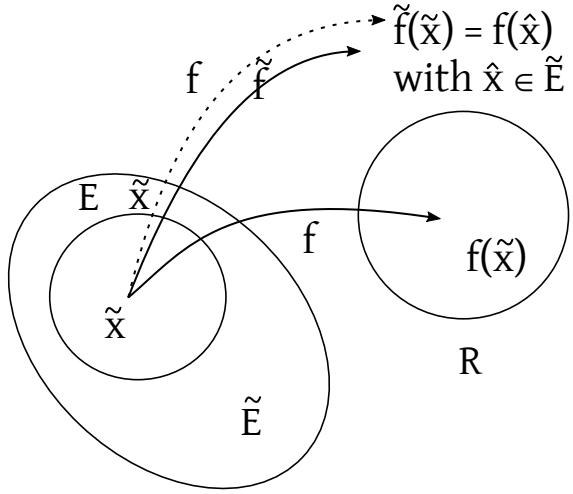

Figure 2: Backwards analysis

The algorithm is called stable *(in regards of backwards analysis) with respect to the relative input error ε if $\eta < \varepsilon \cdot$ number of elementary operations. For given ε we define a* stability indicator *of the backwards analysis as quotient*

$$\sigma_R := \frac{\eta}{\varepsilon}$$

Lemma 2.15. *For stability indicators σ and σ_R, of the forward/backward analysis respectively, it holds that $\sigma \leq \sigma_R$. Especially by backwards stability, forwards stability follows.*

Proof. By definition of the backwards error, for every $\tilde{x} \in E$ there exists some $\hat{x}$ such that $f(\hat{x}) = \tilde{f}(\tilde{x})$ and

$$\frac{\|\hat{x} - \tilde{x}\|}{\|\tilde{x}\|} \dot{\leq} \eta = \sigma_R \cdot \varepsilon \qquad \text{for } \varepsilon \to 0$$

For the relative error in the result, we have

$$\frac{\left\|\tilde{f}(\tilde{x}) - f(\tilde{x})\right\|}{\|f(\tilde{x})\|} = \frac{\|f(\hat{x}) - f(\tilde{x})\|}{\|f(\tilde{x})\|} \dot{\leq} \kappa_{\text{rel}} \frac{\|\hat{x} - \tilde{x}\|}{\|\tilde{x}\|} \dot{\leq} \kappa_{\text{rel}} \cdot \sigma_R \cdot \varepsilon \qquad (\text{for } \varepsilon \to 0)$$

□

Remark. *$Ax = b$ with solution $\tilde{x}$, approximation $\hat{x}$. Error $\|\hat{x} - \tilde{x}\|$ is not computable, because $\tilde{x}$ is unknown. Residue $\|b - A\hat{x}\|$.*

2.7 Backwards analysis for linear equation systems

Question: Is the LU decomposition a backwardsstable algorithm?

In the following, we will use the following notation for vectors and matrices:

$$|A| \leq |B| \iff |a_{ij}| \leq |b_{ij}| \, \forall i, j$$

Lemma 2.16. *The floating point implementation of the scalar product*

$$\langle x, y \rangle \coloneqq x_n y_n + \langle x^{n-1}, y^{n-1} \rangle$$

with $x^{n-1} \coloneqq (x_1, \dots, x_{n-1})^T$ *and* $y^{n-1} \coloneqq (y_1, \dots, y_{n-1})^T$ *denotes some solution* $\langle x, y \rangle_{rd} \coloneqq \mathrm{rd}(\langle x, y \rangle)$ *for some* $x, y \in \mathbb{R}^n$ *such that* $\langle x, y \rangle_{\mathrm{rd}} = \langle \hat{x}, y \rangle$ *for some* $\hat{x} \in \mathbb{R}^n$ *with* $|x - \hat{x}| \dot{\leq} n\varepsilon \, |x|$, *hence the relative componentwise backwards error is* $\eta \leq n \cdot \varepsilon$ *and the scalar product is stable in regards of backwards analysis (with* $2n - 1$ *elementary operations by the scalar product).*

Proof. We need to prove:

$$\forall i : |x_i - \hat{x}_i| \leq \underbrace{n\varepsilon \, |x_i|}_{h(\varepsilon)} + \overbrace{\sigma(\|h(\varepsilon)\|)}^{\sigma} \qquad \text{for } \varepsilon \to 0 \text{ with } \lim_{\varepsilon \to 0} \frac{\|\varphi(\varepsilon)\|}{\|h(\varepsilon)\|} = 0$$

Thus, we look for a function that converges superlinear towards 0 for $\varepsilon \to 0$ ($\mathcal{O}(\varepsilon)$).

We use induction over n to prove this. For $n = 1$, the statement is true by Lemma 2.4. Now let $n > 1$ and let the statement be true for $n - 1$. For floating point implementation of recursion it is true that

$$\langle x, y \rangle_{\mathrm{rd}} = (x_n y_n (1 + \delta) + \langle x^{n-1}, y^{n-1} \rangle_{\mathrm{rd}})(1 + \tilde{\delta})$$

where $|\delta| \leq \varepsilon$ and $|\tilde{\delta}| \leq \varepsilon$ characterize the relative error of multiplication and addition respectively.

By the induction hypothesis, it holds that $\langle x^{n-1}, y^{n-1} \rangle_{\mathrm{rd}} = \langle z, y^{n-1} \rangle$ for some $z \in \mathbb{R}^{n-1}$ with

$$|x^{n-1} - z| \leq (n-1)\varepsilon \, |x^{n-1}| + \mathcal{O}(\varepsilon)$$

By $\hat{x}_n = x_n(1 + \delta)(1 + \tilde{\delta})$ and $\hat{x}_k = z_k(1 + \tilde{\delta})$ by $k = 1, \dots, n - 1$ it follows that

$$\begin{aligned}
\langle x, y \rangle_{\mathrm{rd}} &= x_n y_n (1 + \delta)(1 + \tilde{\delta}) + \langle z(1 + \varepsilon), y^{n-1} \rangle \\
&= \hat{x}_n y_n + \langle \hat{x}^{n-1}, y^{n-1} \rangle = \langle \hat{x}, y \rangle \\
&\text{with } |x_n - \hat{x}_n| \leq 2\varepsilon \, |x_n| + |x_n| \, |y_n| \, \delta\tilde{\delta} \\
&\leq n\varepsilon \, |x_n| + \mathcal{O}(\varepsilon) \\
\text{and } |x_k - \hat{x}_k| &\leq |x_k - z_k| + |x_k - \hat{x}_k| \\
&\leq (n-1)\varepsilon \, |x_k| + \mathcal{O}(\varepsilon) + \varepsilon \, |z_k| \\
&\leq n\varepsilon \, |x_k| + \mathcal{O}(\varepsilon) \text{ for } k = 1, \dots, n - 1
\end{aligned}$$

The last estimate can be made using

$$\begin{aligned}
|z_k| - |x_k| \leq |x_k - z_k| &\leq (n-1)\varepsilon \, |x_k| + \mathcal{O}(\varepsilon) \\
\implies \varepsilon \, |z_k| &\leq \varepsilon \, |x_k| + \mathcal{O}(\varepsilon)
\end{aligned}$$

□

Theorem 2.16.1. *The floating point implementation of forward substitution to solve a linear equation system $Lx = b$ with some lower triangular matrix L computes the solution $\hat{x}$ such that there exists some lower triangular matrix $\hat{L}$ with $\hat{L}\hat{x} = b$ and $\left|L - \hat{L}\right| \dot{\leq} n\varepsilon \left|L\right|$. Hence for componentwise relative backwards errors, we have $\eta \leq n \cdot \varepsilon$ and the forward substitution is stable in regards of backwards analysis.*

Proof. Again, we will use a recursive approach.

$$l_{kk}x_k = b_k - \left\langle l^{k-1}, x^{k-1}\right\rangle$$

For $k = 1, \dots, n$, we have

$$\begin{aligned} x^{k-1} &\coloneqq (x_1, \dots, x_{k-1})^T \text{ and} \\ l^{k-1} &\coloneqq (l_{k,1}, \dots, l_{k,k-1})^T \end{aligned}$$

In floating point arithmetics, the following recursion is yielded,

$$l_{kk}(1+\delta_k)(1+\tilde{\delta}_k)\hat{x}_k = b_k - \left\langle l^{k-1}, x^{k-1}\right\rangle_{\text{rd}} (1+\tilde{\delta}_k)$$

where $\delta_k, \tilde{\delta}_k$ with $|\delta_k| \leq \varepsilon$ and $\left|\tilde{\delta}_k\right| \leq \varepsilon$ are the relative errors of multiplication/addition. By Lemma 2.16, it follows that

$$\left\langle l^{k-1}, \hat{x}^{k-1}\right\rangle_{\text{rd}} = \left\langle \tilde{l}^{k-1}, \hat{x}^{k-1}\right\rangle$$

for some vector $\tilde{l}^{k-1} = (\tilde{l}_{k,1}, \dots, \tilde{l}_{k,k-1})^T$ with $\left|l^{k-1} - \tilde{l}^{k-1}\right| \dot{\leq} (k-1)\varepsilon \left|l^{k-1}\right|$. Now let $\hat{l}_{kk} \coloneqq l_{kk}(1+\delta_k)(1+\tilde{\delta}_k)$ and $\hat{l}^{k-1} \coloneqq (1+\tilde{\delta}_k)\tilde{l}^{k-1}$ then

$$\begin{aligned} \left|\hat{l}^{k-1} - l^{k-1}\right| &\dot{\leq} (k-1)\varepsilon \left|l^{k-1}\right| + \left|\tilde{\delta}_k\right| \left|\tilde{l}^{k-1}\right| \\ &\dot{\leq} k\varepsilon \left|l^{k-1}\right| \\ \left|\hat{l}_{kh} - l_{kj}\right| &\dot{\leq} 2\varepsilon \\ \implies \hat{L}\hat{x} = b \wedge \left|L - \hat{L}\right| &\dot{\leq} n\varepsilon \left|L\right| \end{aligned}$$

□

↓ *This lecture took place on 2018/10/18.*

Lemma 2.17. *Let A have a LU decomposition. Then Gaussian elimination determines $\hat{L}$ and $\hat{U}$ such that $\hat{L}\hat{U} = \hat{A}$ for some matrix $\hat{A}$ with*

$$\left|A - \hat{A}\right| \dot{\leq} n \left|\hat{L}\right| \left|\hat{U}\right| \varepsilon$$

Theorem 2.17.1. *A has some LU-decomposition. Then Gaussian elimination for linear equation system $Ax = b$ gives a solution $\hat{x}$ with $\hat{A}\hat{x} = b$ for some matrix $\hat{A}$ with*

$$\left|A - \hat{A}\right| \dot{\leq} 2n \left|\hat{L}\right| \left|\hat{U}\right| \varepsilon$$

Theorem 2.17.2. *The Gaussian elimination with column pivot strategy for the linear equation system $Ax = b$ determines some $\hat{x}$ such that $\hat{A}\hat{x} = b$ for some matrix $\hat{A}$ with*

$$\frac{\left\|A - \hat{A}\right\|_\infty}{\|A\|_\infty} \dot{\leq} 2n^3 \rho_n(A)\varepsilon \qquad \text{where } \rho_n(A) := \frac{\alpha_{\max}}{\max_{i,j} |a_{ij}|}$$

and where $\alpha_{\max}$ is the largest absolute value of an element that occurs during elimination in the remainder matrices $A^{(1)} = A$ to $A^{(n)} = U$.

$$A = \begin{bmatrix} 1 & 0 & \dots & 0 & 1 \\ -1 & \vdots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & & \\ -1 & \dots & \dots & -1 & 1 \end{bmatrix}$$

For matrix A it holds that $\rho_n(A) = 2^{n-1}$.

Consider that the previous result also provides *no* positive statement about the backwards stability of the LU decomposition.

3 Linear least squares

Goal: Solution of overdetermined linear equation systems

$$Ax = b \qquad A \in \mathbb{R}^{m \times n} \qquad m > n \qquad b \in \mathbb{R}^m$$

by the method of linear least squares.

For this, we are going to use (numerically stable) orthogonal transformations.

3.1 Gaussian process of least squares

Problem: Given m input data points (t_i, b_i) and $t_i, b_i \in \mathbb{R}$ for $i = 1, \dots, m$. They describe the values b_i of an object at timestamp t_i.

We assume that there exists an underlying law such that the dependency of b and t can be represented by some *modelling function* φ with $b(t) = \varphi(t_i; x_1, \dots, x_n)$. The modelling function therefore takes n unknowns p_i.

Example 3.1 (Ohm's law). *Consider Ohm's law $b = xt = \varphi(t_i x)$ where t denotes the current, b the voltage and x resistance.*

So want to approximate data points in the $\mathbb{R}^2$ plane linearly.

If the model would be *exact* and if you have no errors in measurement, then we are supposed to evaluate parameters $x_1, \dots, x_n$ such that $b_i = b(t_i) = \varphi(t_i; x_1, \dots, x_n)$ for $i = 1, \dots, m$.

In reality, measurements are subject to errors and modelling functions are also just approximations of reality. We therefore require that $b_i \approx \varphi(t_i; x_1, \dots, x_n)$ for $i = 1, \dots, m$. We weight the individual deviations

$$\Delta i \coloneqq b_i - \varphi(t_i; x_1, \dots, x_n) \qquad i = 1, \dots, m$$

By Gauss, determine x_i such that $\sum_{i=1}^m \Delta i^2$ becomes minimal. We use the short notation $\Delta^2 \coloneqq \sum_{i=1}^m \delta_i^2 \implies$ min. Often it is more useful to weight the individual deviations differently. For some measurement precision δb, introduce weights

$$\sum_{i=1}^m \left(\frac{\Delta i}{\delta b_i} \right) \to \text{ min}$$

We consider the special case of linear (with respect to x) modelling functions φ, hence $\varphi(t_i; x_1, \dots, x_n) = a_1(t)x_1 + \dots + a_n(t)x_n$ where $a_1, \dots, a_n : \mathbb{R} \to \mathbb{R}$ are arbitrary functions. Furthermore we choose $\|\cdot\| = \|\cdot\|_2$. We can denote the least square problem as

$$\|b - Ax\| \to \text{min} \qquad \text{where } b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}, x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$\text{and } A = (a_{ij}) \in \mathbb{R}^{m \times n} \text{ with } a_{ij} \coloneqq a_j(t_i)$$

Orthogonal equations

We look for some point $z = Ax$ of the image space $U(A)$ of A, that has the smallest distance to some given b. Ax is the orthogonal projection of b to the subspace $U(A) = \text{image}(A)$.

Theorem 3.1.1. *Let V be a vector space of $\dim V < \infty$ with scalar product $\langle , \rangle$. Let $U \subset V$ be a subspace. Let*

$$U^\perp = \{v \in V \mid \langle v, u \rangle = 0 \forall u \in U\}$$

be the orthogonal complement in V. Then for all $v \in V$ with respect to the norm induced by $\|v\| = \sqrt{\langle v, v \rangle}$ such that

$$\|v - u\| = \min_{\tilde{u} \in U} \|v - \tilde{u}\| \iff v - u \in U^\perp$$

Proof sketch:

- $V = U \oplus U^\perp$
- $v = u + w, u \in U, w \in U^\perp$ unique
- $\tilde{u} \in U$ arbitrary:

  $$\|v - \tilde{u}\|^2 = \|(v - u) + (u - \tilde{u})\|^2 = \|v - u\|^2 + \|u - \tilde{u}\|^2 \geq \|v - u\|^2$$

  with $v - u = w$ and $(u - \tilde{u}) \in U$.

↓ *This lecture took place on 2018/10/23.*

The unique solution $u \in U$ of $\min \|v - u\|$ is called *orthogonal projection* of V to U. The map $P : V \to U$ with $v \mapsto Pv$ with $Pv = u$ is linear and is called *orthogonal projection* of V to U.

Remark 3.2. *We get an analogous statement if we replace U by an affine subspace $W = w_0 + U \subseteq V$ where $w_0 \in V$. U is the parallel, to W, linear subspace of V.*

For all $v \in V$, $w \in W$: $\|v - w\| = \min_{\tilde{w} \in w} \|v - \tilde{w}\| \iff v - w \in U^\perp$.

Theorem 3.2.1. *The vector $x \in \mathbb{R}^n$ is a solution of the linear least squares problems $\min \|b - Ax\| \iff x$ satisfies the so-called orthogonal equations $A^T Ax = A^T b$ (dt. "Normalengleichungen").*

Especially, the linear least squares problem is uniquely solvable iff A has full column rank.

Proof. By Theorem 3.1.1, $V := \mathbb{R}^n$, $U := \text{image}(A) = \text{rank}(A)$.

$$\begin{aligned} \implies \max \|b - Ax\| &\iff \langle b - Ax, A\tilde{x} \rangle = 0 \forall \tilde{x} \in \mathbb{R}^n \\ &\iff 0 = \langle A^t(b - Ax), \tilde{x} \rangle \ \forall \tilde{x} \in \mathbb{R}^n \\ &\iff A^t(b - Ax) = 0 \\ &\iff A^t b = A^t Ax \end{aligned}$$

Thus the first statement is proven. How about the second statement?

$$A^t A \text{ regular} \iff \text{rank}(A) = n$$

□

Remark (Geometric interpretation). *$b - Ax$ is orthogonal to* $\text{range}(A)$ *with $A \subseteq \mathbb{R}^n$.*

Solution of the orthogonal equation system: Theoretically, we can create $A^T A$ and retrieve a symmetric positive definite matrix ($\implies$ decomposition with Cholesky process).

Remark (Computing time).

1. *Determination of $A^t A$ (only one half, because of symmetry)*

$$(A^t A)_{ij} = (\langle a_j, a_j \rangle)_{i,j} \qquad a_j = j\text{-th column of } A \to \frac{1}{2} n^2 n$$

2. *Cholesky decomposition $\sim \frac{1}{6} n^3$*

In case of $m \gg n$, the first approach will take over $\implies \sim \frac{1}{2} n^2 n$. $\frac{2}{3} m^3$ for $n \approx m$.

Lemma 3.3. *For some matrix $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and full column rank* ($\text{rank}(A) = n$), *$\kappa_2(A^t A) = (\kappa_2(A))^2$.*

Proof. By definition of condition numbers:

$$
\begin{aligned}
(\kappa_2(A))^2 &= \frac{\max_{\|x\|=1} \|Ax\|_2^2}{\min_{\|x\|=1} \|Ax\|_2^2} \\
&= \frac{\max_{\|x\|=1} \langle Ax, Ax\rangle}{\min_{\|x\|=1} \langle Ax, Ax\rangle} \\
&= \frac{\max \langle A^t Ax, x\rangle}{\min \langle A^t Ax, x\rangle} && \text{“Rayleigh eigenvalue”} \\
&\overset{A^t A \text{sym.}}{=} \frac{\lambda_{\max}(A^t A)}{\lambda_{\min}(A^t A)} \\
&\overset{A^t A \text{sym.}}{=} \kappa_2(A^t A)
\end{aligned}
$$

□

Recognize that the amplification of the error for the orthogonal equation is larger. Therefore it makes sense to avoid computing $A^T A$ and use the original matrices A and A^T instead, e.g.

$$
\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ 0 \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}
$$

$$
\implies A^T r = 0 \wedge r \neq Ax = b
$$

$$
\implies r = b - Ax \qquad \implies 0 = A^T r = A^T(b - Ax)
$$

$$
\|A_3\|_2^2 = \max_{\|x\|_2=1} \|Q_j x\|_2^2 = \max_{\|x\|_2=1} \langle x, Q^T Q x\rangle = 1
$$

Recognize that $Q^T Q = I$.

3.2 Orthogonalization process

We can represent the elimininatation process for linear equation systems (e.g. Gaussian elimination) formally as

$$
A \xrightarrow{f_1} B_1 A \xrightarrow{f_2} B_2 B_1 A \xrightarrow{f_3} \dots \xrightarrow{f_k} B_k \dots B_1 A \cdots \to U
$$

where

$$
U = \begin{pmatrix} 1 & * \\ 0 & \varepsilon \end{pmatrix}
$$

and where matrices B_j describe the operations applied to matrix A. Through this process, the condition numbers of the individual matrices can be amplified, such that instabilities can occur. If we instead use orthogonal transformations Q_j for elimination, we have

$$
\kappa_2(Q_j) = \|Q_j\|_2 \cdot \|Q_j\|_2^{-1} = \|Q_j\|_2 \|Q_j\|_2 = 1
$$

Orthogonalization process are necessarily stable, but have higher computing times compared to Gaussian elimination. If we assume we transformed a given matrix $A \in$

$\mathbb{R}^{m\times n}$ with $m \geq n$ with some orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ into upper triangular matrix, then

$$Q^T A = \begin{bmatrix} * & \dots & x \\ & \ddots & \vdots \\ & & x \\ 0 & & * \\ \vdots & & \vdots \\ 0 & & 0 \end{bmatrix} = \begin{bmatrix} U \\ 0 \end{bmatrix}$$

Theorem 3.3.1. *Let $A \in \mathbb{R}^{m\times n}$, $m \geq n$, of maximum rank $n, b \in \mathbb{R}^m$ and $Q \in \mathbb{R}^{m\times m}$ be an orthogonal matrix with*

$$Q^T A = \begin{bmatrix} U \\ 0 \end{bmatrix} \text{ and } Q^T b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

where $b_1 \in \mathbb{R}^n$, $b_2 \in \mathbb{R}^{m-n}$ and $U \in \mathbb{R}^{n\times n}$ is an (invertible) upper triangular matrix. Then $x = U^{-1}b_1$ is the solution of the linear least squares problems $\min \|b - Ax\|_2$.

Proof. Because Q is orthogonal, $\forall x \in \mathbb{R}^n$ it follows that

$$\|b - Ax\|_2^2 = \left\|Q^T(b - Ax)\right\|_2^2 = \left\|\begin{pmatrix} b_1 - Ux \\ b_2 \end{pmatrix}\right\|_2^2 = \|b_1 - Ux\|_2^2 + \|b_2\|_2^2 \geq \|b_2\|_2^2$$

Because $\operatorname{rank}(A) = \operatorname{rank}(U) = n$, U is invertible. The term $\|b_1 - Ux\|_2^2$ disappear for $x = U^{-1}b_1$ □

Remark. *A:* $0 \neq \|r\| = \|b - Ax\| = \|b_2\|$.

What do orthogonal transformations look like? In general, these transformations are rotations and reflections.

e.g. $x \mapsto \alpha e_1 = Qx$ with $Q = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$.

Alternatively, $x \mapsto \alpha e_1 = x = 2\frac{\langle v,x\rangle}{\langle v,v\rangle} v$ where v is collinear to the difference $x = \alpha e_1$.

Givens rotations

Givens rotations are described by matrices of form

$$Q_{k,l} := \begin{bmatrix} I & & & & \\ & c & & s & \\ & & I & & \\ & -s & & c & \\ & & & & I \end{bmatrix} \in \mathbb{R}^{m\times m}$$

with proper dimension of the unit matrix and $c^2 + s^2 = 1$. For $x \in \mathbb{R}^m$ it holds that

$$x \mapsto y = Q_{k,l}x \text{ with } y_i = (Q_{k,l}x)_i = \begin{cases} cx_k + sx_l & i = k \\ -sx_k + cx_l & i = l \\ x_i & \text{else} \end{cases}$$

Thus for $A = [a_1, \dots, a_n] \in \mathbb{R}^{m \times n}$, we get

$$Q_{k,l} A = [Q_{k,l} a_1, \dots, Q_{k,l} a_n]$$

Hence, only rows k and l of matrix A will be modified.

Question: How to choose c and s to eliminate a component x_l of x? Because $Q_{k,l}$ operates on the (k, l)-layer, we consider without loss of generality $m = 2$. By $x_k^2 + x_l^2 \neq 0$ and $s^2 + c^2 = 1$ it follows that

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} x_k \\ x_l \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}$$

$$\iff r = \pm \sqrt{x_k^2 + x_l^2}, c = \frac{x_k}{r}, s = \frac{x_l}{r}$$

Computing x_k^2 might gives us an exponent overflow.

$$\tau := \frac{x_k}{x_l}, s := \frac{1}{\sqrt{1 + \tau^2}}, c := s\tau \text{ if } |x_l| > |x_k|$$

and accordingly,

$$\tau := \frac{x_l}{x_k}, c = \frac{1}{\sqrt{1 + \tau^2}}, s := c\tau \text{ if } |x_k| \geq |x_l|$$

Now transform A iteratively into triangular form.

For example, consider

$$A = \begin{bmatrix} A & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \xrightarrow{(5,4)} \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \end{bmatrix} \xrightarrow{(4,3)} \dots$$

$$\xrightarrow{(2,1)} \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{bmatrix} \xrightarrow{(5,4)} \dots \xrightarrow{(4,3)} \dots \xrightarrow{(5,4)} \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Remark (Computing time). *Computing time for a dense matrix $A \in \mathbb{R}^{m \times n}$:*

1. $\sim \frac{n^2}{2}$ *square roots and* $\frac{4}{3} n^3$ *multiplications, if* $m \approx n$.
2. $m \cdot n$ *are square root and* $2mn^2$ *multiplications if* $m \ggg n$.

This is an alternative for Gaussian elimination. The algorithm is *more stable*, but *more expensive*.

But it can be advantageous for special matrix, e.g. Hessenberg matrices, which are "almost triangular" (so the first subdiagonal also contains non-zero entries). Here, we need $n - 1$ Givens rotations if we want to achieve upper triangular form.

Householder reflections

↓ *This lecture took place on 2018/10/25.*

For given $0 \neq v \in \mathbb{R}^m$, we define the matrix $Q \in \mathbb{R}^{m \times m}$.

$$Q = I - 2\frac{vv^T}{v^T v}$$

They describe reflections on the mirror plane, which are vertical to v. The following properties hold:

1. symmetry: $Q = Q^T$
2. orthogonality: $QQ^T = I$
3. $Q^2 = \text{id}$, thus Q is *involutional*
4. $\text{spec}(Q) = \{\pm 1\}$, -1 simple, 1 $(n-1)$-times.

For $y \in \mathbb{R}^n$,

$$y \mapsto Qy = \left(I - 2\frac{vv^T}{v^T v}\right) y = y - 2 \cdot \frac{\langle v, y \rangle}{\langle v, v \rangle}$$

Question: How can we achieve that Q maps vector y to $\alpha \cdot e_1$? Hence,

$$\alpha \cdot e_1 = y - 2\frac{\langle v, y \rangle}{\langle v, v \rangle} \cdot v \in \text{span}(e_1) = \alpha(\{e_1\})$$

Consider that

$$|\alpha| = \|\alpha e_1\| = \left\| y - 2\frac{\langle v, y \rangle}{\langle v, v \rangle} v \right\|_2 = \|Qy\|_2 = \|y\|_2$$

and $y \in \text{span}(y - \alpha \cdot e_1)$. This way we can retrieve Q by,

$$v \coloneqq y - \alpha e_1 \qquad \alpha = I \|y\|_2$$

To avoid elimination (or subtraction) in $v = (y_1 - \alpha, y_2, \dots, y_n)^T$, choose $\alpha \coloneqq -\text{sign}(y_1) \cdot \|y\|_2$. Now,

$$\langle v, v \rangle = \langle y - \alpha e_1, y - \alpha e_1 \rangle = \|y\|^2 - 2\alpha \langle e_1, y \rangle + \alpha^2 = 2\alpha^2 - 2\alpha \langle e_1, y \rangle = -2\alpha(y_1 - \alpha)$$

Qy for $y \in \mathbb{R}^m$ can be determined by

$$Qy = y - 2\frac{\langle v, y \rangle}{2(y_1 - \alpha)} \cdot v$$

In a first step choose Q_1 (where $A \in \mathbb{R}^{m \times n}$):

$$[a_1 \dots a_n] = A \to A' = Q_1 \cdot A = \begin{bmatrix} \alpha_1 & \\ 0 & \\ \vdots & a'_2 \dots a'_n \\ 0 & \end{bmatrix}$$

where $Q_1 = I - 2\frac{v_1 v_1^T}{v_1^T v_1}$ with $v_1 := a_1 - \alpha_1 e_1$ and $\alpha_1 := -\operatorname{sign}(a_{11}) \cdot \|a_1\|$.

After the k-th step, we get

$$A^{(k)} = \begin{bmatrix} * & & & * \\ & \ddots & & \\ & & * & \\ 0 & & & T^{(k+1)} \end{bmatrix}$$

with $T^{(k+1)} \in \mathbb{R}^{m-k \times n-k}$. Now we create orthogonal matrix

$$Q_{k+1} := \begin{bmatrix} I & 0 \\ 0 & \tilde{Q}_{k+1} \end{bmatrix}$$

where $\tilde{Q}_{k+1}$ modifies matrix $T^{(k+1)}$. The next subcolumn will be eliminated.

After $p = \min(m, n)$ steps, we retrieve an upper triangular matrix

$$R = Q_p \dots Q_1 \cdot A$$
$$\iff A = QR \text{ where } Q = Q_1 \cdot \dots \cdot Q_p$$

Based on Theorem 2.6, we retrieve the following process:

1. $A = QR$ (QR decomposition using Householder or Givens)
2. $\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = Q^t b$ with $b_1 \in \mathbb{R}^n$ and $b_2 \in \mathbb{R}^{m-n}$ (transformation of b)
3. $Rx = b_1$ (solving by backwards substitution)

Remark (Implementation). *Usually, we store the Householder vectors $v_1, \dots, v_p$ in the lower half of A and the diagonal elements $v_{1i} = \alpha_i$ with $i = 1, \dots, p$ in a separate vector.*

Remark (Computing time).

- $\sim 2n^2 m$ *multiplications if* $m \gg n$
- $\frac{2}{3}n^3$ *if* $m \approx n$

It is important to recognize that this process is numerically stable!

3.3 Singular value decomposition

Literature:

- Gene H. Golub and Charles F. Van Loan: "Matrix computations"

If $A \in \mathbb{R}^{m \times n}$ has non-full rank n, then the linear equation system of $\min \|b - Ax\|$ is not unique, because $\forall \tilde{x} : \|b - A\tilde{x}\| = \|b - A(\tilde{x} + v)\| \; \forall v \in \ker(A)$.

Are there more decompositions of the matrix?

Recognize that the QR decomposition always exists anyways.

Theorem 3.3.2. *Let $A \in \mathbb{R}^{m\times n}$. There $\exists U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n} : U^TU = I, V^TV = I$ and $A = U\Sigma V^T$ with $\Sigma \in \mathbb{R}^{n\times n}$ where $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_p), p = \min(m, n), V = [v_1, \dots, v_n]$ and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p \geq 0$.*

Remark 3.4. *σ_i are called "singular values" of A. u_i and v_i are called left- and right-singular vectors of A.*

↓ *This lecture took place on 2018/10/30.*

Proof. Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ such that $\|x\|_2 = 1 = \|y\|_2$ and $Ax = \sigma y$ where $\sigma = \|A\|_2$. Recognize that x, y exist because $\|A\|_2 = \sup_{\|\tilde{x}\|=1} \|A\tilde{x}\|_2 = \max_{\|\tilde{x}\|_2=1} \|A\tilde{x}\|_2$.

Hence, there exists x with $\|x\|_2 = 1$ and $\|Ax\|_2 = \|A\|_2$. Followingly, we define $y := \frac{Ax}{\|A\|_2}$ and it is true that $\sigma y = Ax$.

$$\|y\|_2 = \frac{\|Ax\|_2}{\|A\|_2} = 1$$

We construct orthonormal bases $\{x, v_2, \dots, v_n\}$ and $\{y, u_2, \dots, u_m\}$. The corresponding matrices $U = \begin{bmatrix} y & u_2 & \dots & u_m \end{bmatrix}$ and $V = \begin{bmatrix} x & v_2 & \dots & v_n \end{bmatrix}$ are orthogonal. $V^TV = I, U^TU = I$. Let $U_2 := (u_2, \dots, u_m)$ and $V_2 := (v_2, \dots, v_n)$. Furthermore,

$$U^TAV = \begin{bmatrix} y^T \\ U_2^T \end{bmatrix} A \begin{bmatrix} X & V_2 \end{bmatrix} = \begin{bmatrix} y^TAx & y^TAV_2 \\ U_2^TAx & U_2^TAV_2 \end{bmatrix} = \begin{bmatrix} \sigma & w^T \\ 0 & B \end{bmatrix} := A_1$$

where $w \in \mathbb{R}^{n-1}$ and $B \in \mathbb{R}^{m-1\times n-1}$ because $\left\| A_1 \left(\begin{bmatrix} \sigma \\ w \end{bmatrix} \right) \right\|_2^2 \geq (\sigma^2 + w^Tw)$ it holds that $\|A_1\|_2^2 \geq (\sigma^2 + w^TW)$. On the other hand, $\sigma^2 = \|A\|_2^2 = \|A_1\|_2^2$. Thus $w = 0$. The statement then follows inductively. □

Properties of the singular value decomposition

Corollary 3.5. *If $A = U\Sigma V^T \in \mathbb{R}^{m\times n}$ and $m \geq n$ is the singular value decomposition of A, then $Av_i = \sigma_i u_i$ and $A^Tu_i = \delta_i v_i$ for all $i = 1, \dots, n$.*

Proof. Consider columnwise the equality $AV = U\Sigma$ and accordingly, $A^TU = V\Sigma^T$. □

By the corollary, it immediately follows that

$$A^TAv_i = \sigma_i^2 v_i \qquad AA^Tu_i = \sigma_i^2 u_i$$

hence the squares of singular values are the eigenvalues of the matrices A^TA, and accordingly AA^T.

Illustrative interpretation: The singular values of a matrix A give the length of the semi-axes of the ellipsis defined by $E = \{Ax \mid \|x\|_2 = 1\}$. The directions of semi-axes are denotes by u_i.

$$\|A\|_2^2 = \sup_{x \neq 0} \frac{\langle Ax, Ax \rangle}{\langle x, x \rangle} = \sup_{x \neq 0} \frac{\langle A^T Ax, x \rangle}{\langle x, x \rangle} = \lambda_{\max}(A^T A)$$

Corollary 3.6. *Let $A \in \mathbb{R}^{m \times n}$. Then,*

$$\|A\|_2 = \sigma_1 \qquad \|A\|_F = \sqrt{\operatorname{trace}(A^T A)} = \sqrt{\sigma_1^2 + \cdots + \sigma_p^2}$$

Proof. Consider that for spectral norm and Frobenius norm, it holds that $\|A\| = \|U^T AV\| = \|\Sigma\|$. □

Corollary 3.7. *If A has exactly r positive singular values, then $rK(A) = r$ and* $\ker(A) = \operatorname{span}\{v_{r+1}, \dots, v_n\}$ *and* $\operatorname{range}(A) = \operatorname{span}\{u_1, \dots, u_r\}$.

Proof. Because U and V are regular, it holds that $\operatorname{rank}(A) = \operatorname{rank}(U\Sigma V^T) = \operatorname{rank}(\Sigma) = r$.

The statements regarding the kernel and image of A result by Corollary 3.5. □

Corollary 3.8. *Let $A \in \mathbb{R}^{m \times n}$ and* $\operatorname{rank}(A) = r$. *Then $A = \sum_{i=1}^r \sigma_i u_i v_i^T$.*

Proof. Again, consider the matrix notation of the singular value decomposition:

$$A = (U\Sigma)V^T = \begin{bmatrix} \sigma_1 u_1 & \sigma_2 u_2 & \dots & \sigma_r u_r & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_n^T \end{bmatrix} = \sum_{r=1}^r \sigma_i u_i v_i^T$$

□

Theorem 3.8.1 (Eckart-Young-Mirsky). *If $k < r = \operatorname{rank}(A)$ and $A_k = \sum_{i=1}^k \sigma_i u_i v_i^T$, then* $\min_{\operatorname{rank}(B)=k} \|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1}$.

Proof. Because $U^T A_k V = \operatorname{diag}(\sigma_1, \dots, \sigma_k, 0, \dots, 0)$, $\operatorname{rank}(A_k) = k$. Furthermore,

$$\left\|U^T (A - A_k) V\right\|_2 = \operatorname{diag}(0, \dots, 0, \sigma_{k+1}, \dots, \sigma_p)$$

and therefore $\|A - A_k\|_2 = \sigma_{k+1}$. Now let $B \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(B) = k$ arbitrary. Then we can find an orthonormal basis $x_1, \dots, x_{n-k}$ such that $\operatorname{kernel}(B) = \operatorname{span}\{x_1, \dots, x_{n-k}\}$. Furthermore it holds that $\operatorname{span}\{x_1, \dots, x_{n-k}\} \cap \operatorname{span}\{v_1, \dots, v_{k+1}\} \neq \{0\}$. Let z with $\|z\|_2 = 1$ of this intersection. Then $Bz = 0$ and $Az = A \sum_{i=1}^{k+1} \alpha_i v_i = \sum_{i=1}^{k+1} \sigma_i \alpha_i u_i$. Because $\|z\|_2 = 1$, $1 = \|z\|_2^2 = \left\|\sum_{i=1}^{k+1} \alpha_i v_i\right\|_2^2 = \sum_{i=1}^{k+1} \alpha_i^2$. Followingly,

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 \alpha_i^2 \geq \sigma_{k+1}^2 \underbrace{\sum_{i=1}^{k+1} \alpha_i^2}_{=1} = \sigma_{k+1}^2$$

□

We go back to the topic of $\min_x \|b - Ax\|_2$.

Consider the equation $b \cdot Ax = d$ and we followingly want to minimize the norm of the residue $\|d\|_2$. Left-sided multiplication with U^T gives

$$U^T(b - Ax) = U^T d \implies U^T b - U^T A V V^T x = U^T d = \tilde{d}$$

Because of orthogonality of U, $\left\|\tilde{d}\right\|_2 = \|d\|_2$ and thus we can minimize $\left\|\tilde{d}\right\|_2$. For this we introduce auxiliary vectors $\tilde{x} = V^T x \in \mathbb{R}^n$ and $\tilde{b} = U^T b \in \mathbb{R}^m$. Because of the singular value decomposition, we retrieve the special form of $U^T b - U^T A V V^T x = U^T d = \tilde{d}$

$$\begin{aligned} -\sigma_i \tilde{x}_i + \tilde{b}_i &= \tilde{d}_i & i &= 1, \dots, r \\ \tilde{b}_i &= \tilde{d}_i & i &= r+1, \dots, m \end{aligned} \tag{3}$$

The last $m - r$ components are independent of $\tilde{x}_i$. The term $\|\tilde{a}\|_2^2 = \sum_{i=1}^m \tilde{d}_i^2$ will become minimal if $\tilde{d}_i = 0$ for $i = 1, \dots, r$. In this case it holds that

$$\min \left\|\tilde{d}\right\|_2^2 = \sum_{i=r+1}^m \tilde{d}_i^2 = \sum_{i=r+1}^m \tilde{b}_i^2 = \sum_{i=r+1}^m (u_i^T b)^2$$

The first r unknowns are given by

$$\tilde{x}_i = \frac{\tilde{b}_i}{\sigma_i} \qquad i = 1, \dots, r$$

because of equations 3 and simultaneously the last $n - r$ unknowns can be chosen freely. The solution vector x therefore has the following representation:

$$x = V\tilde{x} = \sum_{i=1}^n \tilde{x}_i v_i = \sum_{i=1}^r \frac{\tilde{b}_i}{\sigma_i} v_i + \sum_{i=r+1}^n \tilde{x}_i v_i = \sum_{i=1}^r \frac{u_i^T b}{\sigma_i} v_i + \sum_{i=r+1}^n \tilde{x}_i v_i$$

with the free parameters $\tilde{x}_{r+1}, \dots, \tilde{x}_n$. If matrix A does not have full column rank, then the generic solution is representable as sum of a particulate solution in the subspace of r right-singular vectors v_i to *positive* singular values σ_i and an arbitrary vector of kernel(A).

Often, the specific solution x^* with minimal Euclidean norm is used. Because of orthogonality, it is characterized by $\tilde{x}_i = 0$ for $i = r+1, \dots, n$ and we get

$$x^* = \sum_{i=1}^r \frac{u_i^T b}{\sigma_i} v_i \qquad \|x^*\|_2 = \min_{\|b - Ax\|_2^2} \|x\|_2$$

with $\|b - Ax\|_2^2 = \sum_{i=r+1}^n (u_i^T b)^2$.

Generalized inverse

Using the singular value decomposition, we can also give the inverse of a regular matrix $A \in \mathbb{R}^{n \times n}$, because $A^{-1} = (U\Sigma V^T)^{-1} = V\Sigma^{-1}U^T$. For arbitrary matrices

$A \in \mathbb{R}^{m \times n}$ with $\operatorname{rank}(A) = r$, we can define the *generalized inverse* by

$$A^\dagger = [v_1 \quad \dots \quad v_r] \operatorname{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_r}\right) \begin{bmatrix} u_1^T \\ \vdots \\ u_r^T \end{bmatrix}$$

This gives the so-called *Moore-Penrose inverse*.

4 Interpolation

Problem setting: For an unknown function $f : \mathbb{R} \to \mathbb{R}$, construct from given data $f^{(j)}(t_i)$ for $i = 0, \dots, n$ and $j = 0, \dots, c_i$ a (efficiently computable) function φ which approximates f. We require the following properties:

Interpolation property at points t_i, φ and f correspond.

$$\varphi^{(j)}(t_i) = f^{(j)}(t_i) \qquad \forall i, j$$

We also call $f^{(j)}(t_i)$ *supporting points of the function*

Approximation property in a (yet unknown) function space it holds that $\|\varphi - f\| \approx 0$

4.1 Polynomial interpolation

First, we assume that only function values $f_i := f(t_i)$ for $t = 0, \dots, n$ at pairwise different points $t_0, \dots, t_n$ are given.

Goal: Find a polynomial $P \in \mathcal{P}_n$ of degree $\deg(P) \leq n$.

$$P(t) = a_n t^n + a_{n-1} t^{n-1} + \dots + a_n t + a_0$$

with $a_0, \dots, a_n \in \mathbb{R}$ such that f is interpolated at $t_0, \dots, t_n$, thus $P(t_i) = f_i$ for $i = 0, \dots, n$.

Uniqueness and condition

The values $a_0, \dots, a_n$ are uniquely determined by $P(t_i) = f_i$ for $i = 0, \dots, n$, because if $P, Q \in \mathcal{P}_n$ satisfy $P(t_i) = f_i$ then,

$$P(t_i) = Q(t_i) \text{ for } i = 0, \dots, n$$

and thus $R := P - Q \in \mathbb{P}_n$ for some polynomial with $n + 1$ zeros. Thus R is a zero polynomial, thus $P = Q$.

Now consider the map

$$\mathcal{P}_n \to \mathbb{R}^{n+1}, p \mapsto (p(t_0), \dots, p(t_n))$$

The map is linear and by property $\dim(\mathcal{P}_n) = n + 1 = \dim(\mathbb{R}^{n+1})$ with injectivity, surjectivity follows (e.g. by the rank theorem).

Index