Numerical methods for physicists (MSc)

Róbert Horváth

Budapest University of Technology and Economics
Faculty of Natural Sciences
Institute of Mathematics
Department of Analysis

autumn, 2018

Introduction

Course description

Course description

- Contact: e-mail: rhorvath@math.bme.hu, Office: H.24/b
- Course webpage: anal.math.bme.hu/nummeth
- Consultations: office hours: Thursdays 16-17, or by appointment via e-mail
- Course requirements: see the course webpage.
- Lecture notes:

  - slides of the lecture

  - assignments for homework

  - Books:

  Steven C. Chapra, Applied Numerical Methods with MATLAB - for engineers and scientists, McGraw Hill, 2008

  W. Cheney, D. Kincaid, Numerical Mathematics and Computing, Brooks/Cole, Cangage learning, 2013

  - Catch up with Matlab:

  `https://www.mathworks.com/moler/chapters.html`

  `https://web.stanford.edu/class/ee254/software/using_ml.pdf`

  - other readings in Hungarian (listed in the Course requirements)

Introduction to numerical analysis

Introduction

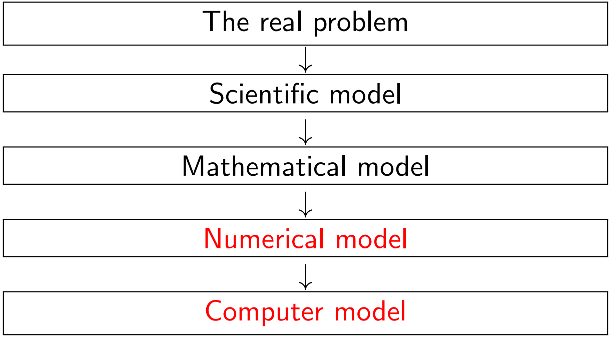
"Numerical analysis is the study of algorithms for the problems of continuous mathematics." (Lloyd N. Trefethen, 1992)

It constructs algorithms and analyses them from the point of view of accuracy, efficiency and its behavior during computer realization.

Problems of continuous mathematics come from different disciplines. They are the mathematical models of e.g. physical, biological, chemical or economical problems.

Introduction

Model construction:



The real problem

Scientific model

Mathematical model

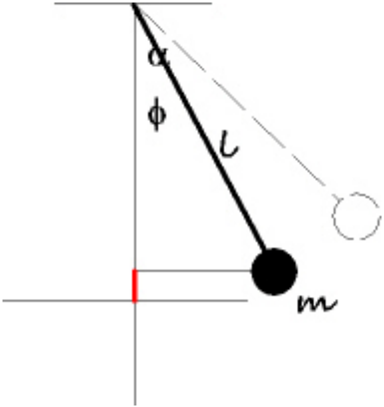
Numerical model

Computer model

Example of the pendulum motion

Problem: Compute the period of a pendulum.

Sci. mod.: Neglect the weight of the string and the drag. Apply the energy conservation principle: $\frac{1}{2}ml^2(\phi'(t))^2 + mgl(1 - \cos\phi(t)) = mgl(1 - \cos\alpha)$.



Math. mod.: The differential equation for the angular velocity:

$$\phi'(t) = \pm\sqrt{\frac{2g}{l}}\sqrt{\cos\phi(t) - \cos\alpha}$$

The period must be computed from this equation.

Example of the pendulum motion

$$\int_0^{T/4} \frac{\phi'(t)}{-\sqrt{\frac{2g}{l}}\sqrt{\cos\phi(t)-\cos\alpha}}\,\mathrm{d}t = T/4.$$

Changing the variable:

$$T = 2\sqrt{2}\sqrt{\frac{l}{g}}\int_0^{\alpha}\frac{1}{\sqrt{\cos\phi-\cos\alpha}}\,\mathrm{d}\phi$$

$$= 4\sqrt{\frac{l}{g}}\int_0^{\pi/2}\frac{1}{\sqrt{1-\sin^2(\alpha/2)\sin^2\vartheta}}\,\mathrm{d}\vartheta.$$

The value of the integral cannot be given in closed form ($\sin\vartheta = \sin(\phi/2)/\sin(\alpha/2)$).

Num. mod.: Let us use numerical integration formulas (see later).

Comp. mod.: $l = 1m$, $g = 9.8m/s^2$
$T = 2.008035541s$ $(\alpha = 5°)$, $T = 2.369049722s$ $(\alpha = 90°)$.

Example of the pendulum motion

Other approach: Let us develop the Taylor series of the function $1/\sqrt{1-x}$ about $x = 0$, and let us apply the series at the point $\sin^2(\alpha/2)\sin^2\vartheta$, then let us integrate the formula:

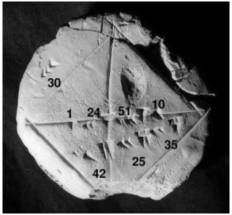
$$T = 2\pi\sqrt{\frac{l}{g}}\left(1 + \frac{1}{4}\sin^2\frac{\alpha}{2} + \dots\right).$$

If we suppose that the initial angular displacement is small, then we obtain the period formula

$$T \approx 2\pi\sqrt{\frac{l}{g}}.$$

This is independent of α. In the example we obtain $T = 2.007089923s$.

History of numerical analysis



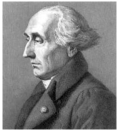
Babylonian stone plate, 1800-1600 B.C.

- Approximation methods were already used in the ancient time (approximation of irrational numbers, interpolation, approximate solution of equations, etc.).
- Some important names:

Newton (approximate solutions of equations, numerical integration, end of 1600s)

History of numerical analysis

Euler (solution of ODEs, 1700s)

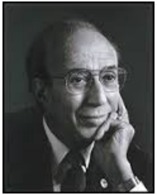


Lagrange (interpolation methods, 1700s)

Gauss (numerical integration, solution of SLAEs, 1800s)

Introduction

- Construction of approximate tables.
- Computer (from the middle of the 20th century)
- 1947: The advent of the modern numerical mathematics. Rounding errors and other scientific considerations.
  John von Neumann, Herman Goldstine, Numerical Inverting of Matrices of High Order, Bulletin of the AMS, Nov. 1947.



J. Neumann, 1903-1957, Hungarian
H. Goldstine, 1913-2004, USA

Basic concepts of numerical analysis

Properly posed problems

Properly posed problems

$$F(x,d)=0 \tag{1}$$

Problem: It is given the data d. Find the solution x. d and x are the elements of some normed spaces and F is an arbitrary function.

Def. 1. (Hadamard, 1902) A problem is a properly posed problem if $\exists\eta>0$ such that the problem has a unique solution ($x_{d+\delta d}$) for all $d+\delta d$ with the property $\|\delta d\|\leq\eta$, and $\exists K(\eta,d)>0$ such that $\|x_{d+\delta d}-x_d\|\leq K(\eta,d)\|\delta d\|$ (the solution depends continuously on the data d).

Example. A simple problem that is not properly posed:
$x-|\{a\in\mathbb{R}\,|\,a^2+a+d/4=0\}|=0$, if $d=1$.

If $\delta d<0$ then we have $x=2$,
if $\delta d=0$ then we have $x=1$, and
if $\delta d>0$ then we have $x=0$.

Properly posed problems

Example. A properly posed problem: Consider the system of linear equations with the data d and with the solution $x = [x_1, x_2]^T$

$$\begin{aligned} dx_1 + x_2 &= 1 \\ x_1 + x_2 &= 0. \end{aligned}$$

Let us choose $d = 0$. In this case the solution is $x_0 = [-1, 1]^T$.

From the general solution of the system we have

$$x_{0+\delta d} = \left[\frac{1}{\delta d - 1}, \frac{-1}{\delta d - 1}\right]^T.$$

The solution is e.g. for $|\delta d| \leq 1/2 =: \eta$. Moreover in this case we have

$$\|x_{0+\delta d} - x_0\| \leq \sqrt{2}\left|\frac{1}{\delta d - 1} - \frac{1}{0 - 1}\right| = \sqrt{2}\left|\frac{\delta d}{\delta d - 1}\right| = \sqrt{2}\left|\frac{\delta d - 0}{\delta d - 1}\right| \leq 2\sqrt{2}\,|\delta d - 0|$$

thus $K(\eta, d) = 2\sqrt{2}$ is a good choice (in Euclidean distance).

Properly posed problems

Example. Let us consider the system below, where d is the data and the solution is $x = [x_1, x_2]^T$. Let $d = 5$, then the solution is $x = [331.7, 5]^T$. It can be checked easily that the problem is properly posed.

$$\begin{aligned} dx_1 - 331x_2 &= 3.5 \\ 6x_1 - 397x_2 &= 5.2 \end{aligned} \tag{2}$$

Let us change the value of d to 4.9 (2%):

$$\begin{aligned} 4.9x_1 - 331x_2 &= 3.5 \\ 6x_1 - 397x_2 &= 5.2. \end{aligned}$$

In this case the solution is $x = [8.1499, 0.1101]^T$ (98%).

The problem is properly posed but, because $K(\eta, d)$ is large, the solution can change very much.

Conditioning of a problem, condition number

Condition numbers, conditioning

Let D be the set of allowable perturbations δd in the problem (1).

Def. 2. The number

$$\kappa(d) = \lim_{\|\delta d\| \to 0} \sup_{\delta d \in D} \frac{\|x_{d+\delta d} - x_d\| / \|x_d\|}{\|\delta d\| / \|d\|}$$

is called the relative condition number of the problem.

Def. 3. The number

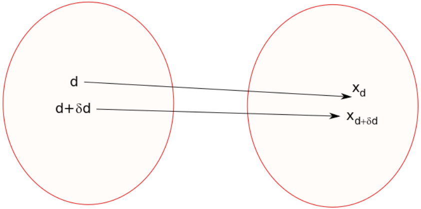
$$\kappa_{abs}(d) = \lim_{\|\delta d\| \to 0} \sup_{\delta d \in D} \frac{\|x_{d+\delta d} - x_d\|}{\|\delta d\|}$$

is called the absolute condition number of the problem.

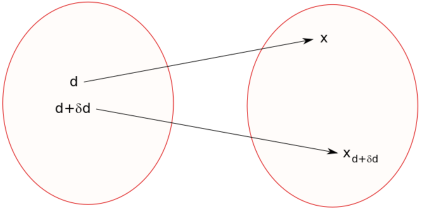
A problem is well-conditioned if the condition number κ is relatively small, and it is badly or ill-conditioned if κ is large.

Condition numbers, conditioning

Well-conditioned problem:



Ill-conditioned problem:



Condition numbers, conditioning

If a problem is properly posed, then the unique solution can be written with the so-called solution function G in the form $x = G(d)$.

If G is differentiable then

$$\kappa(d) = \frac{\|G'(d)\| \cdot \|d\|}{\|G(d)\|}$$

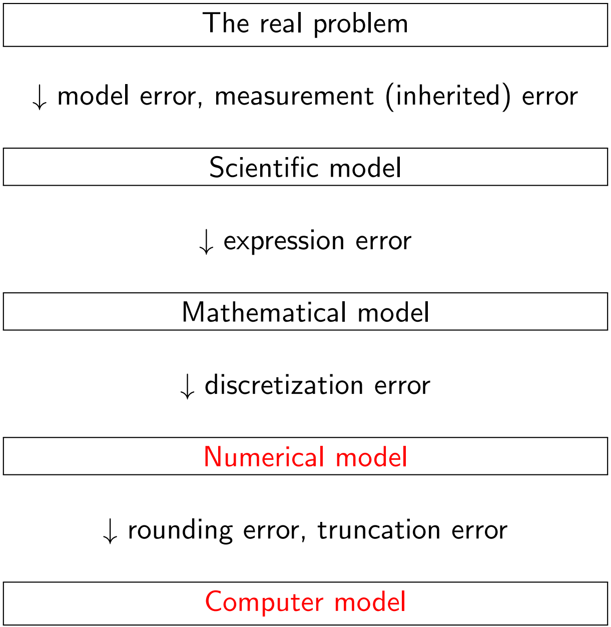
and

$$\kappa_{abs}(d) = \|G'(d)\|.$$

Example. In the example with the system of linear equations (2) $\kappa(5) \approx 1985$ (calculated in 2-norm).

Possible error sources

Possible error sources



The real problem

↓ model error, measurement (inherited) error

Scientific model

↓ expression error

Mathematical model

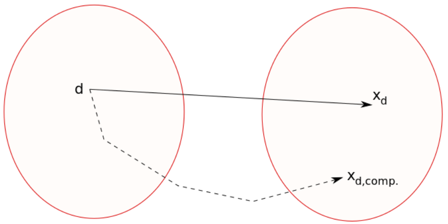
↓ discretization error

Numerical model

↓ rounding error, truncation error

Computer model

Possible error sources

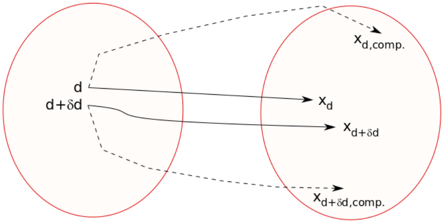


The solution of ill-conditioned problems is very risky.

If the problem is ill-conditioned, then no amount of effort, trickery, or talent used in the computation can produce accurate answers except by chance. (John R. Rice, Matrix Computations and Mathematical Software, McGraw-Hill, 1981)

Possible error sources

We can consider also the conditioning of a computation. Moreover, the computation process can be also ill-conditioned (unstable). We have to avoid these type of methods, too.



Possible error sources

Example. Let us solve the problem on page 17 with Matlab with $d = 0$ and $d = 10^{-16}$ using the Gaussian method (without pivoting).

With $d = 0$, we obtain the exact solution $x = [-1, 1]^T$, but with $d = 10^{-16}$ we obtain the solution $[0, 1]^T$, which is very far from the exact solution

$$x = \left[\frac{1}{10^{-16} - 1}, \frac{-1}{10^{-16} - 1}\right]^T.$$

Rmk.
The cardinal sin of a numerical software is to produce ill-conditioned computations for a well-conditioned problem.

It is highly desirable for a numerical software to recognize that its calculations are ill-conditioned and to report this fact to the user.

Measuring the error with norms

Vector, matrix and function norms

It is highly recommended here to review the summary section about normed spaces → page 214.

If x, y are two elements in a normed space V, then their distance can be measured with the number $\|x - y\|$.

In $\mathbb{R}^n$ we use the following vector norms ($\overline{\mathbf{x}} = [x_1, \ldots, x_n]^T$):

- $\|\overline{\mathbf{x}}\|_1 = |x_1| + \cdots + |x_n|$ (octahedron norm),
- $\|\overline{\mathbf{x}}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$ (Euclidean norm),
- $\|\overline{\mathbf{x}}\|_\infty = \max\{|x_1|, \ldots, |x_n|\}$ (maximum norm, $p \to \infty$).

Norms on $\mathbb{R}^{n \times n}$ are called matrix norms. (For the special properties of matrices see the summary section → page 227) Matrix norms can be defined from vector norms with the expression

$$\|\mathbf{A}\| := \sup_{\overline{\mathbf{x}} \neq \overline{\mathbf{o}}} \frac{\|\mathbf{A}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|}. \tag{3}$$

This is the so-called induced matrix norm

Vector, matrix and function norms

Thm. 4. Suppose that the matrix norm $\|.\|$ was induced by the vector norm $\|.\|$. Then

- $\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{x}\|, \forall \overline{\mathbf{x}} \in \mathbb{R}^n$ (consistency),
- $\|\mathbf{I}\| = 1$ ($\mathbf{I}$ is the identity matrix),
- $\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{B}\|$ (submultiplicity).

Proof. It follows directly from the definition of an induced matrix norm. ■

Thm. 5. The vector norms induce the following matrix norms:

- $p = 1$: $\|\mathbf{A}\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$,
- $p = \infty$: $\|\mathbf{A}\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$,
- $p = 2$: $\|\mathbf{A}\|_2 = \sqrt{\varrho(\mathbf{A}^T\mathbf{A})}$ (ϱ: spectral radius).

Proof. The first two are left as exercises. The case $p = 2$ can be proven as follows. The matrix $\mathbf{A}^T\mathbf{A}$ is symmetric and positive semidefinite, moreover $\mathbf{A}^T\mathbf{A}$ can be written in the form $\mathbf{A}^T\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ (diagonalizable with an orthogonal matrix).

Vector, matrix and function norms

Thus

$$\frac{\|\mathbf{A}\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} = \frac{\overline{\mathbf{x}}^T\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|_2^2} = \frac{\overline{\mathbf{x}}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\overline{\mathbf{x}}}{\|\overline{\mathbf{x}}\|_2^2}$$

$$= \frac{\|\sqrt{\mathbf{\Lambda}}\mathbf{V}^T\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} \le \frac{\varrho(\mathbf{A}^T\mathbf{A})\|\overline{\mathbf{x}}\|_2^2}{\|\overline{\mathbf{x}}\|_2^2} = \varrho(\mathbf{A}^T\mathbf{A}).$$

We get equality in the case if we choose the vector $\overline{\mathbf{x}}$ to be the eigenvector that belongs to the eigenvalue of $\mathbf{A}^T\mathbf{A}$ with the greatest absolute value. Thus the proof is complete. ■

Rmk. In the case of symmetric matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$, we have $\|\mathbf{A}\|_2 = \varrho(\mathbf{A})$.

Rmk. The matrix norm $\|\mathbf{A}\| = \max_{i,j}\{|a_{ij}|\}$ is not an induced norm. The so-called Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ is not an induced norm, too.

The space of the continuous functions defined on $[a,b]$ is denoted with $C[a,b]$. The usual norm of this space, the maximum norm, is defined as follows

$$\|f\|_{C[a,b]} = \max_{x\in[a,b]}\{|f(x)|\}.$$

Norms and eigenvalues

Thm. 6. For quadratic matrices, the estimation $\varrho(\mathbf{A}) \leq \|\mathbf{A}\|$ is satisfied in any induced norm.

Proof.: Let $\overline{\mathbf{x}} \neq \mathbf{0}$ be an eigenvector of $\mathbf{A}$ and λ be the corresponding eigenvaluue. Then $|\lambda| \cdot \|\overline{\mathbf{x}}\| = \|\lambda\overline{\mathbf{x}}\| = \|\mathbf{A}\overline{\mathbf{x}}\| \leq \|\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|$. ■

Thm. 7. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a given matrix. Then for any positive $\varepsilon > 0$, there exists an induced norm $\|.\|$, such that $\|\mathbf{A}\| \leq \varrho(\mathbf{A}) + \varepsilon$.

Thm. 8. Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a given matrix. $\mathbf{A}^k$ tends to $\mathbf{0}$ elementwise if and only if $\varrho(\mathbf{A}) < 1$. Exactly in the same case, the series

$$\sum_{k=0}^{\infty} \mathbf{A}^k$$

converges, moreover its sum is $(\mathbf{I} - \mathbf{A})^{-1}$.

Proof: $\Leftarrow$ Because $\varrho(\mathbf{A}) < 1$, there exists an induced matrix norm such that $\|\mathbf{A}\| < 1$. Thus $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k \to 0$ if $k \to \infty$. Because of the equivalence of the norms, the matrix $\mathbf{A}^k$ tends to zero elementwise.

Norms and eigenvalues

$\Rightarrow$ Let $\overline{\mathbf{v}}$ be an eigenvector of the matrix with the eigenvalue λ. Then $\mathbf{A}^k\overline{\mathbf{v}} = \lambda^k\overline{\mathbf{v}}$. Because $\mathbf{A}^k$ tends to the zero matrix, the vector $\mathbf{A}^k\overline{\mathbf{v}}$ must tend to the zero vector. This can happen only if $|\lambda| < 1$. This implies the condition $\varrho(\mathbf{A}) < 1$.
Let us consider the following identity:

$$(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^l) = \mathbf{I} - \mathbf{A}^{l+1}.$$

$\mathbf{I} - \mathbf{A}$ is regular because its eigenvalues cannot be zero. In this way

$$\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^l = (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} - \mathbf{A}^{l+1}).$$

The series converges only if $\varrho(\mathbf{A}) < 1$ and then its sum is $(\mathbf{I} - \mathbf{A})^{-1}$ indeed. ■

The result of the theorem will be used in the proof of the following two important theorems.

Norms and eigenvalues

Thm. 9. If the relation $\|\mathbf{A}\| < 1$ is valid for the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ in some induced norm, then the following estimation holds

$$\frac{1}{1+\|\mathbf{A}\|} \le \|(\mathbf{I}-\mathbf{A})^{-1}\| \le \frac{1}{1-\|\mathbf{A}\|}.$$

Proof: It follows from the previous theorem that the matrix $\mathbf{I}-\mathbf{A}$ is non-singular.

$$\mathbf{I} = (\mathbf{I}-\mathbf{A})(\mathbf{I}-\mathbf{A})^{-1} \Rightarrow 1 \le \|\mathbf{I}-\mathbf{A}\|\|(\mathbf{I}-\mathbf{A})^{-1}\|$$

$$\le (1+\|\mathbf{A}\|)\|(\mathbf{I}-\mathbf{A})^{-1}\| \Rightarrow \text{estimation on the left hand side.}$$

Let us multiply both sides of the equality $\mathbf{I} = \mathbf{I}-\mathbf{A}+\mathbf{A}$ with the inverse of $\mathbf{I}-\mathbf{A}$, then take the norms on both sides.

$$\|(\mathbf{I}-\mathbf{A})^{-1}\| \le 1 + \|(\mathbf{I}-\mathbf{A})^{-1}\|\,\|\mathbf{A}\|,$$

and after reordering we obtain the inequality on the right hand side. ■

Speed of convergence

Speed of convergence

In iterative methods, the solution is the limit of a specially constructed sequence. Nonlinear equations cannot be solve with direct methods in general. In this case we use iterative methods, that is we generate a sequence that is convergent and its limit is the solution of the equation.

Let us consider the sequence $x_k \to x^\star$. Let $e_k = x_k - x^\star$ be the error of the kth element.

Def. 10. We say that the order of the convergence of the sequence $\{x_k\}$ is the positive real number p if the limit

$$\lim_{k\to\infty} \frac{\|e_{k+1}\|}{\|e_k\|^p} = C \neq 0$$

exists, it is finite and non-zero.

Rmk. If the order of convergence can be defined for a sequence, then it is unique.

Speed of convergence

Rmk. If $p = 1$, then the convergence is linear. If $1 < p < 2$, then the convergence is superlinear. The case $p = 2$ means second order of convergence.

Rmk. If we have a sequence with convergence order p, then for large k values we have the approximation

$$\|e_{k+1}\| \approx C\|e_k\|^p.$$

The logarithm of the equation is

$$\log\|e_{k+1}\| \approx \log C + p\log\|e_k\|.$$

If we graph $\log\|e_{k+1}\|$ against $\log\|e_k\|$, the points falls on a line with slope p that intersects the vertical axis at $\log C$.

This method can be used to check the order of convergence of a sequence (or a method that produces the sequence) empirically.

Example. Both $x_{k+1} = x_k - (2/5)(x_k^2 - 2)$ and $y_{k+1} = y_k - (y_k^2 - 2)/2/y_k$ ($x_0 = y_0 = 3$) tend to $\sqrt{2}$. The first one is order 1 and the second one is order 2.

Speed of convergence

Def. 11. We say that the approximation $\tilde{x}$ of the real number $x^\star$ has h correct digits, if

$$|\tilde{x} - x^\star| \leq \frac{1}{2} 10^{m-h+1} = 5 \times 10^{m-h},$$

where 10^m is the place value of the first significant digit of the number.

Example. Let $x^\star = \pi$ and $\tilde{x} = 3.140$. Then $\tilde{x}$ has 3 correct digits. Indeed $|x^\star - \tilde{x}| = 0.001592654 \leq 10^{0-3+1}/2 = 0.005$. The fourth digit is not correct. In the opposite case we need $0.001592654 \leq 0.0005$, which is not valid.

Example. Let $x^\star = \pi$ and $\tilde{x} = 3.142$. Then $\tilde{x}$ has 4 correct digit because $|x^\star - \tilde{x}| = 0.000407346 \leq 10^{0-4+1}/2 = 0.0005$.

Speed of convergence

Rmk. Let us consider a linearly convergent real sequence. Then the number of correct digits compared to the limit of the sequence increases with $-\lg C$ in each step. Indeed, if the kth element has h correct digits, then for the $(k+1)$th element we have

$$|e^{(k+1)}| \approx C|e^{(k)}| \leq C10^{m-h+1}/2 = 10^{m-(h-\lg C)+1}/2.$$

If $C = 1/2$, then $-\lg C \approx 0.3$, thus the number of correct digits increases with 1 in each three steps.

Rmk. Let us consider a quadratically convergent real sequence. Then the number of correct digits compared to the limit of the sequence increases with $(h + \lg 2 - \lg C - 1 - m)$ in each step. Indeed, if the kth element has h correct digits, then for the $(k+1)$th element we have

$$|e^{(k+1)}| \approx C|e^{(k)}|^2 \leq C(10^{m-h+1}/2)^2$$

$$= 10^{m-(h+(h+\lg 2-\lg C-1-m))+1}/2.$$

Thus the number of correct digits doubles in each step.

Machine number format and its corollaries

Some simple examples

MATLAB results:

- $\tan(\pi/2) = 1.6331e + 016$
- $2^{-1074}/2 = 0$
- $2^{-1074} = 4.94066e - 324$; $2^{-1074} \cdot 1.2 = 4.94066e - 324$
- 10^{310} =Inf
- Let y_k denote the semiperimeter of a regular polygon with 2^k edges inscribed into a circle with radius 1. Then $y_k \to \pi$, if $k \to \infty$. Moreover we have the recursion

$$y_{k+1} = 2^{k+1}\sqrt{\frac{1}{2}\left(1 - \sqrt{1 - (2^{-k}y_k)^2}\right)},$$

where $y_1 = 2$, $y_2 = 2\sqrt{2}$, ... , $y_{10} = 3.14158627$, $y_{12} = 3.14166137$, ..., $y_{19} = 3.70727600$, ... Does not tend to π!

Some simple examples

MATLAB results:

- Calculate the following expression in different ways!

$$y = 333.75b^6 + a^2(11a^2b^2 - b^6 - 121b^4 - 2) + 5.5b^8 + \frac{a}{2b},$$

with $a = 77617$ és $b = 33096$.
- Matlab double precision: $y = -1.1806e + 21$
- Matlab double precision without exponents ($a^2 = a * a$, etc.): $y = 1.1726$
- Matlab single precision: $y = -6.3383e + 29$
- Matlab single precision without exponents ($a^2 = a * a$, etc.): $y = 6.3383e + 29$
- Correct answer:

$$\begin{aligned} z &= 333.75b^6 + a^2(11a^2b^2 - b^6 - 121b^4 - 2) \\ &= -7917111340668961361101134701524942850 \end{aligned}$$

$$x = 5.5b^8 = 7917111340668961361101134701524942848$$

$$y = z + x + \frac{a}{2b} = -2 + \frac{77617}{2 \cdot 33096} = -0.827396059946821$$

Representation of real numbers in floating point systems

(Konrad Zuse, Berlin, 1930s)

$$\pm b^k \left(\frac{a_0}{b^0} + \frac{a_1}{b^1} + \frac{a_2}{b^2} + \cdots + \frac{a_{p-1}}{b^{p-1}} \right) \equiv a_0.a_1a_2 \ldots a_{p-1} \times b^k$$

- b: base of the representation
- p: the number of the digits in the mantissa
- k: exponent or characteristic
- $0 \le a_i < b$ integers, $(i = 0, \ldots, p-1)$
- If $a_0 \neq 0$ then the number is in normal form. This is a unique representation.

Illustrative example

`http://www.binaryconvert.com/result_double.html?decimal=048046049`

Representation of real numbers in floating point systems

In the floating point number system we have:

- Only finite number of rational numbers.
- The numbers do not form a field (e.g. the addition is not associative). (Ex.: $123.4 + 0.04 + 0.03 + 0.02 + 0.01$ in different orders in the case $p = 4$, $b = 10$, $k_{max} = 2$)
- The numbers form a bounded set. In the previous example, the largest number is 999.9 (overflow)
- Around zero, there is a relatively large space. The smallest positive representable number in normal form is 0.01. Without the normal form restriction: 0.00001 (underflow).
- The smallest number that is larger then 1 is denoted by $1 + \varepsilon_m$, where ε_m is the so-called machine epsilon. In the example: 0.001.

Double precision floating point numbers

64 bits, binary number system

- The 1. bit sores the sign of the number $(0 = +, 1 = -)$.
- The bits 2-12. store the characteristic such that we add 1023 to the exponent and we store the binary version of that number (from -1022 to 1023). The characteristic -1023 stores the 0 (if the mantissa is zero) or indicates that number is not in normal form $(0.a_1 \dots a_{52} \times 2^{-1022})$. The characteristic coded with all 1s is used for special purposes (mantissa is not zero - NaN, mantissa is zero - $\pm$Inf (depending on the sign bit)).
- The bits 13-64. store the mantissa (the part after the binary point).

Double precision floating point numbers

The largest exactly representable positive number

$$M = 1.\underbrace{111\dots111}_{\text{52db}}\times 2^{1023} = 1.79769 \times 10^{308}$$

and the smallest positive exactly representable number

$$m = 0.\underbrace{000\dots000}_{\text{51db}}1 \times 2^{-1022} = 4.94066 \times 10^{-324}.$$

The smallest positive exactly representable number in normal form

$$\varepsilon_0 = 1.\underbrace{000\dots000}_{\text{52db}}\times 2^{-1022} = 2.22507 \times 10^{-308}.$$

The smallest exactly representable number next to 1

$$1.\underbrace{000\dots000}_{\text{51db}}1 \times 2^{0},$$

which is greater than 1 with $\varepsilon_g = 2^{-52} = 2.22 \times 10^{-16}$

Rounding to floating points

Thm. 12. Let $0 < x \leq M$. Then

$$|fl(x) - x| \leq \begin{cases} m/2, & \text{if } x < m/2, \\ \frac{\varepsilon_g |x|}{2}, & \text{if } m/2 \leq x \leq M. \end{cases}$$

Proof: The first part is trivial. Let us suppose that x is between the floating point numbers x_i and x_j. Let the number of the digits of the mantissa of x_i be p and the characteristic k. Then

$$|fl(x) - x| \leq \frac{x_j - x_i}{2} = \frac{b^{-p+1} b^k}{2} \leq \frac{\varepsilon_g |x|}{2}. \blacksquare$$

The relative error if $m/2 \leq x \leq M$ is

$$\frac{|fl(x) - x|}{|x|} \leq \frac{\varepsilon_g}{2} =: \mathtt{u} \text{ machine precision.}$$

Error of floating point operations

$$x \boxed{\diamond} y := fl(fl(x) \diamond fl(y))$$

The relative error of the subtraction $(x, y > 0)$.

$$\begin{aligned}
&\frac{|x \boxed{\text{-}} y - (x-y)|}{|x-y|} \\
&= \frac{|(x(1+\delta_x) - y(1+\delta_y))(1+\delta_-) - (x-y)|}{|x-y|} \\
&\leq \frac{|(x\delta_x - y\delta_y)(1+\delta_-)|}{|x-y|} + |\delta_-| \leq \mathtt{u}(1+\mathtt{u})\frac{x+y}{|x-y|} + \mathtt{u},
\end{aligned}$$

where $|\delta_x|, |\delta_y|, |\delta_-| \leq \mathtt{u}$.

If $x \approx y$, then the relative error of the subtraction can be much larger than the machine precision (or than the relative error of x or y).

Catastrophic cancellation

This happens by the subtraction of two numbers that are close to each other:

Example. The case of the sequence that should tend to π. The problem can be eliminated with the following reformulation of the iteration:

$$y_{k+1} = y_k \sqrt{\frac{2}{1+\sqrt{1-(2^{-k}y_k)^2}}}.$$

Example.

$$\sqrt{9876} = 9.937806599 \times 10^1, \qquad \sqrt{9875} = 9.937303457 \times 10^1, \quad \text{error} = 10^{-8}\%$$

$$\downarrow$$

$$\sqrt{9876}-\sqrt{9875} = 0.000503142 \times 10^1 = 5.03142 \quad \underbrace{\mathbf{0000}}_{\text{no information}} \quad \times 10^{-3}$$

error$=10^{-4}\%$

Catastrophic cancellation

Better solution:

$$\sqrt{9876} - \sqrt{9875} = \frac{1}{\sqrt{9876} + \sqrt{9875}}$$

$$= 0.005031418679 = 5.031418679 \times 10^{-3}$$

Catastrophic cancellation can occur in those cases when the result is much smaller than the absolute values of the terms summed up.

Example.

$$e^x = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{x^i}{i!}$$

Let $x = -25$. Then $e^{-25} \approx 1.388794 \times 10^{-11}$. The limit of the above sequence according to Matlab is 8.086559×10^{-7}.

Operation count

If floating point operations are the dominant cost then the computation time is proportional to the number of mathematical operations. This is measured in $flop$s. 1 $flop$ is one floating point operation $(-,+,*,/)$.

Def. 13. We say that the sequence $\{a_n\}$ is of order $O(n^\alpha)$ $(\alpha > 0)$ $(n \to \infty)$, if there are constants $n_0 > 0$ and $K > 0$ such that $|a_n| \leq Kn^\alpha$ if $n \geq n_0$. Notation: $a_n = O(n^\alpha)$.

Introduction to the solution of systems of linear algebraic equations

Systems of linear algebraic equations

Systems of linear algebraic equations (SLAEs)

- General form ($a_{ij},\ b_i$ are known, find the values x_j)

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

- Vector form

$$x_1\overline{\mathbf{a}}_1 + \cdots + x_n\overline{\mathbf{a}}_n = \overline{\mathbf{b}}$$

- Matrix form

$$\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$$

Thm. 14. A SLAE is solvable iff $r(\mathbf{A}) = r(\mathbf{A}|\overline{\mathbf{b}})$. If it is solvable and $r(\mathbf{A}) < n$, then it has infinitely many solutions, if $r(\mathbf{A}) = n$, then the solution is unique.

Sensibility of the solution

The relative error of the solution

Thm. 15. Let us suppose that, instead of the system $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$, we solve the system $(\mathbf{A} + \boldsymbol{\delta}\mathbf{A})\overline{\mathbf{y}} = \overline{\mathbf{b}} + \boldsymbol{\delta}\overline{\mathbf{b}}$. The solution is written in the form $\overline{\mathbf{y}} = \overline{\mathbf{x}} + \boldsymbol{\delta}\overline{\mathbf{x}}$. Moreover, let us suppose that $\|\boldsymbol{\delta}\mathbf{A}\| < 1/\|\mathbf{A}^{-1}\|$ in some induced norm. Then the following estimation is true

$$\frac{\|\boldsymbol{\delta}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} \leq \frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\|\boldsymbol{\delta}\mathbf{A}\|/\|\mathbf{A}\|} \cdot \left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\overline{\mathbf{b}}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\|}{\|\mathbf{A}\|} \right)$$

where $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|$.

Proof. Since $\|\boldsymbol{\delta}\mathbf{A}\| < 1/\|\mathbf{A}^{-1}\|$, the estimation $\|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\| < 1$ holds. Thus, in view of the equality $\mathbf{A} + \boldsymbol{\delta}\mathbf{A} = \mathbf{A}(\mathbf{I} - \mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A})$ the matrix $\mathbf{A} + \boldsymbol{\delta}\mathbf{A}$ is regular (Theorem 8.). Moreover,

$$\boldsymbol{\delta}\overline{\mathbf{x}} = (\mathbf{A} + \boldsymbol{\delta}\mathbf{A})^{-1}(\overline{\mathbf{b}} + \boldsymbol{\delta}\overline{\mathbf{b}}) - \overline{\mathbf{x}} = (\mathbf{A} + \boldsymbol{\delta}\mathbf{A})^{-1}(\overline{\mathbf{b}} + \boldsymbol{\delta}\overline{\mathbf{b}} - (\mathbf{A} + \boldsymbol{\delta}\mathbf{A})\overline{\mathbf{x}})$$

$$= (\mathbf{A} + \boldsymbol{\delta}\mathbf{A})^{-1}(\boldsymbol{\delta}\overline{\mathbf{b}} - \boldsymbol{\delta}\mathbf{A}\overline{\mathbf{x}}) = (\mathbf{I} + \mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A})^{-1}\mathbf{A}^{-1}(\boldsymbol{\delta}\overline{\mathbf{b}} - \boldsymbol{\delta}\mathbf{A}\overline{\mathbf{x}}).$$

The relative error of the solution

Let us apply Theorem 9.

$$\|\boldsymbol{\delta}\overline{\mathbf{x}}\| \leq \frac{\|\mathbf{A}^{-1}\|}{1-\|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|}(\|\boldsymbol{\delta}\overline{\mathbf{b}}\| + \|\boldsymbol{\delta}\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|)$$

$$= \frac{\|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\|}{1-\|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|}\left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\mathbf{A}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|}{\|\mathbf{A}\|}\right).$$

We obtain

$$\frac{\|\boldsymbol{\delta}\overline{\mathbf{x}}\|}{\|\overline{\mathbf{x}}\|} \leq \frac{\|\mathbf{A}^{-1}\| \cdot \|\mathbf{A}\|}{1-\|\mathbf{A}^{-1}\boldsymbol{\delta}\mathbf{A}\|}\left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\mathbf{A}\| \cdot \|\overline{\mathbf{x}}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\|}{\|\mathbf{A}\|}\right)$$

$$\leq \frac{\kappa(\mathbf{A})}{1-\kappa(\mathbf{A})\|\boldsymbol{\delta}\mathbf{A}\|/\|\mathbf{A}\|} \cdot \left(\frac{\|\boldsymbol{\delta}\overline{\mathbf{b}}\|}{\|\overline{\mathbf{b}}\|} + \frac{\|\boldsymbol{\delta}\mathbf{A}\|}{\|\mathbf{A}\|}\right). \blacksquare$$

Condition number of matrices

Condition number of matrices

Let us notice that if the coefficients of a SLAE are changed with a small amount, then the solution can change with a relatively large amount if the parameter $\kappa(\mathbf{A})$ is large.

Def. 16. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a regular matrix. Then the number $\kappa(\mathbf{A}) = \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|$ is called the condition number of the matrix. (Its value depends also on the norm!)

The properties of the condition number in induced norm:

- $\kappa(\mathbf{A}) \geq 1$ $(1 = \|\mathbf{I}\| = \|\mathbf{A}\mathbf{A}^{-1}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{A}^{-1}\|)$,
- $\kappa(\mathbf{A}) = \kappa(\mathbf{A}^{-1})$,
- $\kappa(\alpha\mathbf{A}) = \kappa(\mathbf{A})$, $\alpha \neq 0$,
- For orthogonal matrices: $\kappa_2(\mathbf{A}) = 1$ $(\|\mathbf{A}\|_2 = \|\mathbf{A}^{-1}\|_2 = 1)$,
- For symmetric matrices: $\kappa(\mathbf{A}) \geq |\lambda_{\max}/\lambda_{\min}|$, moreover $\kappa_2(\mathbf{A}) = |\lambda_{\max}/\lambda_{\min}|$.

Hilbert matrix

This is an example for a very badly conditioned matrix:

Hilbert matrix: $\mathbf{H}_n \in \mathbb{R}^{n\times n}$, $(\mathbf{H}_n)_{i,j} = 1/(i+j-1)$.

$$\mathbf{H}_6 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & 1/10 \\ 1/6 & 1/7 & 1/8 & 1/9 & 1/10 & 1/11 \end{bmatrix}$$

Example. $\kappa_2(\mathbf{H}_6) \approx 1.6 \times 10^7$, $\kappa_2(\mathbf{H}_{10}) \approx 3.5 \times 10^{13}$.

Solution methods of SLAEs

Solution methods of SLAEs

- Direct methods: They give exact solutions in finitely many steps. (Cramer rule $x_i = \det \mathbf{A}_i / \det \mathbf{A}$ ($\mathbf{A}_i$-t can be obtained by changing the ith column of $\mathbf{A}$ to $\overline{\mathbf{b}}$), $\overline{\mathbf{x}} = \mathbf{A}^{-1}\overline{\mathbf{b}}$, Gaussian method and its variants)
- Iterative methods: they form a vector sequence that tends to the solution of the system (Gauss–Seidel, Jacobi, SOR). Important question is that when to step the iteration process.

Direct methods of SLAEs

Gaussian method

Gaussian method

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

Carl Friedrich Gauss
(1777-1855)

Possible transformations that do not alter the solution:

- Multiplication of one equation with a constant ($\neq 0$).
- Addition of one equation to another one.
- Interchange of two equations.
- Interchange of two unknowns.

Gaussian method

The coefficient matrix and the right hand side of the system:

$$\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & \dots & a_{3n} & b_3 \\ \vdots & & & & \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{array}$$

Gaussian method

The initial matrix of the elimination $[\mathbf{A}^{(1)}|\overline{\mathbf{b}}^{(1)}]$:

$$\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & \dots & a_{3n}^{(1)} & b_3^{(1)} \\ \vdots & & & & \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)} \end{array}$$

Gaussian method

The elimination of the first column:

$$\begin{array}{cccc|c}
a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\
a_{21}^{(1)} & a_{22}^{(1)} & \dots & a_{2n}^{(1)} & b_2^{(1)} \\
a_{31}^{(1)} & a_{32}^{(1)} & \dots & a_{3n}^{(1)} & b_3^{(1)} \\
\vdots & & & & \\
a_{n1}^{(1)} & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} & b_n^{(1)}
\end{array}$$

$l_{21} = a_{21}^{(1)}/a_{11}^{(1)}, \dots, l_{n1} = a_{n1}^{(1)}/a_{11}^{(1)}$

Gaussian method

The elimination of the first column:

$$\begin{array}{cccc|c}
a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\
0 & a_{22}^{(1)} - l_{21}a_{12}^{(1)} & \dots & a_{2n}^{(1)} - l_{21}a_{1n}^{(1)} & b_2^{(1)} - l_{21}b_1 \\
0 & a_{32}^{(1)} - l_{31}a_{12}^{(1)} & \dots & a_{3n}^{(1)} - l_{31}a_{1n}^{(1)} & b_3^{(1)} - l_{31}b_1 \\
\vdots & & & & \\
0 & a_{n2}^{(1)} - l_{n1}a_{12}^{(1)} & \dots & a_{nn}^{(1)} - l_{n1}a_{1n}^{(1)} & b_n^{(1)} - l_{n1}b_1
\end{array}$$

Gaussian method

The elimination of the first column $[\mathbf{A}^{(2)}|\overline{\mathbf{b}}^{(2)}]$:

$$\left.\begin{array}{cccc|c} {\color{blue}a_{11}^{(1)}} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & a_{32}^{(2)} & \dots & a_{3n}^{(2)} & b_3^{(2)} \\ \vdots & & & & \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} & b_n^{(2)} \end{array}\right.$$

Gaussian method

The elimination of the second column:

$$\left.\begin{array}{cccc|c}
a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\
0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\
0 & a_{32}^{(2)} & \dots & a_{3n}^{(2)} & b_3^{(2)} \\
\vdots & & & & \\
0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} & b_n^{(2)}
\end{array}\right.$$

$l_{32} = a_{32}^{(2)}/a_{22}^{(2)}, \dots, l_{n2} = a_{n2}^{(2)}/a_{22}^{(2)}$

Gaussian method

The elimination of the second column:

$$\begin{array}{cccc|c}
a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\
0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\
0 & 0 & \dots & a_{3n}^{(2)} - l_{32}a_{2n}^{(2)} & b_3^{(2)} - l_{32}b_2^{(2)} \\
\vdots & & & & \\
0 & 0 & \dots & a_{nn}^{(2)} - l_{n2}a_{2n}^{(2)} & b_n^{(2)} - l_{n2}b_2^{(2)}
\end{array}$$

Gaussian method

The elimination of the second column $[\mathbf{A}^{(3)}|\overline{\mathbf{b}}^{(3)}]$:

$$\left.\begin{array}{cccc|c}
a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\
0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\
0 & 0 & \dots & a_{3n}^{(3)} & b_3^{(3)} \\
\vdots & & & & \\
0 & 0 & \dots & a_{nn}^{(3)} & b_n^{(3)}
\end{array}\right.$$

Gaussian method

After the elimination of the $(n-1)$st column, we obtain the form $[\mathbf{A}^{(n)}|\overline{\mathbf{b}}^{(n)}]$:

$$\left[\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \dots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & \dots & a_{3n}^{(3)} & b_3^{(3)} \\ \vdots & & & & \\ 0 & 0 & \dots & a_{nn}^{(n)} & b_n^{(n)} \end{array}\right.$$

Gaussian method

Back substitution:

$$\begin{aligned}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\
a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\
&\vdots \\
a_{nn}^{(n)}x_n &= b_n^{(n)}
\end{aligned}$$

Gaussian method

Back substitution:

$$\begin{aligned}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\
a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\
&\vdots \\
a_{nn}^{(n)}x_n &= b_n^{(n)} \\
\rightarrow x_n = b_n^{(n)}/a_{nn}^{(n)}
\end{aligned}$$

Gaussian method

Back substitution:

$$\begin{aligned}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n &= b_1^{(1)} \\
a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n &= b_2^{(2)} \\
\rightarrow x_2 = (b_2^{(2)} - x_n a_{2n}^{(2)} - \cdots - x_3 a_{23}^{(2)})/a_{22}^{(2)} & \\
\vdots & \\
a_{nn}^{(n)}x_n &= b_n^{(n)} \\
\rightarrow x_n = b_n^{(n)}/a_{nn}^{(n)} &
\end{aligned}$$

Gaussian method

Back substitution:

$$\begin{array}{rcl}
a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n & = & b_1^{(1)} \\
\to x_1 = (b_1^{(1)} - x_n a_{1n}^{(1)} - \cdots - x_2 a_{12}^{(1)})/a_{11}^{(1)} & & \\
a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n & = & b_2^{(2)} \\
\to x_2 = (b_2^{(2)} - x_n a_{2n}^{(2)} - \cdots - x_3 a_{23}^{(2)})/a_{22}^{(2)} & & \\
\vdots & & \\
a_{nn}^{(n)}x_n & = & b_n^{(n)} \\
\to x_n = b_n^{(n)}/a_{nn}^{(n)} & &
\end{array}$$

Gaussian method

The procedure can be carried out in the present form only if the constants $a_{11}^{(1)}, \ldots, a_{nn}^{(n)}$, the so-called pivot elements are not zeros.

The to phase of the algorithm:

- Elimination process
- Back substitution (solution of a SLAE with a triangular coefficient matrix)

Example. Solve the SLAE.

$$\begin{aligned} x_1 + 1/2x_2 + 1/3x_3 &= 11/6 \\ 1/2x_1 + 1/3x_2 + 1/4x_3 &= 13/12 \\ 1/3x_1 + 1/4x_2 + 1/5x_3 &= 47/60 \end{aligned}$$

Solution: $x_1 = x_2 = x_3 = 1$.

Investigation of the Gaussian method

The algorithm of the Gaussian method

Gaussian method, SLAE given with the matrix $[\mathbf{A}|\overline{\mathbf{b}}] = [\bar{a}_{ij}]_{n\times(n+1)}$.

```
for k:=1:n-1 do
   for i:=k+1:n do
      $l_{ik} := \bar{a}_{ik}/\bar{a}_{kk}$
      for j:=k+1:n+1 do
         $\bar{a}_{ij} := \bar{a}_{ij} - l_{ik} \cdot \bar{a}_{kj}$
      end for
   end for
end for
$x_n := \bar{a}_{n,n+1}/\bar{a}_{nn}$
for k:=n-1:-1:1 do
   $x_k := \bar{a}_{k,n+1}$
   for j:=k+1:n do
      $x_k := x_k - \bar{a}_{kj} \cdot x_j$
   end for
   $x_k := x_k/\bar{a}_{kk}$
end for
```

Gauss transformation

Let $\bar{\mathbf{l}}_k = [0, \ldots, 0, l_{k+1,k}, \ldots, l_{n,k}]^T \in \mathbb{R}^n$ $(k = 1, \ldots, n-1)$. Then the kth step of the Gaussian elimination can be written as the matrix multiplication from left with the matrix $\mathbf{L}_k := \mathbf{I} - \bar{\mathbf{l}}_k \overline{\mathbf{e}}_k^T$.

It is easy to see that $(\mathbf{I} - \bar{\mathbf{l}}_k \overline{\mathbf{e}}_k^T)^{-1} = \mathbf{I} + \bar{\mathbf{l}}_k \overline{\mathbf{e}}_k^T$.

The performability of the Gaussian method

Thm. 17. The Gaussian method can be performed with the previous algorithm iff all its principal minors of $\mathbf{A}$ are non-zero, that is $\det(\mathbf{A}(1:k,1:k)) \neq 0$ $(k=1,\ldots,n)$.

Proof: During the Gaussian elimination process we add some rows of the matrix to other rows. This procedure does not modify the determinant of the matrix. Thus

$$\begin{aligned}
\det(\mathbf{A}(1:1,1:1)) = \det(\mathbf{A}^{(1)}(1:1,1:1)) &= a_{11}^{(1)} \neq 0,\\
\det(\mathbf{A}(1:2,1:2)) = \det(\mathbf{A}^{(2)}(1:2,1:2)) &= a_{11}^{(1)}a_{22}^{(2)} \neq 0,\\
&\vdots\\
\det(\mathbf{A}(1:n,1:n)) = \det(\mathbf{A}^{(n)}(1:n,1:n)) &= a_{11}^{(1)}a_{22}^{(2)}\ldots a_{nn}^{(n)} \neq 0.
\end{aligned}$$

(We need the last condition because of the back substitution.)
This implies the statement. ■.

Performance of the Gaussian method

Thm. 18. If the coefficient matrix $\mathbf{A}$ of the SLAE

- has a strictly dominant diagonal,
- is symmetric positive definite,
- M-matrix,

then the Gaussian method can be realized with the previous algorithm.

Before the proof, we introduce M-matrices.

Def. 19. We call a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ to be an M-matrix if all its offdiagonal elements are nonpositive, it is regular and $\mathbf{A}^{-1} \geq \mathbf{0}$.

Example.

$$\mathbf{A} = \begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix}, \quad \mathbf{A}^{-1} = \begin{bmatrix} 3/4 & 1/2 & 1/4 \\ 1/2 & 1 & 1/2 \\ 1/4 & 1/2 & 3/4 \end{bmatrix}.$$

Performance of the Gaussian method - M-matrices

Thm. 20. The elements of the main diagonal of an M-matrix are positive.

Proof: If $a_{ii} \leq 0$, then $\mathbf{A}\overline{\mathbf{e}}_i \leq \mathbf{0}$. In this case $\overline{\mathbf{e}}_i \leq \mathbf{0}$, because $\mathbf{A}^{-1} \geq \mathbf{0}$, which is a contradiction. ■

Thm. 21. If $\mathbf{A}$ is an M-matrix, then there is a positive vector $\overline{\mathbf{g}} > \mathbf{0}$ such that $\mathbf{A}\overline{\mathbf{g}} > \mathbf{0}$.

Proof: Let $\overline{\mathbf{e}} = [1, \ldots, 1]^T$. Then $\overline{\mathbf{g}} = \mathbf{A}^{-1}\overline{\mathbf{e}}$ is a good choice because all elements are positive and $\mathbf{A}\overline{\mathbf{g}} = \mathbf{A}\mathbf{A}^{-1}\overline{\mathbf{e}} = \overline{\mathbf{e}} > \mathbf{0}$. ■

The converse of the theorem is also true in the following form.

Performance of the Gaussian method - M-matrices

Thm. 22. If a vector $\overline{\mathbf{g}} > 0$ exists with the property $\mathbf{A}\overline{\mathbf{g}} > 0$ and the offdiagonal of $\mathbf{A}$ is non-positive, then $\mathbf{A}$ is an M-matrix.

Proof: Let $\mathbf{G} = \operatorname{diag}(g_1, \ldots, g_n)$ and $\mathbf{D} = \operatorname{diag}(a_{11}g_1, \ldots, a_{nn}g_n)$. Then the offdiagonal of $\mathbf{D}^{-1}\mathbf{A}\mathbf{G}$ is non-positive, moreover there are ones in the main diagonal. In this way the matrix can be written in the form $\mathbf{D}^{-1}\mathbf{A}\mathbf{G} = \mathbf{I} - \mathbf{B}$, where $\mathbf{B}$ is a nonnegative matrix with zeros in the main diagonal. Because $\mathbf{D}^{-1}\mathbf{A}\mathbf{G}\overline{\mathbf{e}} = \mathbf{D}^{-1}\mathbf{A}\overline{\mathbf{g}} > 0$, $(\mathbf{I} - \mathbf{B})\overline{\mathbf{e}} > 0$. This shows that the maximum norm of $\mathbf{B}$ is less than one. Thus its spectral radius is also less than one. In this way $\mathbf{I} - \mathbf{B}$ is invertible and (see Thm. 8.) $0 \leq \mathbf{I} + \mathbf{B} + \mathbf{B}^2 + \cdots = (\mathbf{I} - \mathbf{B})^{-1}$. Thus $\mathbf{A}$ is invertible and the inverse is nonnegative. ■

Performance of the Gaussian method - M-matrices

Thm. 23. Let $\mathbf{A}$ be an M-matrix and $\overline{\mathbf{g}}$ a vector for which the condition of the above theorem is valid. Then

$$\|\mathbf{A}^{-1}\|_\infty \leq \frac{\|\overline{\mathbf{g}}\|_\infty}{\min_i(\mathbf{A}\overline{\mathbf{g}})_i}.$$

Proof: Let $\mathbf{A}\overline{\mathbf{g}} = \overline{\mathbf{y}} > \mathbf{0}$. Then

$$(\min_i y_i)\|\mathbf{A}^{-1}\|_\infty \leq \|\mathbf{A}^{-1}\overline{\mathbf{y}}\|_\infty = \|\overline{\mathbf{g}}\|_\infty,$$

from which the statement follows directly. ■

Performance of the Gaussian method

Proof:

- One step of the Gaussian method does not spoil the dominance.
- All principal minors of symmetric positive definite matrices are positive.
- Let $\mathbf{A}$ be an M-matrix ($\exists \overline{\mathbf{g}} > 0$ such that $\mathbf{A}\overline{\mathbf{g}} > 0$). Let us perform one Gauss transformation (this can be done because the diagonal is positive). Then all the offdiagonal elements remain nonpositive. $\mathbf{A}^{-1}\overline{\mathbf{g}} > 0$ is a positive vector with the property $\mathbf{A}^{(2)}\mathbf{A}^{-1}\overline{\mathbf{g}} = \mathbf{L}_1\mathbf{A}\mathbf{A}^{-1}\overline{\mathbf{g}} = \mathbf{L}_1\overline{\mathbf{g}} > 0$. This follows form the fact that $\mathbf{L}_1$ is a nonnegative matrix with 1s in the main diagonal. Thus, one step of the method preserves the M-matrix property of the coefficient matrix. Thus the diagonal is positive again, and we can step further with the method similarly. ■

Operation count for the Gaussian method

Operation count

Operation count for the elimination:

$$\frac{2(n-1)n(2n-1)}{6} + \frac{3(n-1)n}{2}$$

$$= \frac{4n^3 + 3n^2 - 7n}{6} = \frac{2}{3}n^3 + O(n^2)\ flop$$

Operation count for the back substitution: $1 + 3 + \cdots + 2n - 1 = n^2\ flop$

Altogether:

$$\frac{2}{3}n^3 + O(n^2)$$

For large matrices the number of operations for the back substitution is negligible compared to that for the elimination.

Operation count

For triangular matrices: n^2 (only back substitution).

For tridiagonal matrices: $8n - 7$.

Rmk. If we computed the solution $\overline{\mathbf{x}}$ with the formula $\overline{\mathbf{x}} = \mathbf{A}^{-1}\overline{\mathbf{b}}$ (suppose that we know the inverse somehow), then the number of operations would be $2n^2 - n$.

LU decomposition

LU decomposition

Thm. 24. Let us suppose that for the matrix $\mathbf{A}$ the condition $\det(\mathbf{A}(1:k,1:k)) \neq 0$ $(k = 1, \ldots, n-1)$ is fulfilled, that is the Gaussian elimination method can be performed for this matrix. Then there exist a normed lower triangular matrix $\mathbf{L}$ (lower) (1s are in the main diagonal) and an upper triangular matrix $\mathbf{U}$ such that $\mathbf{A} = \mathbf{LU}$ (LU decomposition). If the regular matrix $\mathbf{A}$ has an LU decomposition, then the decomposition is unique, moreover $\det(\mathbf{A}) = u_{11} \ldots u_{nn}$.

Proof: During the Gaussian elimination process the Gauss transformations change the matrix $\mathbf{A}$ as follows:

$$\mathbf{L}_{n-1}\mathbf{L}_{n-2} \ldots \mathbf{L}_1\mathbf{A} = \mathbf{U},$$

where $\mathbf{U}$ is the upper triangular matrix obtained after the elimination process.

LU decomposition

Because $(\mathbf{I} - \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T)^{-1} = \mathbf{I} + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T$ and $\bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T \bar{\mathbf{l}}_l \bar{\mathbf{e}}_l^T = \mathbf{0}$ if $l > k$, the matrix $\mathbf{A}$ can be written in the form

$$\mathbf{A} = \mathbf{L}_1^{-1} \dots \mathbf{L}_{n-2}^{-1} \mathbf{L}_{n-1}^{-1} \mathbf{U} = \left(\prod_{k=1}^{n-1} (\mathbf{I} + \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T) \right) \mathbf{U}$$

$$= \underbrace{\left(\mathbf{I} + \sum_{k=1}^{n-1} \bar{\mathbf{l}}_k \bar{\mathbf{e}}_k^T \right)}_{\text{alsó normált háromszögmátrix}} \mathbf{U} = \mathbf{L}\mathbf{U}.$$

The calculation of the determinant of the matrix $\mathbf{A}$:

$$\det(\mathbf{A}) = \det(\mathbf{L})\det(\mathbf{U}) = u_{11} \dots u_{nn}.$$

LU decomposition

Uniqueness:
Let us suppose that there are two different decompositions: $\mathbf{A} = \tilde{\mathbf{L}}\tilde{\mathbf{U}} = \mathbf{LU}$. Then

$$\tilde{\mathbf{L}}^{-1}\mathbf{L} = \tilde{\mathbf{U}}\mathbf{U}^{-1} = \mathbf{I},$$

because the product of normed lower triangular matrices is normed lower triangular and similar statement is true for upper triangular matrices. ■

Rmk. The matrix $\mathbf{U}$ is the upper triangular matrix that is formed during the elimination process, matrix $\mathbf{L}$ is the matrix of the L_{ij} coefficients

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \dots & 0 \\ l_{31} & l_{32} & \dots & 0 \\ \vdots & & & \\ l_{n1} & l_{n2} & \dots & 1 \end{bmatrix}.$$

Corollary: If one of the main minors of a regular matrix is zero, then the matrix does not have LU decomposition.

LU decomposition

Example.

$$\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/3 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1/2 & 1/3 \\ 0 & 1/12 & 1/12 \\ 0 & 0 & \frac{1}{180} \end{bmatrix}$$

Remarks:

- If we have computed the LU decomposition of $\mathbf{A}$, then the matrices $\mathbf{L}$ and $\mathbf{U}$ can be stored in the computer memory in the place of $\mathbf{A}$. The SLAE $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ can be solved with the solution of two SLAEs with triangular coefficient matrices. Operation: $2n^2 << 2n^3/3$.
- We generally do not calculate the inverse of matrices! If we need to do this, then we can perform this task with the expression $\mathbf{U}^{-1}\mathbf{L}^{-1}$ or using the Gauss–Jordan method. The number of operations is $2n^3 + O(h^2)$ in both cases.

The effect of rounding errors

We see earlier that $|\mathit{fl}(a_{ij}) - a_{ij}| \leq \mathtt{u}|a_{ij}|$.

For a matrix $\mathbf{A} = [a_{ij}]$, let us denote the matrix $[|a_{ij}|]$ by $|\mathbf{A}|$. Thus we can write that

$$|\mathit{fl}(\mathbf{A}) - \mathbf{A}| \leq \mathtt{u}|\mathbf{A}|.$$

Thm. 25. [GL, page 105] Let us suppose that after the application of the Gaussian method using floating point numbers, we obtain the matrices $\hat{\mathbf{L}}$ and $\hat{\mathbf{U}}$, for which $\hat{\mathbf{L}}\hat{\mathbf{U}} - \mathbf{A} = \boldsymbol{\delta}\mathbf{A}$. Then the estimation

$$|\boldsymbol{\delta}\mathbf{A}| \leq 3(n-1)\mathtt{u}(|\mathbf{A}| + |\hat{\mathbf{L}}| \cdot |\hat{\mathbf{U}}|) + K\mathtt{u}^2$$

holds, where K is a positive constant.

Pivoting

Pivoting

The Gaussian method can be performed only if the pivot elements are not zero. What should we do if $a_{kk}^{(k)}$ is zero?

- Let us choose a non-zero element from the column $\mathbf{A}(k+1:n,k)$. Let us denote this element by s. Let us swap the two rows (change of indexes), then let us continue the elimination.
- If there is no non-zero element in the column $\mathbf{A}(k+1:n,k)$, then the first k columns are linearly dependent, thus $\det(\mathbf{A})=0$. In this case there is not unique solution.
- Partial pivoting: The matrix $\mathbf{L}$ appears in the error estimation in Theorem 25. It can be a good idea to decrease the elements of $\mathbf{L}$ in absolute value. In view of the form $l_{sk}=a_{sk}^{(k)}/a_{kk}^{(k)}$, the error can be decreased by choosing the largest element in absolute value to be the pivot element. The number of the required operations is $(n^2-n)/2$ comparison.

Pivoting

- Full pivoting: In the kth step we choose the greatest element in absolute value from the sub-matrix $\mathbf{A}(k:n,k:n)$. This is $n(n+1)(2n+1)/6-1=n^3/3+O(n^2)$ comparison.

Let us consider the problem, and let us round to 4 significant digits.

$$\begin{aligned}0.003x_1+59.14x_2 &= 59.17\\ 5.291x_1-6.13x_2 &= 46.78\end{aligned}$$

Exact solution $x_1=10.00$, $x_2=1.000$. Without pivoting, we obtain $x_1=-10$, $x_2=1.001$ (cancellation), with partial pivoting we obtain the exact solution.

LU decomposition for general matrices

LU decomposition for general matrices

Thm. 26. (LU decomposition for general matrices) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be an arbitrary matrix. Then there is a normed lower triangular matrix $\mathbf{L}$ with elements not greater than 1 in absolute value, an upper triangular matrix $\mathbf{U}$, and a permutation matrix $\mathbf{P}$ such that $\mathbf{PA} = \mathbf{LU}$.

Error analysis

Let us suppose that we perform the Gaussian method with partial pivoting on a computer that uses floating point numbers. Let us suppose that we arrive at the matrices $\mathbf{P}$, $\hat{\mathbf{L}}$ and $\hat{\mathrm{U}}$, for which we have $\mathbf{P}^T\hat{\mathbf{L}}\hat{\mathrm{U}} - \mathbf{A} = \boldsymbol{\delta}\mathbf{A}$. Then the following estimation is true:

$$|\boldsymbol{\delta}\mathbf{A}| \leq 3(n-1)\mathbf{u}(|\mathbf{A}| + |\mathbf{P}^T| \cdot |\hat{\mathbf{L}}| \cdot |\hat{\mathrm{U}}|) + O(\mathrm{u}^2).$$

Thus

$$\|\boldsymbol{\delta}\mathbf{A}\|_\infty \leq 3(n-1)\mathbf{u}(\|\mathbf{A}\|_\infty + n\|\hat{\mathrm{U}}\|_\infty) + O(\mathrm{u}^2).$$

With the notation

$$\rho = \max_{i,j,k} \frac{|\hat{a}_{ij}^{(k)}|}{\|\mathbf{A}\|_\infty} \text{ (in practice } \leq 10 \text{ but can be also } 2^{n-1}\text{)},$$

this is the so-called growth factor, we obtain

$$\begin{aligned}\|\boldsymbol{\delta}\mathbf{A}\|_\infty &\leq 3(n-1)\mathbf{u}(\|\mathbf{A}\|_\infty + n^2\rho\|\mathbf{A}\|_\infty) + O(\mathrm{u}^2)\\ &\leq 6n^3\rho\|\mathbf{A}\|_\infty\mathbf{u} + O(\mathrm{u}^2).\end{aligned}$$

$\mathbf{LDM}^T$ decomposition

$\mathbf{LDM}^T$ decomposition

Thm. 27. Let us suppose that all main minors of $\mathbf{A}$ are non-zero. Then there exist the unique normed lower triangular matrices $\mathbf{L}$ and $\mathbf{M}$ and the diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{LDM}^T$.

Proof: The $\mathbf{LU}$ decomposition is performable. Let $\mathbf{D}$ be such that $d_{ii} = u_{ii} (\neq 0)$. Then the matrix $\mathbf{M} = (\mathbf{D}^{-1}\mathbf{U})^T$ is a normed lower triangular matrix. Moreover $\mathbf{LD}(\mathbf{D}^{-1}\mathbf{U}) = \mathbf{LU} = \mathbf{A}$. The uniqueness follows from the uniqueness of the LU decomposition. ■

Thm. 28. For symmetric matrices $\mathbf{A}$, there exists a unique normed lower triangular matrix $\mathbf{L}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{A} = \mathbf{LDL}^T$.

Proof: The matrix $\mathbf{M}^{-1}\mathbf{A}\mathbf{M}^{-\top} = \mathbf{M}^{-1}\mathbf{LD}$ is symmetric and lower triangular $\Rightarrow$ diagonal. $\det(\mathbf{D}) \neq 0 \Rightarrow \mathbf{M}^{-1}\mathbf{L}$ is also diagonal but also normed lower triangular. That is $\mathbf{M}^{-1}\mathbf{L} = \mathbf{I}$, and $\mathbf{M} = \mathbf{L}$. ■

Cholesky decomposition

Cholesky decomposition

Thm. 29. Let us suppose that $\mathbf{A}$ is a symmetric and positive definite matrix. Then there exist a unique lower triangular matrix $\mathbf{G}$ with positive diagonal such that $\mathbf{A} = \mathbf{G}\mathbf{G}^T$.

Proof: The matrix $\mathbf{A}$ can be written uniquely in the form $\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^T$. The diagonal matrix $\mathbf{D}$ has positive diagonal. Let $\mathbf{G} = \mathbf{L} \cdot \mathrm{diag}(\sqrt{d_{11}}, \ldots, \sqrt{d_{nn}})$, which is a lower triangular matrix with positive diagonal. Moreover $\mathbf{G}\mathbf{G}^T = \mathbf{A}$. ■

Rmk. In practice, the Cholesky decomposition is not calculated with the above expression but the elements of $\mathbf{G}$ are calculated directly from above and from left by the help of the expression $\mathbf{A} = \mathbf{G}\mathbf{G}^T$. The number of operations is $n^3/3 + O(n^2)$ flop $+ \, n$ square root.

André-Louis Cholesky, 1875–1918, French

Iterative solutions of SLAEs

Linear iterative methods

When do we use iterative methods?

We would like to define a linear iteration

$$\overline{\mathbf{x}}_{k+1} = \mathbf{B}\overline{\mathbf{x}}_k + \overline{\mathbf{f}},\ \ k = 0, 1, \dots$$

such that the limit of the vector sequence is the solution of the system $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$.

The number of operations in one iteration step is $2n^2\ flop$. Thus, we can perform $n/3$ iteration steps in order to not to exceed the number of operations of the Gauss method. The method is mainly used for sparse matrices, when the number of nonzero elements is $O(n)$ (e.g. band matrices).

Questions:

- When does the sequence converge to the solution?
- How fast is the convergence?
- How to choose the matrix $\mathbf{B}$ and the vectors $\overline{\mathbf{f}}$, $\overline{\mathbf{x}}^{(0)}$?
- When to stop the iteration?

Convergence of iterative methods

Because of the inequality

$$\|\mathbf{B}\overline{\mathbf{x}}' - \overline{\mathbf{f}} - (\mathbf{B}\overline{\mathbf{x}}'' - \overline{\mathbf{f}})\| \leq \|\mathbf{B}\| \cdot \|\overline{\mathbf{x}}' - \overline{\mathbf{x}}'\|$$

and the Banach fixed point theorem, if $\|\mathbf{B}\| < 1$ in some induced norm ($\Leftrightarrow \varrho(\mathbf{B}) < 1$), and the solution $\overline{\mathbf{x}}^\star$ of the system is a fixed point of the map $\overline{\mathbf{x}} \mapsto \mathbf{B}\overline{\mathbf{x}} + \overline{\mathbf{f}}$ then starting the iteration from an arbitrary vector, it will tend to the solution of the system. Moreover

$$\|\overline{\mathbf{x}}_k - \overline{\mathbf{x}}^\star\| \leq \frac{\|\mathbf{B}\|^k}{1 - \|\mathbf{B}\|} \|\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_0\|.$$

Rmk. The smaller the spectral radius the faster the convergence.

The construction of the iteration

The iteration can be constructed as follows. Let $\mathbf{A} = \mathbf{S} - \mathbf{T}$ and let $\mathbf{S}$ be nonsingular. Then

$$\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}} \;\rightarrow\; (\mathbf{S} - \mathbf{T})\overline{\mathbf{x}} = \overline{\mathbf{b}} \;\rightarrow\; \overline{\mathbf{x}} = \mathbf{S}^{-1}\mathbf{T}\overline{\mathbf{x}} + \mathbf{S}^{-1}\overline{\mathbf{b}}.$$

$$\overline{\mathbf{x}}_{k+1} = \underbrace{\mathbf{S}^{-1}\mathbf{T}}_{\mathbf{B}}\,\overline{\mathbf{x}}_k + \underbrace{\mathbf{S}^{-1}\overline{\mathbf{b}}}_{\overline{\mathbf{f}}}.$$

The matrix $\mathbf{S}$ is called preconditioner. Because $\mathbf{B} = \mathbf{I} - \mathbf{S}^{-1}\mathbf{A}$, a good preconditioner must be

- close to $\mathbf{A}$, hence the norm of $\mathbf{B}$ can be small in this case (see later).
- and easily invertible.

Example.

- $\mathbf{S} = \mathbf{A}$: it is close to $\mathbf{A}$ but the computation of its inverse is as difficult as that of $\mathbf{A}$. The method converges in one step.
- $\mathbf{S} = \mathbf{I}$: inverse is easy, but it has nothing to do with $\mathbf{A}$.

Jacobi iteration

Jacobi iteration

Let $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{R}$, where $\mathbf{D}$ is the diagonal matrix of $\mathbf{A}$ (suppose that there are no zeros in the diagonal). $\mathbf{L}$ is the matrix of the elements below the diagonal, while $\mathbf{R}$ is constructed from the elements above the diagonal, and both multiplied by -1. Let $\mathbf{S} = \mathbf{D}$ and $\mathbf{T} = \mathbf{R} + \mathbf{L}$.

Def. 30. The iteration

$$\overline{\mathbf{x}}_{k+1} = \underbrace{\mathbf{D}^{-1}(\mathbf{L} + \mathbf{R})}_{:=\mathbf{B}_J} \overline{\mathbf{x}}_k + \mathbf{D}^{-1}\overline{\mathbf{b}}$$

constructed with the above splitting ($\overline{\mathbf{x}}_0$ is arbitrary) is called Jacobi iteration.

Jacobi iteration

Carl Gustav Jacob Jacobi (1804-1851, German)

Componentwise:

$$(\overline{\mathbf{x}}_{k+1})_i = -\frac{1}{a_{ii}}\left(\sum_{j=1,\neq i}^{n} a_{ij}(\overline{\mathbf{x}}_k)_j - b_i\right), \quad i = 1, \ldots, n.$$

Gauss–Seidel iteration

Gauss–Seidel iteration

Let us modify the previous iteration! Let us use the newly computed components!

$$(\overline{\mathbf{x}}_{k+1})_i = -\frac{1}{a_{ii}}\left(\sum_{j=1}^{i-1} a_{ij}(\overline{\mathbf{x}}_{k+1})_j + \sum_{j=i+1}^{n} a_{ij}(\overline{\mathbf{x}}_k)_j - b_i\right).$$

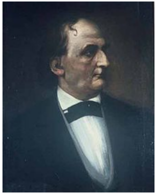
Matrix form:

$$\overline{\mathbf{x}}_{k+1} = \mathbf{D}^{-1}(\mathbf{L}\overline{\mathbf{x}}_{k+1} + \mathbf{R}\overline{\mathbf{x}}_k + \overline{\mathbf{b}}),$$

that is

$$\overline{\mathbf{x}}_{k+1} = \underbrace{(\mathbf{D}-\mathbf{L})^{-1}\mathbf{R}}_{\mathbf{B}_{GS}}\,\overline{\mathbf{x}}_k + (\mathbf{D}-\mathbf{L})^{-1}\overline{\mathbf{b}}.$$

Def. 31. The iteration constructed with the splitting $\mathbf{S} = \mathbf{D} - \mathbf{L}$, $\mathbf{T} = \mathbf{R}$ ($\overline{\mathbf{x}}_0$ is arbitrary) is called Gauss–Seidel iteration.



Philipp Ludwig von Seidel (1821-1896, German)

Comparison of the Jacobi and Gauss–Seidel iterations

The Gauss–Seidel seams to be better, because we always use the updated components, but if

$$\mathbf{A} = \begin{bmatrix} 1 & 1/2 & 1 \\ 1/2 & 1 & 1 \\ -2 & 2 & 1 \end{bmatrix}$$

then

$$\mathbf{B}_J = \begin{bmatrix} 0 & -1/2 & -1 \\ -1/2 & 0 & -1 \\ 2 & -2 & 0 \end{bmatrix}, \quad \mathbf{B}_{GS} \begin{bmatrix} 0 & -1/2 & -1 \\ 0 & 1/4 & -1/2 \\ 0 & -3/2 & -1 \end{bmatrix}.$$

Thus $\varrho(\mathbf{B}_J) = 1/2 < 1$ and $\varrho(\mathbf{B}_{GS}) = |-3/8 - \sqrt{73}/8| \approx 1.443 > 1$.

Relaxation methods

Relaxation methods

The Jacobi method fulfills the equality:

$$(\overline{\mathbf{x}}_{k+1})_i = (\overline{\mathbf{x}}_k)_i + (\overline{\mathbf{x}}_{k+1})_i - (\overline{\mathbf{x}}_k)_i.$$

The main idea of the relaxation for the Jacobi method:

$$(\tilde{\overline{\mathbf{x}}}_{k+1})_i = (\tilde{\overline{\mathbf{x}}}_k)_i + \omega((\tilde{\overline{\mathbf{x}}}_{k+1})_{i,J} - (\tilde{\overline{\mathbf{x}}}_k)_i), \;\; 0 \neq \omega \in \mathbb{R},$$

where $(\tilde{\overline{\mathbf{x}}}_0)_i = (\overline{\mathbf{x}}_0)_i$, $(\tilde{\overline{\mathbf{x}}}_{k+1})_{i,J}$ is the value where the Jacobi method would step from $(\tilde{\overline{\mathbf{x}}}^k)_i$ $(i = 1, \ldots, n)$, and ω is a so-called relaxation parameter.

Main goal: how to choose ω in order to make the convergence faster?

- $\omega = 1$: we get back the Jacobi iteration.
- $0 < \omega < 1$: under-relaxation.
- $\omega > 1$: over-relaxation.

JOR method (Jacobi over-relaxation, $J(\omega)$)

The componentwise form of the JOR method (without $\tilde{}$):

$$(\overline{\mathbf{x}}_{k+1})_i = (\overline{\mathbf{x}}_k)_i + \omega \left(-\frac{1}{a_{ii}} \left(\sum_{j=1,\neq i}^{n} a_{ij} (\overline{\mathbf{x}}_k)_j - b_i \right) - (\overline{\mathbf{x}}_k)_i \right)$$

$$= (1-\omega)(\overline{\mathbf{x}}_k)_i - \frac{\omega}{a_{ii}} \left[\sum_{j=1, j\neq i}^{n} a_{ij} (\overline{\mathbf{x}}_k)_j - b_i \right].$$

Thus we arrive at the vector form

$$\overline{\mathbf{x}}_{k+1} = \underbrace{((1-\omega)\mathbf{I} + \omega \mathbf{D}^{-1}(\mathbf{L}+\mathbf{R}))}_{\mathbf{B}_{J(\omega)}} \overline{\mathbf{x}}_k + \omega \mathbf{D}^{-1} \overline{\mathbf{b}},$$

where the iteration matrix is

$$\mathbf{B}_{J(\omega)} = \omega \mathbf{B}_J + (1-\omega)\mathbf{I}. \qquad (4)$$

SOR method (Successive over-relaxation, $GS(\omega)$)

This method is the relaxation of the Gauss–Seidel method:

We apply the relaxation elementwise:

$$(\overline{\mathbf{x}}_{k+1})_i = (1-\omega)(\overline{\mathbf{x}}_k)_i - \frac{\omega}{a_{ii}}\left[\sum_{j=1}^{i-1} a_{ij}(\overline{\mathbf{x}}_{k+1})_j + \sum_{j=i+1}^{n} a_{ij}(\overline{\mathbf{x}}_k)_j - b_i\right].$$

In matrix form:

$$\overline{\mathbf{x}}_{k+1} = \underbrace{(\mathbf{D}-\omega\mathbf{L})^{-1}((1-\omega)\mathbf{D}+\omega\mathbf{R})}_{\mathbf{B}_{GS(\omega)}}\overline{\mathbf{x}}_k + \omega(\mathbf{D}-\omega\mathbf{L})^{-1}\overline{\mathbf{b}}.$$

Convergence

Convergence of regular splitting

Def. 32. The splitting $\mathbf{A} = \mathbf{S} - \mathbf{T}$ of the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called regular splitting, if $\mathbf{S}$ is non-singular, $\mathbf{S}^{-1} \geq \mathbf{0}$ and $\mathbf{T} \geq \mathbf{0}$.

Thm. 33. If $\mathbf{A} = \mathbf{S} - \mathbf{T}$ is a regular splitting of a non-singular matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ with the property $\mathbf{A}^{-1} \geq \mathbf{0}$ then $\varrho(\mathbf{S}^{-1}\mathbf{T}) < 1$.

Proof. Let $\mathbf{B} = \mathbf{S}^{-1}\mathbf{T} \geq \mathbf{0}$. Then

$$0 \leq \left(\sum_{i=0}^{k} \mathbf{B}^i\right) \mathbf{S}^{-1} = \left(\sum_{i=0}^{k} \mathbf{B}^i\right) (\mathbf{I} - \mathbf{B})\mathbf{A}^{-1}$$

$$= (\mathbf{I} - \underbrace{\mathbf{B}^{k+1}}_{\geq \mathbf{0}})\underbrace{\mathbf{A}^{-1}}_{\geq \mathbf{0}} \leq \mathbf{A}^{-1}.$$

That is the series is convergent, thus $\varrho(\mathbf{B}) < 1$. ■

Convergence of regular splitting

Thm. 34. Let $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{R}$ (with the previous splitting), where we have $\mathbf{L} + \mathbf{R} \geq \mathbf{0}$. Then the matrix $\mathbf{A}$ has a regular splitting $\mathbf{A} = \mathbf{S} - \mathbf{T}$ with the property $\varrho(\mathbf{S}^{-1}\mathbf{T}) < 1$ iff $\mathbf{A}$ is an M-matrix.

Proof ($\Leftarrow$) Let $\mathbf{S} = \mathbf{D} > \mathbf{0}$ and $\mathbf{T} = \mathbf{L} + \mathbf{R}$. This is a regular splitting, moreover, because $\mathbf{A}^{-1} \geq \mathbf{0}$. Thus $\varrho(\mathbf{S}^{-1}\mathbf{T}) < 1$ because of the previous theorem.
($\Rightarrow$) The signs of the elements are OK. We have to show that $\mathbf{A}$ is non-singular and its inverse is nonnegative.

$$\mathbf{A}^{-1} = (\mathbf{S} - \mathbf{T})^{-1} = (\mathbf{S}(\mathbf{I} - \mathbf{S}^{-1}\mathbf{T}))^{-1} = (\mathbf{I} - \underbrace{\mathbf{S}^{-1}\mathbf{T}}_{\varrho<1})^{-1}\mathbf{S}^{-1}$$

$$= \sum_{k=0}^{\infty} (\mathbf{S}^{-1}\mathbf{T})^k \mathbf{S}^{-1} \geq \mathbf{0}. \blacksquare$$

Convergence of the Jacobi and Gauss–Seidel iterations

Thm. 35. For M-matrices, the J, J(ω), GS and GS(ω) ($\omega \in (0,1]$) methods are all convergent.

Proof. If $\mathbf{A}$ is an M-matrix then $\mathbf{A}^{-1} \geq \mathbf{0}$. In the case of the JOR method, the choice

$$\mathbf{S} = \frac{1}{\omega}\mathbf{D}, \ \mathbf{T} = \frac{1-\omega}{\omega}\mathbf{D} + \mathbf{L} + \mathbf{R}$$

gives a regular splitting for $\omega \in (0,1]$. Thus the iteration is convergent based on the previous theorem.
In the case of the SOR method, the choice

$$\mathbf{S} = \frac{1}{\omega}\mathbf{D} - \mathbf{L}, \ \mathbf{T} = \frac{1-\omega}{\omega}\mathbf{D} + \mathbf{R}$$

gives regular splitting for all $\omega \in (0,1]$. The case $\omega = 1$ gives back the Jacobi and Gauss–Seidel methods. ■

Convergence of the Jacobi and Gauss–Seidel iterations

Thm. 36. For matrices with strictly dominant diagonal, the Jacobi iteration is convergent. (Similar theorem is true for the Gauss–Seidel iteration.)

Proof.

$$\varrho(\mathbf{B}_J) \leq \|\mathbf{B}_J\|_\infty = \max_{i=1,\dots,n} \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1. \blacksquare$$

Thm. 37. If $\mathbf{A}$ is symmetric and positive definite then the Gauss–Seidel iteration is convergent.

Thm. 38. [Ostrowski, Reich] If $\mathbf{A}$ is symmetric and $\omega \in (0, 2)$ then

$$\varrho(\mathbf{B}_{GS(\omega)}) < 1,$$

that is the SOR method is convergent.

Convergence of the Jacobi and Gauss–Seidel iterations

Thm. 39. [Kahan] For the SOR method we have

$$\varrho(\mathbf{B}_{GS(\omega)}) \geq |1-\omega|,$$

that is the necessary condition of the convergence is $\omega \in (0,2)$.

Proof.

$$\prod_{i=1}^{n} |\lambda_i| = |\det(\mathbf{B}_{GS(\omega)})| =$$

$$= |\det((\mathbf{D}-\omega\mathbf{L})^{-1})| \cdot |\det((1-\omega)\mathbf{D}+\omega\mathbf{R})| = |1-\omega|^n.$$

Thus

$$\varrho(\mathbf{B}_{GS(\omega)}) = \max_{i=1,\ldots,n} |\lambda_i| \qquad \underbrace{\geq}_{\text{arithm. and geom. mean}}$$

$$\geq \left(\prod_{i=1}^{n} |\lambda_i|\right)^{1/n} = |1-\omega|. \blacksquare$$

Stopping conditions

Stopping conditions

When to stop the iteration?

- If $\|\mathbf{B}\| < 1$ in some norm then based on the Banach fixed point theorem we have
$$\|\overline{\mathbf{x}} - \overline{\mathbf{x}}^{(j)}\| \leq \frac{\|\mathbf{B}\|^j}{1 - \|\mathbf{B}\|} \|\overline{\mathbf{x}}^{(1)} - \overline{\mathbf{x}}^{(0)}\|.$$
From the value $\|\mathbf{B}\|$ and the result of the first iteration, we can calculate that how many iteration we need to achieve a prescribed accuracy in a certain norm.
- Consider the results of two consecutive iterations. If $\|\overline{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_k\|$ is sufficiently small then we stop the iteration.
- We compute the so-called remainders: $\overline{\mathbf{r}}_k = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k$. If $\|\overline{\mathbf{r}}_{k+1} - \overline{\mathbf{r}}_k\| / \|\overline{\mathbf{r}}^{(0)}\|$ is sufficiently small then we stop the iteration.
- We fix a value $k_{\max}$ where we stop the iteration at all events.

Gradient methods

Minimizing property

Minimizing property

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a symmetric, positive definite matrix and let us consider the multivariable function

$$\phi(\overline{\mathbf{x}}) = \frac{1}{2}\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{x}}^T\overline{\mathbf{b}}$$

with n unknowns.

Thm. 40. The function $\phi(\overline{\mathbf{x}})$ has exactly one stationary point, the point $\overline{\mathbf{x}}^\star = \mathbf{A}^{-1}\overline{\mathbf{b}}$ (the solution of the system $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$).

Proof. We have

$$\phi(\overline{\mathbf{x}}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j - \sum_{j=1}^{n} b_jx_j.$$

After getting rid of the terms that do not contain x_k we obtain:

Minimizing property

$$\frac{\partial \phi}{\partial x_k}(\overline{\mathbf{x}}) =$$

$$= \left(\frac{1}{2}\left(\sum_{k\neq i=1}^{n} a_{ik}x_i x_k + \sum_{k\neq j=1}^{n} a_{kj}x_k x_j + a_{kk}x_k^2\right) - b_k x_k\right)'_{x_k}$$

$$= \sum_{j=1}^{n} a_{kj}x_j - b_k.$$

That is $\operatorname{grad}\phi(\overline{\mathbf{x}}) = \mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}$. Thus the only stationary point is $\overline{\mathbf{x}}^{\star}$ indeed. ■

Minimizing property

Thm. 41. The absolute minimizer of $\phi(\overline{\mathbf{x}})$ is $\overline{\mathbf{x}}^\star = \mathbf{A}^{-1}\overline{\mathbf{b}}$. The minimum value is $-\overline{\mathbf{b}}^T\mathbf{A}^{-1}\overline{\mathbf{b}}/2$.

Proof. Let $\overline{\mathbf{x}} = \overline{\mathbf{x}}^\star + \Delta\overline{\mathbf{x}}$ be an arbitrary vector.

$$\phi(\overline{\mathbf{x}}^\star + \Delta\overline{\mathbf{x}}) = \phi(\overline{\mathbf{x}}^\star) + \frac{1}{2}\Delta\overline{\mathbf{x}}^T\mathbf{A}\Delta\overline{\mathbf{x}}.$$

The statement follows from the positive definiteness of the matrix $\mathbf{A}$. The minimum value comes with simple substitution. ■.

When we change $\Delta\overline{\mathbf{x}}$ with $\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star$ in the above formula then

$$\phi(\overline{\mathbf{x}}) = -\frac{1}{2}\overline{\mathbf{b}}^T\mathbf{A}^{-1}\overline{\mathbf{b}} + \frac{1}{2}(\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star)^T\mathbf{A}(\overline{\mathbf{x}} - \overline{\mathbf{x}}^\star).$$

It can be seen from this, that the level curves of the function (if $c \geq -\overline{\mathbf{b}}^T\mathbf{A}^{-1}\overline{\mathbf{b}}/2$) are hyperellipses with center $\overline{\mathbf{x}}^\star$. (If $\mathbf{A} = \mathbf{I}$ then we obtain concentric circles.)

An example in two variables

Let us consider the SLAE $2x_1 = 4$, $8x_2 = 8$. Its solution is $x_1^\star = 2, x_2^\star = 1$. Then

$$\phi(\overline{\mathbf{x}}) = x_1^2 + 4x_2^2 - 4x_1 - 8x_2 = (x_1 - 2)^2 + 4(x_2 - 1)^2 - 8.$$

Thus the minimizer of this function is $\overline{\mathbf{x}}^\star$ indeed. The minimum value is -8.

The equation of the level curve to the value $c = 0$

$$\frac{(x_1 - 2)^2}{8} + \frac{(x_2 - 1)^2}{2} = 1,$$

which is the equation of an ellipse with center $\overline{\mathbf{x}}^\star$ and semi-axis $\sqrt{8}$ and $\sqrt{2}$.

Equivalent reformulations

The search for the solution $\overline{\mathbf{x}}^{\star}$ of $\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}$ is equivalent with

- the search for the minimizer of the function $\phi(\overline{\mathbf{x}})$,
- the search for the lowest point of a surface, if the level curves of the surface are hyperellipsoids.

Directional minimizers

Let us introduce the residual vector $\overline{\mathbf{r}} = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}$.

Let us investigate the following general question: Let us move from the point $\overline{\mathbf{x}}$ in the direction of the vector $\overline{\mathbf{p}} \neq \mathbf{0}$. When will we at the lowest point? That is we search for the real parameter α that minimizes the one-variable function

$$\begin{aligned}\phi(\overline{\mathbf{x}} + \alpha\overline{\mathbf{p}}) &= \phi(\overline{\mathbf{x}}) - \alpha\overline{\mathbf{p}}^T\overline{\mathbf{r}} + \frac{1}{2}\alpha^2\overline{\mathbf{p}}^T\mathbf{A}\overline{\mathbf{p}} \\ &= \phi(\overline{\mathbf{x}}) + \alpha\left(\frac{1}{2}\alpha\overline{\mathbf{p}}^T\mathbf{A}\overline{\mathbf{p}} - \overline{\mathbf{p}}^T\overline{\mathbf{r}}\right).\end{aligned}$$

We obtain minimum if $\alpha = \overline{\mathbf{p}}^T\overline{\mathbf{r}}/(\overline{\mathbf{p}}^T\mathbf{A}\overline{\mathbf{p}})$.

Basic algorithm

Basic algorithm, $\mathbf{A} \in \mathbb{R}^{n \times n}$ SPD, $\overline{\mathbf{b}} \in \mathbb{R}^{n}$ given.

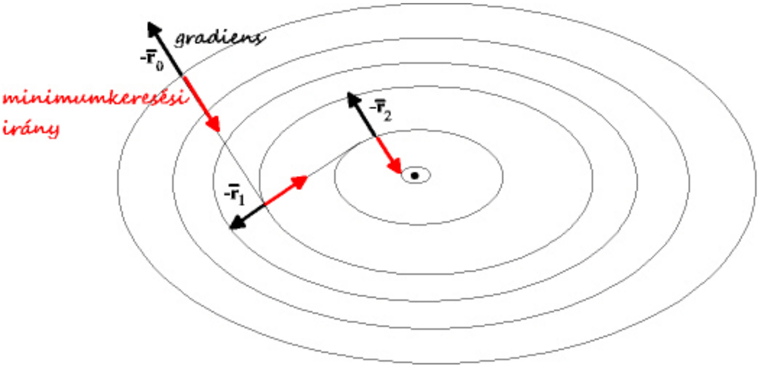
```
  $k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$
  while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ do
      $k := k + 1$
      Let chosse a $k$th search direction: $\overline{\mathbf{p}}_k \neq \mathbf{0}$
      $\alpha_k := \overline{\mathbf{p}}_k^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k)$
      $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{p}}_k$
      $\overline{\mathbf{r}}_k := \overline{\mathbf{b}} - \mathbf{A} \overline{\mathbf{x}}_k$
  end while
```

How to choose the search directions to achieve fast convergence to the solution of the SLAE?

Gradient method

Gradient method (steepest descend)

Illustration of the method:


gradiens
$-\overline{\mathbf{r}}_0$
minimumkeresési irány
$-\overline{\mathbf{r}}_2$
$-\overline{\mathbf{r}}_1$

When we choose the steepest direction down from the point $\overline{\mathbf{x}}$ (direction opposite to the gradient vector, that is $\overline{\mathbf{p}} = \overline{\mathbf{r}} = -(\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}})$), and we want to get to the lowest point in this direction, then we must step from the point $\overline{\mathbf{x}}$ to the point $\overline{\mathbf{x}} + (\overline{\mathbf{r}}^T\overline{\mathbf{r}}/(\overline{\mathbf{r}}^T\mathbf{A}\overline{\mathbf{r}}))\overline{\mathbf{r}}$. This is the gradient method.

Gradient method

Gradient method, $\mathbf{A} \in \mathbb{R}^{n \times n}$ SPD, $\overline{\mathbf{b}} \in \mathbb{R}^n$ given.

```
$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$
while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ do
    $k := k + 1$
    $\alpha_k := \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{r}}_{k-1}^T \mathbf{A} \overline{\mathbf{r}}_{k-1})$
    $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{r}}_{k-1}$
    $\overline{\mathbf{r}}_k := \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k$
end while
```

Thm. 42.

$$\frac{\phi(\overline{\mathbf{x}}_{k+1}) + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}}}{\phi(\overline{\mathbf{x}}_k) + (1/2)\overline{\mathbf{b}}^T \mathbf{A}^{-1} \overline{\mathbf{b}}} \leq 1 - \frac{1}{\kappa_2(\mathbf{A})}$$

This shows a relatively slow convergence, especially if $\kappa_2(\mathbf{A})$ is large.

Conjugate gradient method (CGM)

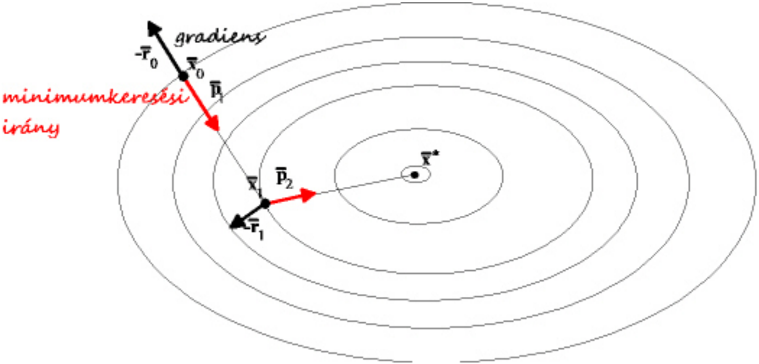
Conjugate gradient method

Early 1950s.

Magnus Hestenes

Eduard Stiefel

Main idea of the conjugate gradient method


gradiens
$-\overline{\mathbf{r}}_0$
$\overline{\mathbf{x}}_0$
$\overline{\mathbf{p}}_1$
minimumkeresési
irány
$\overline{\mathbf{x}}^\star$
$\overline{\mathbf{x}}_1$
$\overline{\mathbf{p}}_2$
$-\overline{\mathbf{r}}_1$

The vector $\overline{\mathbf{p}}_2$ must point in the direction of the vector $\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}_1$.

Def. 43. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a given symmetric, positive definite matrix. We say that the vectors $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are **A**-orthogonal (**A**-conjugate), if $\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{y}} = 0$.

Main idea of the conjugate gradient method

$$0 = \overline{\mathbf{p}}_1^T \overline{\mathbf{r}}_1 = \overline{\mathbf{p}}_1^T (\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_1) = \overline{\mathbf{p}}_1^T (\mathbf{A}\overline{\mathbf{x}}^\star - \mathbf{A}\overline{\mathbf{x}}_1) = \overline{\mathbf{p}}_1^T \mathbf{A}(\overline{\mathbf{x}}^\star - \overline{\mathbf{x}}_1).$$

The vector $\overline{\mathbf{p}}_2$ must be $\mathbf{A}$-orthogonal to $\overline{\mathbf{p}}_1$. Let us search $\overline{\mathbf{p}}_2$ in the form

$$\overline{\mathbf{p}}_2 = \overline{\mathbf{r}}_1 - \beta_1 \overline{\mathbf{p}}_1.$$

We have

$$\beta_1 = \frac{\overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{r}}_1}{\overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{p}}_1}.$$

Moreover we can simplify the calculations as follows

$$\overline{\mathbf{r}}_{k+1} - \overline{\mathbf{r}}_k = \overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_{k+1} - (\overline{\mathbf{b}} - \mathbf{A}\overline{\mathbf{x}}_k)$$

$$= -\mathbf{A}(\overline{\mathbf{x}}_{k+1} - \overline{\mathbf{x}}_k) = -\mathbf{A}\alpha_{k+1}\overline{\mathbf{p}}_{k+1}.$$

Main idea of the conjugate gradient method

Thus, we have

$$\overline{\mathbf{r}}_{k+1} = \overline{\mathbf{r}}_k - \mathbf{A}\alpha_{k+1}\overline{\mathbf{p}}_{k+1}.$$

Thus we arrive at the algorithm: $\overline{\mathbf{x}}_0 = \mathbf{0}$, $\overline{\mathbf{r}}_0 = \overline{\mathbf{b}}$ given.

1. Let $\overline{\mathbf{p}}_1 = \overline{\mathbf{r}}_0$.
2. $\alpha_1 := \overline{\mathbf{p}}_1^T \overline{\mathbf{r}}_0 / (\overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{p}}_1)$.
3. $\overline{\mathbf{x}}_1 := \overline{\mathbf{x}}_0 + \alpha_1 \overline{\mathbf{p}}_1$.
4. $\overline{\mathbf{r}}_1 := \overline{\mathbf{r}}_0 - \alpha_1 \mathbf{A} \overline{\mathbf{p}}_1$.
5. $\beta_1 := \overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{r}}_1 / (\overline{\mathbf{p}}_1^T \mathbf{A} \overline{\mathbf{p}}_1)$.
6. $\overline{\mathbf{p}}_2 = \overline{\mathbf{r}}_1 - \beta_1 \overline{\mathbf{p}}_1$.
7. $\alpha_2 := \overline{\mathbf{p}}_2^T \overline{\mathbf{r}}_1 / (\overline{\mathbf{p}}_2^T \mathbf{A} \overline{\mathbf{p}}_2)$.
8. $\overline{\mathbf{x}}_2 := \overline{\mathbf{x}}_1 + \alpha_2 \overline{\mathbf{p}}_2$ ($= \overline{\mathbf{x}}^\star$ exact solution).

Main idea of the conjugate gradient method

How can we generalize the previous result for larger systems?

CGM, $\mathbf{A} \in \mathbb{R}^{n\times n}$ SPD, $\overline{\mathbf{b}} \in \mathbb{R}^n$ given.

```
$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$, $\overline{\mathbf{p}}_1 = \overline{\mathbf{r}}_0$
while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ do
    $k := k + 1$
    $\alpha_k := \overline{\mathbf{p}}_k^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k)$
    $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{p}}_k$
    $\overline{\mathbf{r}}_k := \overline{\mathbf{r}}_{k-1} - \alpha_k \mathbf{A} \overline{\mathbf{p}}_k$
    $\beta_k := \overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{r}}_k / (\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k)$
    $\overline{\mathbf{p}}_{k+1} := \overline{\mathbf{r}}_k - \beta_k \overline{\mathbf{p}}_k$
end while
```

Convergence of the conjugate gradient method

Convergence of the conjugate gradient method

Thm. 44. If $\overline{\mathbf{r}}_{k-1} \neq 0$ for a given k (that is the algorithm is not terminated in the $(k-1)$th step) then

$$\overline{\mathbf{x}}_k \in \mathrm{lin}\{\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_k\} = \mathrm{lin}\{\overline{\mathbf{r}}_0, \dots, \overline{\mathbf{r}}_{k-1}\} =: V_k,$$

moreover for $k \geq 2$ we have

$$\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_j = 0, \ \ j = 0, \dots, k-2,$$

and

$$\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_j = 0, \ \ j = 1, \dots, k-1.$$

Proof. Tedious proof with induction. ■

Convergence of the conjugate gradient method

Rmk. It can be shown that $\overline{\mathbf{p}}_k^T \overline{\mathbf{r}}_{k-1} = \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1}$, thus

$$\alpha_k = \frac{\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1}}{\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k}.$$

Moreover

$$\overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k = \overline{\mathbf{r}}_k^T (\overline{\mathbf{r}}_{k-1} - \alpha_k \mathbf{A} \overline{\mathbf{p}}_k) = -\alpha_k \overline{\mathbf{r}}_k^T \mathbf{A} \overline{\mathbf{p}}_k,$$

and

$$\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} = \overline{\mathbf{r}}_{k-1}^T (\overline{\mathbf{r}}_k + \alpha_k \mathbf{A} \overline{\mathbf{p}}_k) = \alpha_k \overline{\mathbf{r}}_{k-1}^T \mathbf{A} \overline{\mathbf{p}}_k = \alpha_k \overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k,$$

so

$$\beta_k = -\frac{\overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k}{\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1}}.$$

Convergence of the conjugate gradient method

The final algorithm. With the previous results the algorithm can be simplified as follows:

CGM, $\mathbf{A} \in \mathbb{R}^{n\times n}$ SPD, $\overline{\mathbf{b}} \in \mathbb{R}^n$ given.

```
$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$, $\overline{\mathbf{p}}_1 = \overline{\mathbf{r}}_0$
while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ do
    $k := k + 1$
    $\alpha_k := \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1} / (\overline{\mathbf{p}}_k^T \mathbf{A} \overline{\mathbf{p}}_k)$ ($2n - 1 + 2n^2 + n - 1$ flop)
    $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{p}}_k$ ($2n$ flop)
    $\overline{\mathbf{r}}_k := \overline{\mathbf{r}}_{k-1} - \alpha_k \mathbf{A} \overline{\mathbf{p}}_k$ ($2n$ flop)
    $\beta_k' := \overline{\mathbf{r}}_k^T \overline{\mathbf{r}}_k / (\overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{r}}_{k-1})$ ($2n - 1$ flop)
    $\overline{\mathbf{p}}_{k+1} := \overline{\mathbf{r}}_k + \beta_k' \overline{\mathbf{p}}_k$ ($2n$ flop)
end while
```

Rmk. The number of operation is $2n^2 + 11n - 3 = 2n^2 + O(n)$ per iteration. If it need more than $n/3$ steps then the Gauss method is faster.

Convergence of the conjugate gradient method

Def. 45. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a SPD matrix. We define the $\mathbf{A}$-norm of a vector $\overline{\mathbf{x}} \in \mathbb{R}^n$ as $\|\overline{\mathbf{x}}\|_{\mathbf{A}} = \sqrt{\overline{\mathbf{x}}^T \mathbf{A} \overline{\mathbf{x}}}$.

Let us introduce the notation $\overline{\mathbf{e}}_k = \overline{\mathbf{x}}_k - \overline{\mathbf{x}}^\star$.

Thm. 46. If $\overline{\mathbf{r}}_{k-1} \neq \mathbf{0}$ then $\overline{\mathbf{x}}_k$ is the unique point in V_k that minimizes $\|\overline{\mathbf{e}}_k\|_{\mathbf{A}}$.

$$\|\overline{\mathbf{e}}_1\|_{\mathbf{A}} \geq \|\overline{\mathbf{e}}_2\|_{\mathbf{A}} \geq \cdots \geq \|\overline{\mathbf{e}}_k\|_{\mathbf{A}},$$

moreover $\overline{\mathbf{e}}_k = \mathbf{0}$ for some index $k \leq n$.

Proof. Consider the vectors $\overline{\mathbf{x}}_k - \overline{\mathbf{x}}^\star + \Delta\overline{\mathbf{x}} = \overline{\mathbf{e}}_k + \Delta\overline{\mathbf{x}}$, where $\Delta\overline{\mathbf{x}}$ is an arbitrary vector in V_k.

Convergence of the conjugate gradient method

$$\|\overline{\mathbf{e}}_k + \Delta\overline{\mathbf{x}}\|_{\mathbf{A}}^2 = (\overline{\mathbf{e}}_k + \Delta\overline{\mathbf{x}})^T\mathbf{A}(\overline{\mathbf{e}}_k + \Delta\overline{\mathbf{x}})$$

$$= (\overline{\mathbf{e}}_k)^T\mathbf{A}\overline{\mathbf{e}}_k + \Delta\overline{\mathbf{x}}^T\mathbf{A}\Delta\overline{\mathbf{x}} + 2\cdot\overbrace{\underbrace{(\overline{\mathbf{e}}_k)^T\mathbf{A}}_{=(\mathbf{A}\overline{\mathbf{e}}_k)^T=(\mathbf{A}\overline{\mathbf{x}}_k-\overline{\mathbf{b}})^T=-\overline{\mathbf{r}}_k^T}\cdot\underbrace{\Delta\overline{\mathbf{x}}}_{\in V_k}}^{=0}$$

$$= (\overline{\mathbf{e}}_k)^T\mathbf{A}\overline{\mathbf{e}}_k + \Delta\overline{\mathbf{x}}^T\mathbf{A}\Delta\overline{\mathbf{x}}.$$

It can be seen that $\Delta\overline{\mathbf{x}} = \mathbf{0}$ minimizes the norm, that is $\overline{\mathbf{x}}_k$ is the best approximation in $\mathbf{A}$-norm of the solution V_k.

It follows from the inclusion $V_1 \subset V_2 \subset \cdots \subset V_k$ that the $\mathbf{A}$-norm of the error vector decreases monotonically.

If the procedure has not been terminated earlier then $V_n = \mathbb{R}^n$, since the vectors $\overline{\mathbf{r}}_k$ are orthogonal. Moreover $\overline{\mathbf{x}}_n$ is the best approximation in $\mathbf{A}$-norm in $\mathbb{R}^n$, that is the solution is $\overline{\mathbf{x}}^\star$ itself. ■

Convergence of the conjugate gradient method

Thm. 47. Let $\mathbf{A}$ be an SPD matrix with condition number $\kappa(\mathbf{A})$. Then it is true the error estimate

$$\|\overline{\mathbf{e}}^{(k)}\|_{\mathbf{A}} \leq 2\left(\frac{\sqrt{\kappa_2(\mathbf{A})}-1}{\sqrt{\kappa_2(\mathbf{A})}+1}\right)^k \|\overline{\mathbf{e}}^{(0)}\|_{\mathbf{A}}.$$

Thm. 48. If the matrix $\mathbf{A}$ has s distinct eigenvalues then the conjugate gradient method delivers the solution at least in s steps.

Rmk. The method is efficient if

- $\mathbf{A}$ is well-conditioned,
- $\mathbf{A}$ has only few distinct eigenvalues.

Remarks

Some remarks

- Because of the relatively large number of operations, the method is used for sparse matrices. We do not need to find optimal relaxation parameters unlike in the case of the SOR method.
- In exact arithmetic, CGM is a direct method but it is an iterative method in practice due to rounding errors.
- When we terminate the iteration after the kth step then we obtain the best approximation in $\mathbf{A}$-norm in the subspace V_k.

Some remarks

- Preconditioning: Let $\mathbf{C}$ be an SPD matrix such that $\mathbf{C}^2 \approx \mathbf{A}$ and $\mathbf{C}^2$ can be invert easily. Let us consider the system

  $$\underbrace{(\mathbf{C}^{-1}\mathbf{A}\mathbf{C}^{-1})}_{\tilde{\mathbf{A}}}\underbrace{(\mathbf{C}\overline{\mathbf{x}})}_{\tilde{\overline{\mathbf{x}}}} = \underbrace{\mathbf{C}^{-1}\overline{\mathbf{b}}}_{\tilde{\overline{\mathbf{b}}}}.$$

  If we solve this system with the CG method then, albeit we have to solve a system with the coefficient matrix $\mathbf{C}^2$ in each step, the method converges quickly due to the well-conditioned coefficient matrix.

Some remarks

The full algorithm with the notation $\overline{\mathbf{z}}_k = \mathbf{C}^{-2}\overline{\mathbf{r}}_k$.

Preconditioned CGM.

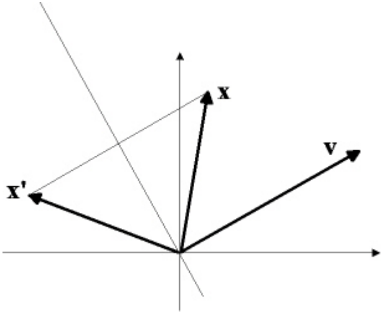
```
$k := 0$, $\overline{\mathbf{r}}_0 := \overline{\mathbf{b}}$, $\overline{\mathbf{x}}_0 := \mathbf{0}$, solution of $\mathbf{C}^2\overline{\mathbf{p}}_1 = \overline{\mathbf{r}}_0$, $\overline{\mathbf{z}}_0 = \overline{\mathbf{p}}_1$.
while $\overline{\mathbf{r}}_k \neq \mathbf{0}$ do
    $k := k + 1$
    $\alpha_k := \overline{\mathbf{r}}_{k-1}^T \overline{\mathbf{z}}_{k-1} / \overline{\mathbf{p}}_k^T \overline{\mathbf{p}}_k$
    $\overline{\mathbf{x}}_k := \overline{\mathbf{x}}_{k-1} + \alpha_k \overline{\mathbf{p}}_k$
    $\overline{\mathbf{r}}_k := \overline{\mathbf{r}}_{k-1} - \alpha_k \mathbf{A} \overline{\mathbf{p}}_k$
    solution of $\mathbf{C}^2\overline{\mathbf{z}}_k = \overline{\mathbf{r}}_k$
    $\beta'_k := \overline{\mathbf{r}}_k^T \overline{\mathbf{z}}_k / (\overline{\mathbf{r}}_{k-1})^T \overline{\mathbf{z}}_{k-1}$
    $\overline{\mathbf{p}}_{k+1} := \overline{\mathbf{z}}_k + \beta'_k \overline{\mathbf{p}}_k$
end while
```

QR decomposition

Householder reflection

Householder reflection

How can we give the reflection image of a vector $\overline{\mathbf{x}}$ across a line through the origin that is perpendicular to the vector $\overline{\mathbf{v}}$ in $\mathbb{R}^2$?



$$\overline{\mathbf{x}}' = \overline{\mathbf{x}} - \frac{2\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{v}} = \overline{\mathbf{x}} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}} = (\mathbf{I} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}})\overline{\mathbf{x}}.$$

Householder reflection

Let $\overline{\mathbf{v}} \in \mathbb{R}^n$ be an arbitrary nonzero vector. Then the multiplication with the matrix

$$\mathbf{H} = \mathbf{I} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}$$

reflects each vector $\overline{\mathbf{x}}$ to the plain that goes through the origin and perpendicular to the vector $\overline{\mathbf{v}}$.

Thm. 49. $\mathbf{H}$ is a symmetric and orthogonal matrix.

Proof. The symmetry is trivial.

$$\left(\mathbf{I} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right)\left(\mathbf{I} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right) = \mathbf{I} - 4\frac{\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}} + 4\frac{\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\frac{\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}} = \mathbf{I}. \blacksquare$$

Householder reflection

Question: How to choose the vector $\overline{\mathbf{v}}$ to reflect the vector $\overline{\mathbf{x}}$ to the axes x_1, that is parallel to the vector $\overline{\mathbf{e}}_1$?

$$\underbrace{\mathbf{H}\overline{\mathbf{x}}}_{\in lin(\overline{\mathbf{e}}_1)} = \overline{\mathbf{x}} - \frac{2\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{v}},$$

thus $\overline{\mathbf{v}} \in lin(\overline{\mathbf{x}}, \overline{\mathbf{e}}_1)$. Let $\overline{\mathbf{v}} = \overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1$.
Then

$$\begin{aligned}\mathbf{H}\overline{\mathbf{x}} &= \overline{\mathbf{x}} - \frac{2(\overline{\mathbf{x}}^T + \alpha\overline{\mathbf{e}}_1^T)\overline{\mathbf{x}}}{(\overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1)^T(\overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1)}(\overline{\mathbf{x}} + \alpha\overline{\mathbf{e}}_1)\\
&= \overline{\mathbf{x}} - 2\frac{\overline{\mathbf{x}}^T\overline{\mathbf{x}} + \alpha x_1}{\overline{\mathbf{x}}^T\overline{\mathbf{x}} + 2\alpha x_1 + \alpha^2}\overline{\mathbf{x}} - \alpha\frac{2\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{e}}_1\\
&= \left(1 - 2\frac{\|\overline{\mathbf{x}}\|_2^2 + \alpha x_1}{\|\overline{\mathbf{x}}\|_2^2 + 2\alpha x_1 + \alpha^2}\right)\overline{\mathbf{x}} - \alpha\frac{2\overline{\mathbf{v}}^T\overline{\mathbf{x}}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\overline{\mathbf{e}}_1.\end{aligned}$$

If $\alpha = \pm\|\overline{\mathbf{x}}\|_2$ then the coefficient of $\overline{\mathbf{x}}$ is zero.

Householder reflection

Thus, if a vector $\overline{\mathbf{x}} \neq \mathbf{0}$ is given then $\overline{\mathbf{v}} = \overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1$ is a good choice. Then

$$\mathbf{H}\overline{\mathbf{x}} = \mp\|\overline{\mathbf{x}}\|_2 \frac{2(\overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1)^T \overline{\mathbf{x}}}{(\overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1)^T (\overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1)} \overline{\mathbf{e}}_1$$

$$= \mp\|\overline{\mathbf{x}}\|_2 \frac{2\|\overline{\mathbf{x}}\|_2^2 \pm 2\|\overline{\mathbf{x}}\|_2 x_1}{2\|\overline{\mathbf{x}}\|_2^2 \pm 2\|\overline{\mathbf{x}}\|_2 x_1} \overline{\mathbf{e}}_1 = \mp\|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1.$$

Def. 50. The reflection matrix $\mathbf{H}$ that reflects a given vector $\overline{\mathbf{x}}$ through a plane that goes through the origin such a way that the reflection is on the first coordinate axes, is called Householder reflection (that belong to the vector $\overline{\mathbf{x}}$).

Application: Based on the above considerations, the Householder reflection that belongs to the vector $\overline{\mathbf{x}}$ can be determined as follows:
- We determine the normal vector of the plane of reflection: $\overline{\mathbf{v}} = \overline{\mathbf{x}} \pm \|\overline{\mathbf{x}}\|_2 \overline{\mathbf{e}}_1$,
- then we construct the reflection matrix with the vector $\overline{\mathbf{v}}$:

$$\mathbf{H} = \mathbf{I} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}.$$

Householder reflection

$$\mathbf{H}\overline{\mathbf{x}} = \mathbf{H}\begin{bmatrix} * \\ * \\ \vdots \\ * \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Rmk. If $x_1 \neq 0$ then it is practical to choose the normal vector as $\overline{\mathbf{v}} = \overline{\mathbf{x}} + \operatorname{sgn}(x_1)\|\overline{\mathbf{x}}\|_2\overline{\mathbf{e}}_1$.

Rmk. It is practical to norm the vector $\overline{\mathbf{v}}$ such that the first element of the vector will be 1. Then $\overline{\mathbf{v}}$ can be stored in the place of the eliminated elements of $\overline{\mathbf{x}}$.

Rmk. Let $\mathbf{C}$ be an arbitrary matrix. Then the calculation of $\mathbf{HC}$ can be performed as follows:

$$\mathbf{HC} = \left(\mathbf{I} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right)\mathbf{C} = \mathbf{C} - \frac{2\overline{\mathbf{v}}\overline{\mathbf{v}}^T}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\mathbf{C}$$

$$= \mathbf{C} + \overline{\mathbf{v}}\underbrace{\left(-\frac{2\overline{\mathbf{v}}^T\mathbf{C}}{\overline{\mathbf{v}}^T\overline{\mathbf{v}}}\right)}_{=:\overline{\mathbf{w}}^T} = \mathbf{C} + \overline{\mathbf{v}}\overline{\mathbf{w}}^T.$$

QR decomposition

QR decomposition

Thm. 51. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ $(m \geq n)$ be a full rank matrix. Then there exists an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{m\times m}$ and an upper triangular matrix $\mathbf{R} \in \mathbb{R}^{m\times n}$ such that $\mathbf{A} = \mathbf{QR}$.

Proof. Let $\mathbf{H}_1$ be the Householder reflection that belongs to the column $\mathbf{A}(1:m,1)$. Then the $2:m$ elements of the first column of $\mathbf{A}^{(2)} := \mathbf{H}_1\mathbf{A}$ are zero. Let $\tilde{\mathbf{H}}_2$ be the Householder reflection that belongs to the column $\mathbf{A}^{(2)}(2:m,2)$. Moreover, let $\mathbf{H}_2 = \mathrm{diag}(1, \tilde{\mathbf{H}}_2)$. Then the $2:m$ elements of the first column of $\mathbf{A}^{(3)} := \mathbf{H}_2\mathbf{A}^{(2)}$ and the $3:m$ elements of the second column are zero, etc. Based on the full rank, this procedure can be continued further. We obtain the representation

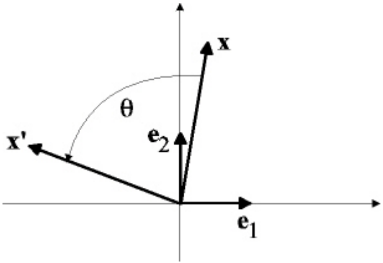
$$\mathbf{H}_n \cdot \cdots \cdot \mathbf{H}_1 \cdot \mathbf{A} = \mathbf{R},$$

where $\mathbf{R}$ is an upper triangular matrix. The matrix $\mathbf{Q}^T := \mathbf{H}_n \cdot \cdots \cdot \mathbf{H}_1$ is orthogonal, so with the above notations we have $\mathbf{A} = \mathbf{QR}$. ■

Givens rotation

Givens rotation

Rotation with angle θ in $\mathbb{R}^2$.



$$\overline{\mathbf{x}}' = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \overline{\mathbf{x}}.$$

This matrix is orthogonal. Moreover with the choice $s = \sin\theta$ and $c = \cos\theta$, the vector $[x_1, x_2]^T$ $(x_1 \neq 0)$ is transformed to the form $[*, 0]^T$.

Givens rotation

- If $x_2 = 0$ then $s = 0$, $c = 1$ is a good choice.
- If $x_2 \neq 0$ then from the solution of the SLAE $sx_1 + cx_2 = 0$, $s^2 + c^2 = 1$ we obtain the parameters

$$s = \frac{\pm x_2}{\sqrt{x_1^2 + x_2^2}}, \quad c = \frac{\mp x_1}{\sqrt{x_1^2 + x_2^2}}.$$

Generally: rotation in the hyperplane (i, j) with angle θ

$$\mathbf{G}(i, j, \theta) = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & c & & & & -s & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & s & & & & c & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{bmatrix}$$

Application of the Givens rotation

QR decomposition (schematically):

$$\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \\ 0 & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ * & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{bmatrix}$$

$$\begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \end{bmatrix}$$

Rmk. The number of operations of the Householder QR decomposition is $2n^2(m - n/3)$, while for the Givens QR decomposition we have $3n^2(m - n/3)$.

Application of Givens rotation

The QR decomposition of an upper Hessenberg matrix (schematically):

$$\begin{bmatrix} * & * & * \\ * & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \end{bmatrix}$$

Householder and Givens

,

Alston Scott Householder, 1904-1993 (USA), Wallace Givens, 1910-1993 (USA)

Solution of full rank over-determined systems

Solution of over-determined systems

Over-determined systems

$$\mathbf{A}\overline{\mathbf{x}} = \overline{\mathbf{b}}, \quad \mathbf{A} \in \mathbb{R}^{m\times n}, \; m \geq n, \; r(\mathbf{A}) = n$$

The above system generally does not have solution (or only one). Then we can search for the vector $\overline{\mathbf{x}}$ (denoted by $\overline{\mathbf{x}}_{LS}$) that minimizes the norm $\|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2$.
Let

$$\phi(\overline{\mathbf{x}}) = \|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2,$$

and let $\overline{\mathbf{z}} \in \mathbb{R}^n$ be an arbitrary vector. Because of the full column rank, $\|\mathbf{A}\overline{\mathbf{z}}\|_2 = 0$ can hold on if $\overline{\mathbf{z}} = \mathbf{0}$. Then

$$\begin{aligned}\phi(\overline{\mathbf{x}} + \overline{\mathbf{z}}) &= \|\mathbf{A}(\overline{\mathbf{x}} + \overline{\mathbf{z}}) - \overline{\mathbf{b}}\|_2^2 \\ &= \|\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}\|_2^2 + \|\mathbf{A}\overline{\mathbf{z}}\|_2^2 + 2\overline{\mathbf{z}}^T\mathbf{A}^T(\mathbf{A}\overline{\mathbf{x}} - \overline{\mathbf{b}}).\end{aligned}$$

Let $\overline{\mathbf{x}}_{LS}$ be the solution of the SLAE $\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}} = \mathbf{A}^T\overline{\mathbf{b}}$ ($\overline{\mathbf{z}}^T\mathbf{A}^T\mathbf{A}\overline{\mathbf{z}} = \|\mathbf{A}\overline{\mathbf{z}}\|_2^2 \neq 0$ provided that $\overline{\mathbf{z}} \neq \mathbf{0}$, thus $\mathbf{A}^T\mathbf{A}$ is SPD, thus it is non-singular). Then

$$\phi(\overline{\mathbf{x}}_{LS} + \overline{\mathbf{z}}) = \|\mathbf{A}\overline{\mathbf{x}}_{LS} - \overline{\mathbf{b}}\|_2^2 + \|\mathbf{A}\overline{\mathbf{z}}\|_2^2 = \phi(\overline{\mathbf{x}}_{LS}) + \|\mathbf{A}\overline{\mathbf{z}}\|_2^2,$$

that shows that $\overline{\mathbf{x}}_{LS}$ uniquely minimizes ϕ indeed.

Over-determined systems

We have to solve the so-called normal equation

$$\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}} = \mathbf{A}^T\overline{\mathbf{b}}.$$

It has unique solution due to the full rank, thus the solution can be written in the form $\overline{\mathbf{x}}_{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\overline{\mathbf{b}}$. This is not efficient in practice.

Computation of $\overline{\mathbf{x}}_{LS}$ with the normal equation

- $\mathbf{A}^T\mathbf{A}$ is SPD.
- Let us compute its Cholesky decomposition $\mathbf{L}\mathbf{L}^T$.
- Let us solve the system $\mathbf{L}\overline{\mathbf{y}} = \mathbf{A}^T\overline{\mathbf{b}}$.
- We get $\overline{\mathbf{x}}_{LS}$ as the solution of $\mathbf{L}^T\overline{\mathbf{x}} = \overline{\mathbf{y}}$.

Number of operations: $(m + n/3)n^2$ flop

Over-determined systems

Computation of $\overline{\mathbf{x}}_{LS}$ with QR decomposition

$$\begin{aligned}\|\mathbf{A}\overline{\mathbf{x}}-\overline{\mathbf{b}}\|_2^2 &= \|\mathbf{Q}\mathbf{R}\overline{\mathbf{x}}-\overline{\mathbf{b}}\|_2^2 = \|\mathbf{Q}^T(\mathbf{Q}\mathbf{R}\overline{\mathbf{x}}-\overline{\mathbf{b}})\|_2^2 \\ &= \|\mathbf{R}\overline{\mathbf{x}}-\mathbf{Q}^T\overline{\mathbf{b}}\|_2^2 = \|\mathbf{R}_1\overline{\mathbf{x}}-\overline{\mathbf{c}}\|_2^2+\|\overline{\mathbf{d}}\|_2^2,\end{aligned}$$

where $\mathbf{R}_1=\mathbf{R}(1:n,1:n)$, $\overline{\mathbf{c}}=(\mathbf{Q}^T\overline{\mathbf{b}})(1:n,:)$, $\overline{\mathbf{d}}=(\mathbf{Q}^T\overline{\mathbf{b}})(n+1:m,:)$.

- Compute the QR decomposition of $\mathbf{A}$.
- Determine the matrix $\mathbf{R}_1=\mathbf{R}(1:n,1:n)$.
- Determine the vector $\overline{\mathbf{c}}=(\mathbf{Q}^T\overline{\mathbf{b}})(1:n,:)$.
- $\overline{\mathbf{x}}_{LS}$ is the solution of the SLAE $\mathbf{R}_1\overline{\mathbf{x}}=\overline{\mathbf{c}}$.

Number of operations: $2(m-n/3)n^2$ flop

Over-determined systems

Rmk.

- If $m >> n$ then the number of operations of the solution with the QR decomposition is approximately the double of that of the other.
- For quadratic full rank matrices, the number of operations is the same in both cases: $4n^3/3$, which is the double of that of the Gauss method. When we take into the account also the memory usage, then the total solution time may be comparable with that of the Gauss method, moreover, in this case there is no growth factor, that is the method is stable.
- We cannot use these methods for (nearly) rank deficient matrices.
- For the normal equation, we can use the CG method but the condition number of the new system will be the square of that of the original system.

Eigenvalue problems

Conditioning

Conditioning

Thm. 52. [Bauer-Fike, 1960] Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a diagonalizable matrix ($\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^{-1}$). Let μ be an eigenvalue of the matrix $\mathbf{A} + \boldsymbol{\delta}\mathbf{A}$. Then it is valid the estimation

$$\min_{\lambda \text{ eigenv. of } \mathbf{A}} |\lambda - \mu| \leq \kappa_p(\mathbf{V})\|\boldsymbol{\delta}\mathbf{A}\|_p.$$

Friedrich Ludwig Bauer (1923-, German)

Conditioning

Proof. If μ is an eigenvalue of $\mathbf{A}$, then the statement is trivial. Otherwise, because $\mathbf{A} + \boldsymbol{\delta}\mathbf{A} - \mu\mathbf{I}$ is singular, the matrix

$$\mathbf{V}^{-1}(\mathbf{A} + \boldsymbol{\delta}\mathbf{A} - \mu\mathbf{I})\mathbf{V} = \mathbf{D} + \mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V} - \mu\mathbf{I}$$

is also singular. Thus, there is a vector $\overline{\mathbf{x}} \neq \mathbf{0}$ (an eigenvector) such that

$$(\mathbf{D} - \mu\mathbf{I} + \mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V})\overline{\mathbf{x}} = \mathbf{0} \quad /(\mathbf{D} - \mu\mathbf{I})^{-1} \cdot .$$

Conditioning

$$(\mathbf{I} + (\mathbf{D} - \mu\mathbf{I})^{-1}\mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V})\overline{\mathbf{x}} = \mathbf{0},$$

that is

$$\overline{\mathbf{x}} = -(\mathbf{D} - \mu\mathbf{I})^{-1}\mathbf{V}^{-1}\boldsymbol{\delta}\mathbf{A}\mathbf{V}\overline{\mathbf{x}}.$$

Hence

$$\|\overline{\mathbf{x}}\|_p \leq \|(\mathbf{D} - \mu\mathbf{I})^{-1}\|_p\|\mathbf{V}^{-1}\|_p\|\boldsymbol{\delta}\mathbf{A}\|_p\|\mathbf{V}\|_p\|\overline{\mathbf{x}}\|_p$$

and

$$1 \leq \|(\mathbf{D} - \mu\mathbf{I})^{-1}\|_p\kappa_p(\mathbf{V})\|\boldsymbol{\delta}\mathbf{A}\|_p$$

$$= \max_i \frac{1}{|\lambda_i - \mu|}\kappa_p(\mathbf{V})\|\boldsymbol{\delta}\mathbf{A}\|_p = \frac{1}{\min_i |\lambda_i - \mu|}\kappa_p(\mathbf{V})\|\boldsymbol{\delta}\mathbf{A}\|_p.$$

From this the statement follows already. ■

Conditioning

Rmk. Thus the condition of the eigenvalue problem is determined by the condition number of the diagonalizing matrix.

Rmk. The Hilbert matrix $\mathbf{H}_n$ is symmetric, thus it can be diagonalized with an orthogonal matrix. The 2-norm of orthogonal matrices is 1, hence

$$\min_{\lambda \text{ eigenv. of } \mathbf{H}} |\lambda - \mu| \leq \|\boldsymbol{\delta}\mathbf{H}_n\|_2.$$

The Hilbert matrix behaves badly in case of the solution of a SLAE, but it behaves well in case of an eigenvalue problem.

Conditioning

Example. The eigenvalue change of a non-symmetric, non-diagonalizable matrix. Let

$$\mathbf{A} = \begin{bmatrix} 0 & & & & \varepsilon \\ 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix} \in \mathbb{R}^{40\times 40}.$$

The characteristic polynomial of the matrix is $\lambda^{40} - \varepsilon$. Thus if $\varepsilon = 0$, then all the eigenvalues are zeros. If $\varepsilon > 0$, then there is a real eigenvalue $\sqrt[40]{\varepsilon}$ and the other 39 eigenvalues are complex.

If ε changes from 0 to 10^{-40}, then the eigenvalue changes from 0 to 0.1. Thus, the change in the eigenvalue is $10^{39}\varepsilon$.

The power method

The idea of the power method

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a normal matrix, and let us suppose that $\mathbf{A}$ have a strictly dominant eigenvalue, that is

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

Then the eigenvalue $\lambda_1 \in \mathbb{R}$ and the corresponding eigenvector $\overline{\mathbf{v}}_1$ can be chosen to be real. Let $\overline{\mathbf{v}}_1, \ldots, \overline{\mathbf{v}}_n$ be the normed eigenvectors, and because $\mathbf{A}$ is normal, they form an orthonormal basis. Let $\overline{\mathbf{x}} \in \mathbb{R}^n$ be such that $\alpha_1 \neq 0$ $(\alpha_1 \in \mathbb{R})$ is not zero in the form $\overline{\mathbf{x}} = \alpha_1\overline{\mathbf{v}}_1 + \alpha_2\overline{\mathbf{v}}_2 + \cdots + \alpha_n\overline{\mathbf{v}}_n$.

Then

$$\begin{aligned}\mathbf{A}^k\overline{\mathbf{x}} &= \alpha_1\lambda_1^k\overline{\mathbf{v}}_1 + \alpha_2\lambda_2^k\overline{\mathbf{v}}_2 + \cdots + \alpha_n\lambda_n^k\overline{\mathbf{v}}_n \\ &= \lambda_1^k\left(\alpha_1\overline{\mathbf{v}}_1 + \underbrace{\alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\overline{\mathbf{v}}_2}_{\to 0} + \cdots + \underbrace{\alpha_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\overline{\mathbf{v}}_n}_{\to 0}\right).\end{aligned}$$

The power method

```
The power method, $\overline{\mathbf{v}}_1^T\overline{\mathbf{y}}^{(0)} \neq 0$, $\|\overline{\mathbf{y}}^{(0)}\|_2 = 1$
  for $k := 1 : k_{\max}$ do
    $\overline{\mathbf{x}}^{(k)} := \mathbf{A}\overline{\mathbf{y}}^{(k-1)}$
    $\overline{\mathbf{y}}^{(k)} := \overline{\mathbf{x}}^{(k)}/\|\overline{\mathbf{x}}^{(k)}\|_2$
    $\nu^{(k)} := (\overline{\mathbf{y}}^{(k)})^T\mathbf{A}\overline{\mathbf{y}}^{(k)}$
  end for
```

Thm. 53.

$$\overline{\mathbf{y}}^{(k)} = \frac{\mathbf{A}^k\overline{\mathbf{y}}^{(0)}}{\|\mathbf{A}^k\overline{\mathbf{y}}^{(0)}\|_2},$$

$\nu^{(k)} \to \lambda_1$, moreover there exists a sequence $\{\gamma_k\} \subset \mathbb{R}$ such that $|\gamma_k| = 1$ $(k = 1, \dots)$ and

$$\gamma_k\overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1.$$

The power method

Proof.

The first part can be proven with induction.

Parseval's inequality: $\|\overline{\mathbf{x}}\|_2 = \sqrt{\sum_{i=1}^{n} |\alpha_i|^2}$.

Namely:

$$\overline{\mathbf{x}}^H \overline{\mathbf{x}} = \left(\sum_{i=1}^{n} \overline{\alpha}_i \overline{\mathbf{v}}_i^H \right) \left(\sum_{i=1}^{n} \alpha_i \overline{\mathbf{v}}_i \right) = \sum_{i=1}^{n} |\alpha_i|^2.$$

Let $\overline{\mathbf{y}}^{(0)} = \alpha_1 \overline{\mathbf{v}}_1 + \alpha_2 \overline{\mathbf{v}}_2 + \cdots + \alpha_n \overline{\mathbf{v}}_n$ and we know that $\alpha_1 \neq 0$. Hence

$$\overline{\mathbf{y}}^{(k)} = \frac{\lambda_1^k \left(\alpha_1 \overline{\mathbf{v}}_1 + \alpha_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k \overline{\mathbf{v}}_2 + \cdots + \alpha_n \left(\frac{\lambda_n}{\lambda_1} \right)^k \overline{\mathbf{v}}_n \right)}{\sqrt{\sum_{i=1}^{n} |\alpha_i|^2 |\lambda_i|^{2k}}}$$

$$= \frac{\lambda_1^k \alpha_1 \left(\overline{\mathbf{v}}_1 + \frac{\alpha_2}{\alpha_1} \left(\frac{\lambda_2}{\lambda_1} \right)^k \overline{\mathbf{v}}_2 + \cdots + \frac{\alpha_n}{\alpha_1} \left(\frac{\lambda_n}{\lambda_1} \right)^k \overline{\mathbf{v}}_n \right)}{|\lambda_1|^k |\alpha_1| \sqrt{1 + \sum_{i=2}^{n} |\frac{\alpha_i}{\alpha_1}|^2 |\frac{\lambda_i}{\lambda_1}|^{2k}}}.$$

The power method

Thus

$$\overbrace{\frac{|\lambda_1|^k|\alpha_1|}{\lambda_1^k\alpha_1}}^{=:\gamma_k}\overline{\mathbf{y}}^{(k)}$$

$$= \frac{\left(\overline{\mathbf{v}}_1 + \frac{\alpha_2}{\alpha_1}\left(\frac{\lambda_2}{\lambda_1}\right)^k\overline{\mathbf{v}}_2 + \cdots + \frac{\alpha_n}{\alpha_1}\left(\frac{\lambda_n}{\lambda_1}\right)^k\overline{\mathbf{v}}_n\right)}{\sqrt{1+\sum_{i=2}^n|\frac{\alpha_i}{\alpha_1}|^2||\frac{\lambda_i}{\lambda_1}|^{2k}}} \to \overline{\mathbf{v}}_1,$$

where $|\gamma_k| = 1$ $(k = 1, \dots)$.

$$\begin{aligned} 0 \leftarrow (\gamma_k\overline{\mathbf{y}}^{(k)})^T\mathbf{A}(\gamma_k\overline{\mathbf{y}}^{(k)}) - \overline{\mathbf{v}}_1^T\mathbf{A}\overline{\mathbf{v}}_1 &= |\gamma_k|^2(\overline{\mathbf{y}}^{(k)})^T\mathbf{A}\overline{\mathbf{y}}^{(k)} - \lambda_1 \\ &= (\overline{\mathbf{y}}^{(k)})^T\mathbf{A}\overline{\mathbf{y}}^{(k)} - \lambda_1 = \nu^{(k)} - \lambda_1. \blacksquare \end{aligned}$$

The power method

Rmk.

- If $\lambda_1, \alpha_1 > 0$, then $\overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.
- If $\lambda_1 > 0, \alpha_1 < 0$, then $-\overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.
- If $\lambda_1 < 0, \alpha_1 > 0$, then $(-1)^k \overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.
- If $\lambda_1 < 0, \alpha_1 < 0$, then $(-1)^{k+1} \overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$.

Rmk. Let $\overline{\mathbf{e}}^{(k)} = \overline{\mathbf{y}}^{(k)} - \overline{\mathbf{v}}_1$ be the error of the kth iteration vector. Then, for sufficiently large values k we have $\|\overline{\mathbf{e}}^{(k+1)}\|_2 \approx |\lambda_2/\lambda_1| \|\overline{\mathbf{e}}^{(k)}\|_2$ (linear convergence).

Rmk. If $\overline{\mathbf{x}}$ is an approximation of the eigenvector that belongs to the dominant eigenvalue of $\mathbf{A}$, then we have $\overline{\mathbf{x}}^T(\mathbf{A}\overline{\mathbf{x}}) \approx \overline{\mathbf{x}}^T(\lambda\overline{\mathbf{x}})$ and

$$\lambda \approx \frac{\overline{\mathbf{x}}^T \mathbf{A} \overline{\mathbf{x}}}{\overline{\mathbf{x}}^T \overline{\mathbf{x}}}$$

is an approximation of the eigenvalue.

Rayleigh’s coefficient

Rayleigh's coefficient

Def. 54. Let $\mathbf{0} \neq \overline{\mathbf{x}} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$. The number

$$R(\overline{\mathbf{x}}) = \frac{\overline{\mathbf{x}}^T \mathbf{A} \overline{\mathbf{x}}}{\overline{\mathbf{x}}^T \overline{\mathbf{x}}}$$

is called the Rayleigh's coefficient to the vector $\overline{\mathbf{x}}$.

Thm. 55. Let the $\mathbf{0} \neq \overline{\mathbf{x}} \in \mathbb{R}^n$ be a given vector. Then

$$\min_{\alpha\in\mathbb{R}} \|\mathbf{A}\overline{\mathbf{x}} - \alpha\overline{\mathbf{x}}\|_2 = \|\mathbf{A}\overline{\mathbf{x}} - R(\overline{\mathbf{x}})\overline{\mathbf{x}}\|_2.$$

Proof.

$$\begin{aligned} \|\mathbf{A}\overline{\mathbf{x}} - \alpha\overline{\mathbf{x}}\|_2^2 &= (\overline{\mathbf{x}}^T\mathbf{A}^T - \alpha\overline{\mathbf{x}}^T)(\mathbf{A}\overline{\mathbf{x}} - \alpha\overline{\mathbf{x}}) \\ &= \overline{\mathbf{x}}^T\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}} - 2\alpha\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}} + \alpha^2\overline{\mathbf{x}}^T\overline{\mathbf{x}} \\ &= \alpha^2\overline{\mathbf{x}}^T\overline{\mathbf{x}} - 2\alpha\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}} + \overline{\mathbf{x}}^T\mathbf{A}^T\mathbf{A}\overline{\mathbf{x}}. \end{aligned}$$

Rayleigh's coefficient

Because $\overline{\mathbf{x}}^T\overline{\mathbf{x}} > 0$ if $\overline{\mathbf{x}} \neq \mathbf{0}$, hence the function takes its minimum az

$$\alpha_{\min} = \frac{\overline{\mathbf{x}}^T\mathbf{A}\overline{\mathbf{x}}}{\overline{\mathbf{x}}^T\overline{\mathbf{x}}} = R(\overline{\mathbf{x}}). \blacksquare$$

Rmk. For symmetric matrices

$$\lambda_{\min} \leq R(\overline{\mathbf{x}}) \leq \lambda_{\max}.$$

Rmk. For symmetric matrices

$$\lambda_{\max} = \max_{\overline{\mathbf{x}} \in \mathbb{R}^n \neq 0} R(\overline{\mathbf{x}}), \quad \lambda_{\min} = \min_{\overline{\mathbf{x}} \in \mathbb{R}^n \neq 0} R(\overline{\mathbf{x}})$$

(Courant-Fischer theorem).

From now on, we will consider only symmetric matrices in the eigenvalue problems!

Inverse iteration

Inverse iteration

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a non-singular symmetric matrix with the eigenvalues λ_i and with the eigenvectors $\overline{\mathbf{v}}_i$. Then, if $\mu \neq \lambda_i$, then the matrix $\mathbf{A} - \mu\mathbf{I}$ is invertible and the eigenvectors of $(\mathbf{A} - \mu\mathbf{I})^{-1}$ are identical with those of $\mathbf{A}$, its eigenvalues are $(\lambda_i - \mu)^{-1}$.

If the number μ is sufficiently close to λ_j, then the dominant eigenvalue will be $(\lambda_j - \mu)^{-1}$, thus executing the power method with the matrix $(\mathbf{A} - \mu\mathbf{I})^{-1}$, λ_j and $\overline{\mathbf{v}}_j$ can be approximated.

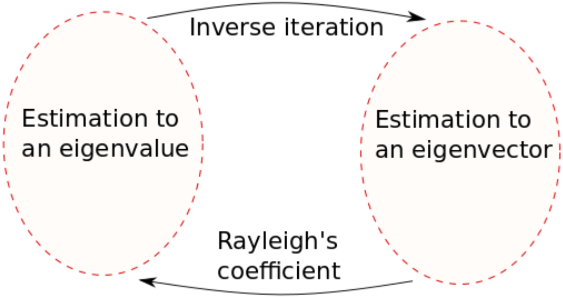
Inverse iteration

```
Inverse iteration, $\overline{\mathbf{v}}_1^T\overline{\mathbf{y}}^{(0)} \neq 0$, $\|\overline{\mathbf{y}}^{(0)}\|_2 = 1$
    for $k := 1 : k_{\max}$ do
        $\overline{\mathbf{x}}^{(k)} := (\mathbf{A} - \mu\mathbf{I})^{-1}\overline{\mathbf{y}}^{(k-1)}$
        (solution of $(\mathbf{A} - \mu\mathbf{I})\overline{\mathbf{x}}^{(k)} = \overline{\mathbf{y}}^{(k-1)}$)
        $\overline{\mathbf{y}}^{(k)} := \overline{\mathbf{x}}^{(k)}/\|\overline{\mathbf{x}}^{(k)}\|_2$
        $\nu^{(k)} := (\overline{\mathbf{y}}^{(k)})^T\mathbf{A}\overline{\mathbf{y}}^{(k)}$
    end for
```

Rmk.

- First we compute the LU-decomposition of the matrix $\mathbf{A} - \mu\mathbf{I}$. This makes possible to solve the system with $2n^2$ flops in each iteration.
- Much more expensive than the power method, but it can converge to any eigenvalue.
- The condition $\overline{\mathbf{v}}_1^T\overline{\mathbf{y}}^{(0)} \neq 0$ is not too strict. If it does not hold initially, then it will be satisfied after sufficiently large number of iterations due to the rounding errors. Thus, the method will converge in this case, too.

Approximation of eigenvalues and eigenvectors



Inverse iteration

Estimation to an eigenvalue

Estimation to an eigenvector

Rayleigh's coefficient

Rayleigh quotient iteration

Rayleigh quotient iteration

Let us use Rayleigh's quotient in the approximation of the eigenvalue! If $\overline{\mathbf{y}}^{(k)} \to \overline{\mathbf{v}}_1$, then $R(\overline{\mathbf{y}}^{(k)}) \to \lambda_j$.

Rayleigh quotient it., $\overline{\mathbf{v}}_1^T \overline{\mathbf{y}}^{(0)} \neq 0$, $\|\overline{\mathbf{y}}^{(0)}\|_2 = 1$

```
for $k := 1 : k_{\max}$ do
    compute $R(\overline{\mathbf{y}}^{(k-1)})$
    solution of $(\mathbf{A} - R(\overline{\mathbf{y}}^{(k-1)})\mathbf{I})\overline{\mathbf{x}}^{(k)} = \overline{\mathbf{y}}^{(k-1)}$
    $\overline{\mathbf{y}}^{(k)} := \overline{\mathbf{x}}^{(k)} / \|\overline{\mathbf{x}}^{(k)}\|_2$
end for
```

Rmk. We have to solve a new SLAE in every step.

Rmk. The convergence is of order 3.

Householder's deflation

Householder's deflation

- Let us suppose that we have already determined the strictly dominant eigenvalue λ_1 and the corresponding eigenvector $\overline{\mathbf{v}}_1$ to the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.
- Let us compute the Householder matrix for which we have $\mathbf{H}\overline{\mathbf{v}}_1 = \alpha\overline{\mathbf{e}}_1$ $(\alpha \neq 0)$.
- Then

$$\begin{aligned}\mathbf{HAH}\overline{\mathbf{e}}_1 &= \frac{1}{\alpha}\mathbf{HAH}\alpha\overline{\mathbf{e}}_1 = \frac{1}{\alpha}\mathbf{HAHH}\overline{\mathbf{v}}_1 = \frac{1}{\alpha}\mathbf{HA}\overline{\mathbf{v}}_1 \\ &= \frac{1}{\alpha}\mathbf{H}\lambda_1\overline{\mathbf{v}}_1 = \frac{1}{\alpha}\lambda_1\mathbf{H}\overline{\mathbf{v}}_1 = \frac{1}{\alpha}\lambda_1\alpha\overline{\mathbf{e}}_1 = \lambda_1\overline{\mathbf{e}}_1.\end{aligned}$$

- Thus

$$\mathbf{HAH} = \begin{bmatrix} \lambda_1 & \overline{\mathbf{b}}^T \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}.$$

Householder's deflation

- The eigenvalues of $\mathbf{A}_2$ equal the eigenvalues of $\mathbf{A}$ except for the eigenvalue λ_1. If $|\lambda_2| > |\lambda_3|$, then when we execute the power method with the matrix $\mathbf{A}_2$, we can compute the approximation of the eigenvalue λ_2: $\tilde{\lambda}_2$.
- Execute the inverse iteration with the matrix $(\mathbf{A} - \tilde{\lambda}_2\mathbf{I})^{-1}$. This result in the eigenvector $\overline{\mathbf{v}}_2$.
- Because

  $$\mathbf{HAH}(\mathbf{H}\overline{\mathbf{v}}_2) = \mathbf{HA}\overline{\mathbf{v}}_2 = \lambda_2(\mathbf{H}\overline{\mathbf{v}}_2),$$

  the vector $\mathbf{H}\overline{\mathbf{v}}_2$ is an eigenvector of $\mathbf{HAH}$ with the eigenvalue λ_2. Thus $\mathbf{H}\overline{\mathbf{v}}_2(2:n)$ is the eigenvector of $\mathbf{A}_2$ with the dominant eigenvalue λ_2.
- In a similar way, we perform a similar procedure with matrix $\mathbf{A}_2$ instead of the matrix $\mathbf{A}$.

Rank deflation

Rank deflation

- Let us suppose that we have computed already the strictly dominant eigenvalue λ_1 and the corresponding eigenvector $\overline{\mathbf{v}}_1$ of the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$.
- Let us consider the matrix $\mathbf{A} - \lambda_1 \overline{\mathbf{v}}_1 \overline{\mathbf{v}}_1^T$. The eigenvalues of this matrix equal the eigenvalues of $\mathbf{A}$, with the only difference that zero stands instead of λ_1. The eigenvectors are the same.
- When λ_2 is strictly dominant, then executing the power method with the above matrix, we can obtain λ_2 and $\overline{\mathbf{v}}_2$.

QR-iteráció

QR-iteráció

Main idea: If we could find a matrix $\mathbf{V}$ to the matrix $\mathbf{A}$ such that $\mathbf{V}^{-1}\mathbf{A}\mathbf{V}$ is an upper triangular matrix, then the diagonal of this upper triangular matrix would contain the eigenvalues of the matrix. Unfortunately such a matrix $\mathbf{V}$ cannot be constructed directly.

Let us approximate this matrix with the orthogonal matrices of the QR decomposition.

QR iteration, $\mathbf{A}$ is a given symmetric matrix, $\mathbf{A}^{(0)} := \mathbf{A}$

```
for $k := 1 : k_{\max}$ do
    Construct the $QR$ decomposition of $\mathbf{A}^{(k-1)}$: $\mathbf{A}^{(k-1)} = \mathbf{Q}^{(k-1)}\mathbf{R}^{(k-1)}$
    $\mathbf{A}^{(k)} := (\mathbf{Q}^{(k-1)})^T\mathbf{A}^{(k-1)}\mathbf{Q}^{(k-1)}$
end for
```

QR iteration

Thus

$$\mathbf{A}^{(k)} = (\mathbf{Q}^{(k-1)})^T \dots (\mathbf{Q}^{(0)})^T \mathbf{A}\, \underbrace{\mathbf{Q}^{(0)} \dots \mathbf{Q}^{(k-1)}}_{=:\mathbf{Q}_k} = \mathbf{Q}_k^T \mathbf{A} \mathbf{Q}_k,$$

and the eigenvalues of $\mathbf{A}^{(k)}$ will be the same as the eigenvalues of $\mathbf{A}$.

Thm. 56. a) If all the eigenvalues of $\mathbf{A}$ are real and different in absolute values, then the matrix sequence $\{\mathbf{A}^{(k)}\}$ tends to an upper triangular matrix.
b) If all the eigenvalues of a symmetric matrix $\mathbf{A}$ are different in absolute values, then the matrix sequence $\{\mathbf{A}^{(k)}\}$ tends to a diagonal matrix.

Rmk. In both cases the eigenvalues appear in the diagonal of the limit matrix.

Remarks

Rmk. Let $\mathbf{A} = \mathbf{QR}$ be an upper triangular matrix. Then the matrix

$$\mathbf{A}^{(1)} = \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{Q}^T\mathbf{QRQ} = \mathbf{RQ} = \mathbf{RQRR}^{-1} = \mathbf{RAR}^{-1}$$

is also upper triangular.

Rmk. Every QR decomposition is $4n^3/3$ flops, thus the method converges very slowly. The solution for this can be the conversion of the original matrix to Hessenberg form, e.g. with Householder reflections ($4n^3/3$ flop, the eigenvalues do not change): $\mathbf{A} \to \mathbf{H}_1\mathbf{A}\mathbf{H}_1 \to \mathbf{H}_2\mathbf{H}_1\mathbf{A}\mathbf{H}_1\mathbf{H}_2$, etc., schematically

$$\begin{bmatrix} * & * & * & * \\ {\color{red}*} & * & * & * \\ {\color{red}*} & * & * & * \\ {\color{red}*} & * & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ 0 & {\color{red}*} & * & * \\ 0 & {\color{red}*} & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \end{bmatrix}$$

For Hessenberg matrices, the QR decomposition can be performed with Givens rotations very fast ($3n^2$ flop).

Rmk. For symmetric matrices the Hessenberg form will be tridiagonal.

Summary of some main concepts

- Normed spaces (norms, normed spaces, equivalence of norms, Banach spaces, Banach fixed point theorem)
- Vector and matrix norms
- Euclidean spaces (scalar product, euclidean space, orthogonality, Gram–Schmidt orthogonalization, orthogonal polynomials)
- Special properties of matrices
- Eigenvalues and eigenvectors of matrices
- Diagonalizability of matrices

Normed spaces

Vector space (linear space)

Def. 57. A set $V \neq \emptyset$ is called (real) vector space, if an addition and a multiplication with scalar operation is defined on it with the properties:

1. $x + y = y + x$, $\forall a, b \in V$;
2. $(x + y) + z = x + (y + z)$, $\forall x, y, z \in V$;
3. $\exists o \in V$, $x + o = x$, $\forall x \in V$;
4. $\forall x \in V$, $\exists \hat{x} \in V$, $x + \hat{x} = o$;
5. $1 \cdot x = x$, $\forall x \in V$;
6. $\alpha(x + y) = \alpha x + \alpha y$, $\forall x, y \in V$, $\forall \alpha \in \mathbb{R}$;
7. $(\alpha + \beta)x = \alpha x + \beta x$, $\forall x \in V$, $\forall \alpha, \beta \in \mathbb{R}$;
8. $\alpha(\beta x) = (\alpha\beta)x$, $\forall x \in V$, $\forall \alpha, \beta \in \mathbb{R}$.

Ex.: Vectors on the plane and in space, $\mathbb{R}^n$, $\mathbb{R}^{m \times n}$, $C[a, b]$, P_n etc. with the usual operations.

Special vector systems in vector spaces

Def. 58. A vector $x \in V$ is called the linear combination of the vectors $x_1, \ldots, x_k \in V$, if $\exists\ \alpha_1, \ldots \alpha_k \in \mathbb{R}$ such that $x = \alpha_1 x_1 + \cdots + \alpha_k x_k$.
If $W \subset V$ then we denote
$Lin(W) := \{x \in V \,|\, x$ is the linear combination of the vectors in $W\}$

Def. 59. The vectors $x_1, \ldots, x_k \in V$ $(k \in \mathbb{N})$ are called lin. independent if $\alpha_1 x_1 + \cdots + \alpha_k x_k = o \Rightarrow \alpha_i = 0$ $(i = 1, \ldots, k)$. If we have infinitely many vectors, then we require the above property for all finite subset. ($\leftrightarrow$ lin. dependent)

Def. 60. The vector system $\mathcal{B} \subset V$ is called the basis of V if it is linearly independent and $Lin(\mathcal{B}) = V$.

If V possesses a bases with finitely many elements, then V is called finite dimensional vector space. In finite dimensional vector spaces the number of elements in each basis are equal. This is the dimension of the vector space.

Normed spaces

Def. 61. The pair $(V, \|.\|)$ is called normed space if V is a vector space and $\|.\| : V \to \mathbb{R}$ is a given function (so-called norm) with the properties:

1. $\|x\| = 0 \Leftrightarrow x = o$;
2. $\|\alpha x\| = |\alpha| \cdot \|x\|, \forall x \in V, \forall \alpha \in \mathbb{R}$;
3. $\|x + y\| \leq \|x\| + \|y\|, \forall x, y \in V$.

Ex.

- Vectors on the plane and in the space, $\|\vec{v}\| =$ is the usual length of the vectors.
- $\mathbb{R}^n$, $\mathbf{x} = [x_1, \dots, x_n]^T$:
  $\|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n|$,
  $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$,
  $\|\mathbf{x}\|_\infty = \max\{|x_1|, \dots, |x_n|\}$.
- $C[a, b]$, f
  $\|f\|_{C[a,b]} = \max_{x\in[a,b]}\{|f(x)|\}$
- $\mathbb{R}^{m\times n}$, $\mathbf{A} = [a_{ij}] \in \mathbb{R}^{m\times n}$
  $\|\mathbf{A}\| = \max_{i=1:m, j=1:n}\{|a_{ij}|\}$ (see later).

Convergence in normed spaces, $V = (V, \|.\|)$

Def. 62. The distance of the elements $x, y \in V$ is the value $\|x - y\|$.

Thm. 63.

- $\|x - y\| \geq 0$, $\forall x, y \in V$, $\|x - y\| = 0 \Leftrightarrow x = y$,
- $\|x - y\| = \|y - x\|$, $\forall x, y \in V$,
- $\|x - y\| \leq \|x - z\| + \|z - y\|$, $\forall x, y, z \in V$.

Def. 64. We say that the sequence $\{x_k\} \subset V$ tends to the element $x \in V$ if the real number sequence $\{\|x_k - x\|\}$ tends to zero. Notation: $x_k \to x$.

Def. 65. The norms $\|.\|_1$ és $\|.\|_2$ defined on the same vector space are called equivalent if $\exists c_1, c_2 > 0$ such that

$$c_1\|x\|_1 \leq \|x\|_2 \leq c_2\|x\|_1, \ \forall x \in V.$$

Convergence in normed spaces, $V = (V, \|.\|)$

Rmk. Equivalent norms define the same convergence. In finite dimensional vector spaces all norms are equivalent.

Def. 66. We say that the sequence $\{x_k\} \subset V$ is a Cauchy sequence if $\forall \varepsilon > 0$, $\exists M \in \mathbb{N}$, $\forall n, m \geq M$ $\|x_n - x_m\| < \varepsilon$.

Thm. 67. All convergent sequences in V are Cauchy sequences.

Rmk. The converse of the theorem is not true.

Def. 68. We say that the normed space $(V, \|.\|)$ is a Banach space if all Cauchy sequences in V are convergent.

Example. The examples listed for normed spaces are examples also for Banach spaces.

Banach fixed point theorem

Thm. 69. Let $(V, \|.\|)$ be a Banach space and $\emptyset \neq H \subset (V, \|.\|)$ a closed subset ($\{x_k\} \subset H$, $x_k \to x$ implies $x \in H$). Let $F : H \to H$ be a contraction ($\exists\ 0 \leq q < 1$, $\|F(x) - F(y)\| \leq q\|x - y\|$, $\forall\ x, y \in H$).

- Then F possesses one and only one fixed point in H, that is an element $x^\star \in H$ such that $F(x^\star) = x^\star$.
- With arbitrary initial element $x_0 \in H$, the sequence produced with the iteration $x_{k+1} = F(x_k)$ tends to $x^\star$.
- It is valid the estimation

$$\|x^\star - x_m\| \leq \frac{q^m}{1-q}\|x_1 - x_0\|. \tag{5}$$

Euclidean spaces

Euclidean spaces

Def. 70. The pair $(V, \langle ., . \rangle)$ is called euclidean space if V is a vector space and $\langle ., . \rangle : (V \times V) \to \mathbb{R}$ is a so-called scalar product with the properties:

1. $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in V$,
2. $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$, for all $x, y \in V,\ \alpha \in \mathbb{R}$,
3. $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$, for all $x, y, z \in V$,
4. $\langle x, x \rangle > 0$, for all $o \neq x \in V$.

Two important examples

- In the space of the column vectors $\mathbb{R}^n$: with the notations $\overline{\mathbf{x}} = [x_1, \ldots, x_n]^T$ and $\overline{\mathbf{y}} = [y_1, \ldots, y_n]^T$, the assignment $\langle \overline{\mathbf{x}}, \overline{\mathbf{y}} \rangle = x_1 y_1 + \ldots + x_n y_n$ defines a scalar product ($\overline{\mathbf{x}}^T \overline{\mathbf{y}}$).
- In the vector space $C[a, b]$, the assignment

$$\langle f, g \rangle = \int_a^b s(x) f(x) g(x)\, \mathrm{d}x$$

defines a scalar product for all positive weight function $s \in C[a, b]$.

Euclidean spaces

Thm. 71. In a euclidean space $(V, \langle ., . \rangle)$, the assignment $\|x\| = \sqrt{\langle x, x \rangle}$ defines a norm (norm induced be the scalar product).

Def. 72.

- $x, y \in V$ orthogonal if $\langle x, y \rangle = 0$,
- $x_1, x_2, \ldots \in V$ orthogonal vector system if the vectors are pairwise orthogonal,
- $x \in V$ is normed if $\|x\| = 1$ is fulfilled in the norm induced by the scalar product.
- $x_1, x_2, \ldots \in V$ is an orthonormal vector system if the vectors are pairwise orthogonal and each vector is normed.

Gram–Schmidt orthogonalization

Thm. 73. Let $x_1, \dots, x_k$ be a linearly independent vector system in a euclidean space. Then we can set an orthonormal vector system $q_1, \dots, q_k$ with the properties $\operatorname{lin}(q_1, q_2, \dots, q_l) = \operatorname{lin}(x_1, x_2, \dots, x_l)$ for all $l = 1, \dots, k$.

Rmk. The polynomials p, q are called orthogonal on the interval $[a, b]$ with respect to the positive weight function s if

$$\int_a^b s(x)p(x)q(x)\,\mathrm{d}x = 0.$$

Def. 74. Let us consider the polynomials $1, x, x^2$ on the interval $[-1, 1]$. Then the polynomials obtained with the Gram–Schmidt orthogonalization using the weight function $s(x) \equiv 1$ in the scalar product are called Legendre polynomials, while with the weight function $s(x) = 1/\sqrt{1 - x^2}$ we obtain the so-called Chebyshev polynomials.

Orthogonal polynomials

Degree	Legendre	Chebyshev
0	1	1
1	x	x
2	$(3x^2-1)/2$	$2x^2-1$
3	$(5x^3-3x)/2$	$4x^3-3x$
4	$(35x^4-30x^2+3)/8$	$8x^4-8x^2+1$

$T_0 = 1$, $T_1 = x$
Chebyshev: $T_{k+1} = 2xT_k - T_{k-1}$.
Legendre: $(k+1)T_{k+1} = (2k+1)xT_k - kT_{k-1}$.

Orthogonal polynomials

Thm. 75. Let us suppose that the polynomials $p_0, p_1, \ldots$ (subscripts denote the degrees) are pairwise orthogonal on the interval $[a, b]$ with respect to the positive weight function s. Then all roots of the polynomial are real, single and located in the interval $[a, b]$.

Proof. Let us consider the polynomial p_l and denote the distinct real roots from $[a, b]$ with odd multiplicity by $z_1, \ldots, z_k$. If $k = l$, then the statement is true, if $k < l$, then let us consider the polynomial $p(x) = (x - z_1)\ldots(x - z_k)$ ($p \equiv 1$ if $k = 0$), which has degree k. The polynomial $p_l \cdot p$ has degree $(l + k)$ and it does not change sign in the interval $[a, b]$. Thus the condition

$$\int_a^b p_l(x)p(x)s(x)\,\mathrm{d}x = 0$$

cannot hold. This completes the proof. ■

Special properties of matrices

Special matrices

- Band matrix: $\exists p, q \in \mathbb{N}$, $a_{i,j} = 0$ if $j < i - p$ or $i < j - q$. $1 + p + q$ is the so-called bandwidth.
- Diagonal matrix: offdiagonal elements are zero $(p = 0, q = 0)$, $\mathbf{I}$ identity matrix.
- Upper triangular matrix: elements "below" the diagonal are zero $(p = 0)$.
- Lower triangular matrix: elements "above" the diagonal are zero $(q = 0)$.
- Upper Hessenberg matrix: elements "below" the subdiagonal are zero $(p = 1)$.
- Lower Hessenberg matrix: elements "above" the superdiagonal are zero $(q = 1)$.

Special matrices

- Tridiagonal matrix: all elements outside the main, sub- and superdiagonals are zero. $(p = q = 1)$.
- Symmetric matrix: $\mathbf{A}^T = \mathbf{A}$
- Skew-symmetric matrix: $\mathbf{A}^T = -\mathbf{A}$
- The vectors $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}} \in \mathbb{R}^n$ are called orthogonal if $\overline{\mathbf{x}}^T\overline{\mathbf{y}} = 0$. Moreover, we trivially have $\overline{\mathbf{y}}^T\overline{\mathbf{x}} = 0$. If $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are orthogonal, then $\|\overline{\mathbf{x}} + \overline{\mathbf{y}}\|_2^2 = \|\overline{\mathbf{x}}\|_2^2 + \|\overline{\mathbf{y}}\|_2^2$ (Pythagorean theorem).

  Orthogonal matrix: $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$

  $(\|\mathbf{A}\mathbf{x}\|_2^2 = \mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \|\mathbf{x}\|_2^2,\ \|\mathbf{A}\|_2 = 1,\ \|\mathbf{A}\mathbf{B}\|_2 = \|\mathbf{B}\|_2)$

Special matrices

- $\mathbf{P}$ is a permutation matrix if, with the notation $\overline{\mathbf{e}}_k = [0,\dots,0,\overbrace{1}^{k\text{-adik}},0,\dots,0]^T$ $(k=1,\dots,n)$, $\mathbf{P} = [\overline{\mathbf{e}}_{i_1},\dots,\overline{\mathbf{e}}_{i_n}]$, where $i_1,\dots,i_n$ is a permutation of the numbers $1,2,\dots,n$. The product $\mathbf{AP}$ rearranges the columns of $\mathbf{A}$ in the order $i_1,\dots,i_n$, while the product $\mathbf{P}^T\mathbf{A}$ does the same with the rows of $\mathbf{A}$. It is valid the relation $\mathbf{PP}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}$.
- Let $\mathbf{A}$ be a symmetric matrix, and we investigate the possible values of the expression $f(\mathbf{x}) := \mathbf{x}^T\mathbf{Ax}$ if $\mathbf{x} \neq \mathbf{0}$:
  – always positive (negative): $\mathbf{A}$ positive (negative) definite,
  – always nonnegative (nonpositive): $\mathbf{A}$ positive (negative) semidefinite,
  – can be both positive and negative: $\mathbf{A}$ indefinite.
- Diagonally dominant matrix: $|a_{ii}| \geq \sum_{j=1,j\neq i}^{n} |a_{ij}|$, $\forall i = 1,\dots,n$. Strictly diagonally dominant matrix if "$>$" is valid.

Eigenvalues and eigenvectors of matrices

Eigenvalues and eigenvectors

Def. 76. Suppose that there is a vector $\overline{\mathbf{v}} \neq \mathbf{0}$ and a number λ to the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ such that $\mathbf{A}\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}$. Then the number λ is called the eigenvalue of the matrix $\mathbf{A}$, while the vector $\overline{\mathbf{v}}$ is called an eigenvector corresponding to the eigenvalue λ.

Thm. 77. Eigenvalues are the solutions of the so-called characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. (Real values or complex conjugate pairs.) The number of eigenvalues counted with multiplicity is n (algebraic multiplicity). Proof. Trivial. ■

Thm. 78. The linear combinations of eigenvectors are also eigenvectors ($\neq \mathbf{0}$). Proof. Trivial. ■

Thm. 79. $\exists \mathbf{A}^{-1} \Leftrightarrow \lambda_i \neq 0, \ \forall\ i = 1, \ldots, n$. Proof. Trivial. ■

Eigenvalues and eigenvectors

Thm. 80.

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i, \quad \mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i.$$

Proof. It can be proven with investigation of the coefficients of the characteristic polynomial. ■

Rmk. The eigenvalues can be complex numbers. In this case the eigenvectors also have complex elements.

Def. 81. For complex matrices $\mathbf{A}$, $\mathbf{A}^H$ denotes the transpose conjugate of the matrix. If $\mathbf{A}^H = \mathbf{A}$ is valid, then the matrix is called hermitian matrix. A matrix in unitary if $\mathbf{A}^H\mathbf{A} = \mathbf{A}\mathbf{A}^H = \mathbf{I}$.

Eigenvalues and eigenvectors

Thm. 82. All eigenvalues of symmetric (real) matrices are real, the eigenvectors can be chosen to real vectors.

Proof. Let $\overline{\mathbf{v}}$ be an eigenvector with the eigenvalue λ. Then $\overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}} = \overline{\mathbf{v}}^H\lambda\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}^H\overline{\mathbf{v}}$. Trivially

$$(\overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}})^H = \overline{\mathbf{v}}^H\mathbf{A}\overline{\mathbf{v}}, \quad (\overline{\mathbf{v}}^H\overline{\mathbf{v}})^H = \overline{\mathbf{v}}^H\overline{\mathbf{v}},$$

that is these are 1×1 matrices. The conjugate transpose of these matrices are themselves. Thus λ must be real. The eigenvectors are the solutions of the system of equations $(\mathbf{A} - \lambda\mathbf{I})\overline{\mathbf{x}} = \mathbf{0}$, which can be chosen to be real. ■

Eigenvalues and eigenvectors

Thm. 83. All eigenvalues of symmetric, positive (semi)definite matrices are (nonnegative) positive.

Proof. Let $\overline{\mathbf{v}}$ be an eigenvector with the eigenvalue λ (real). Then the statement follows from the equalities $\overline{\mathbf{v}}^T \mathbf{A} \overline{\mathbf{v}} = \overline{\mathbf{v}}^T \lambda \overline{\mathbf{v}} = \lambda \overline{\mathbf{v}}^T \overline{\mathbf{v}} > 0$ and $\overline{\mathbf{v}}^T \overline{\mathbf{v}} > 0$ (the proof is similar for semidefinite matrices). ■

Def. 84. The greatest absolute value of the eigenvalues of the matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is called the spectral radius of $\mathbf{A}$. Notation: $\varrho(\mathbf{A})$. That is

$$\varrho(\mathbf{A}) = \max\{|\lambda_i| \,|\, \lambda_i \text{ is an eigenvalue of } \mathbf{A}\}.$$

Gershgorin theorem

Thm. 85. Let us consider the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. Let K_i be the closed circle on the complex plane defined as follows. Its center is a_{ii} and its radius is $\sum_{j=1,j\neq i}^{n} |a_{ij}|$ $(i = 1, \ldots, n)$. Then all the eigenvalues of the matrix are in the set $\cup_i K_i$.

Proof. Let λ be an eigenvalue of the matrix. If λ equals one of the diagonal elements, then the statement is true for this eigenvalue. Otherwise, let us write $\mathbf{A}$ in the form $\mathbf{A} = \mathbf{D} + \mathbf{T}$, where $\mathbf{D}$ is the diagonal matrix of $\mathbf{A}$. $\mathbf{A} - \lambda\mathbf{I}$ is singular, thus there exists a vector $\overline{\mathbf{x}} \neq \mathbf{0}$, with which $(\mathbf{A} - \lambda\mathbf{I})\overline{\mathbf{x}} = \mathbf{0}$, that is $(\mathbf{D} - \lambda\mathbf{I})\overline{\mathbf{x}} = -\mathbf{T}\overline{\mathbf{x}}$.

Gershgorin theorem

Hence

$$\|\overline{\mathbf{x}}\|_\infty \leq \|(\mathbf{D} - \lambda\mathbf{I})^{-1}\mathbf{T}\|_\infty \|\overline{\mathbf{x}}\|_\infty,$$

that is

$$1 \leq \frac{\sum_{j=1, j\neq k}^{n} |a_{kj}|}{|a_{kk} - \lambda|}$$

for some index $k = 1, \ldots, n$. Thus $\lambda \in K_k$. ■

Rmk. When the union of s Gershgorin circles is disjoint from the other circles, then the union contains exactly s eigenvalues (2. Gershgorin theorem).

Diagonalizability of matrices

Diagonalizability

Def. 86. Two quadratic matrices $(\mathbf{A}, \mathbf{B})$ are similar if $\exists$ $\mathbf{S}$ nonsingular matrix, for which $\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$.

Thm. 87. The eigenvalues of similar matrices are equal.

Proof.

$$\det(\mathbf{B} - \lambda\mathbf{I}) = \det(\mathbf{S}^{-1}\mathbf{A}\mathbf{S} - \lambda\mathbf{I})$$

$$= \det(\mathbf{S}^{-1})\det(\mathbf{A} - \lambda\mathbf{I})\det(\mathbf{S}) = \det(\mathbf{A} - \lambda\mathbf{I}). \blacksquare$$

Rmk. If $\overline{\mathbf{v}}$ is an eigenvector of $\mathbf{B}$ then $\mathbf{S}\overline{\mathbf{v}}$ is an eigenvector of $\mathbf{A}$.

Def. 88. A matrix $\mathbf{A}$ is called diagonalizable if it is similar to a diagonal matrix.

Diagonalizability

Ex.: Not diagonalizable:

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

. 1 is double eigenvalue, thus it must be similar to the identity matrix but then $\mathbf{A} = \mathbf{S}^{-1}\mathbf{I}\mathbf{S} = \mathbf{I}$, which is not true.

Thm. 89. Eigenvectors that belong to different eigenvalues are linearly independent.

Proof. Suppose $\mathbf{A}\overline{\mathbf{v}} = \lambda\overline{\mathbf{v}}$ és $\mathbf{A}\overline{\mathbf{w}}_i = \mu\overline{\mathbf{w}}_i$ $(i = 1, \ldots, l)$, $\lambda \neq \mu$ and $\overline{\mathbf{v}} = \sum_{i=1}^{l} \alpha_i \overline{\mathbf{w}}_i$ for some constant $\alpha_i \neq 0$. Then

$$\lambda\overline{\mathbf{v}} = \mathbf{A}\overline{\mathbf{v}} = \mathbf{A}\sum_{i=1}^{l} \alpha_i \overline{\mathbf{w}}_i = \mu \sum_{i=1}^{l} \alpha_i \overline{\mathbf{w}}_i = \mu\overline{\mathbf{v}},$$

which implies the equality $\lambda = \mu$. ■

Cor.: When all the eigenvectors of a matrix are different, then the matrix has a linearly independent eigenvector system.

Diagonalizability

Thm. 90. An $n \times n$ matrix is diagonalizable if and only if it has a linearly independent eigenvector system with n vectors.

Proof. $\Leftarrow$ $\mathbf{A}\overline{\mathbf{v}}_j = \lambda_j \overline{\mathbf{v}}_j \ (j = 1, \dots, n)$

$$\mathbf{A} \underbrace{\begin{bmatrix} \overline{\mathbf{v}}_1 & \dots & \overline{\mathbf{v}}_n \end{bmatrix}}_{:=\mathbf{S}} = \begin{bmatrix} \overline{\mathbf{v}}_1 & \dots & \overline{\mathbf{v}}_n \end{bmatrix} \underbrace{\begin{bmatrix} \lambda_1 & 0 & 0 & \dots \\ 0 & \lambda_2 & 0 & \dots \\ & & \ddots & \end{bmatrix}}_{:=\mathbf{\Lambda}}$$

Thus $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$, that is the matrix is diagonalizable.

$\Rightarrow$ $\exists \mathbf{S}$ regular matrix, with which $\mathbf{S}^{-1}\mathbf{A}\mathbf{S} = \mathbf{\Lambda}$ for some diagonal matrix $\mathbf{\Lambda}$. Then the eigenvalues of $\mathbf{A}$ equal the elements of $\mathbf{\Lambda}$. Since the system $\overline{\mathbf{e}}_j$ is an eigenvector system of $\mathbf{\Lambda}$, $\mathbf{S}\overline{\mathbf{e}}_j$ is an eigenvector system of $\mathbf{A}$. These are linearly independent vectors because of the regularity of $\mathbf{S}$. ■

Diagonalizability

Def. 91. A matrix $\mathbf{A}$ is called normal if $\mathbf{A}^H\mathbf{A} = \mathbf{A}\mathbf{A}^H$.

Thm. 92. Normal matrices are diagonalizable.

Proof. Let λ_1 and $\overline{\mathbf{v}}_1$ be an eigenvalue and the corresponding eigenvector of the matrix (these always exist - they can be complex). Let $\overline{\mathbf{v}}_1$ satisfy the condition $\overline{\mathbf{v}}_1^H\overline{\mathbf{v}}_1 = 1$ (the vector is normed). Let us extend this vector to a unitary system $(\overline{\mathbf{v}}_2, \dots, \overline{\mathbf{v}}_n)$. Then

$$\mathbf{A}\underbrace{\begin{bmatrix} \overline{\mathbf{v}}_1 & \dots & \overline{\mathbf{v}}_n \end{bmatrix}}_{:=\mathbf{S}_1 \text{ unitér}} = \begin{bmatrix} \overline{\mathbf{v}}_1 & \dots & \overline{\mathbf{v}}_n \end{bmatrix} \begin{bmatrix} \lambda_1 & * & * & \dots \\ 0 & * & * & \dots \\ & & \ddots & \\ 0 & * & * & \dots \end{bmatrix}.$$

Thus

$$\mathbf{S}_1^H\mathbf{A}\mathbf{S}_1 = \begin{bmatrix} \lambda_1 & * \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}.$$

Diagonalizability

Let us repeat the previous procedure for the matrix $\mathbf{A}_2$! There exists a unitary matrix $\tilde{\mathbf{S}}_2$ such that

$$\tilde{\mathbf{S}}_2^H \mathbf{A}_2 \tilde{\mathbf{S}}_2 = \begin{bmatrix} \lambda_2 & * & * & \dots \\ 0 & * & * & \dots \\ & & \ddots & \\ 0 & * & * & \dots \end{bmatrix}.$$

Let

$$\mathbf{S}_2 = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{S}}_2 \end{bmatrix}.$$

Then

$$\mathbf{S}_2^H \mathbf{S}_1^H \mathbf{A} \mathbf{S}_1 \mathbf{S}_2 = \begin{bmatrix} \lambda_1 & * & * & \dots \\ 0 & \lambda_2 & * & \dots \\ & & \ddots & \\ 0 & 0 & * & \dots \end{bmatrix}.$$

Diagonalizability

Similarly, we can obtain the unitary matrices $\mathbf{S}_3, \dots, \mathbf{S}_{n-1}$. With these matrices we have

$$\mathbf{S}_{n-1}^H \dots \mathbf{S}_2^H \mathbf{S}_1^H \mathbf{A} \mathbf{S}_1 \mathbf{S}_2 \dots \mathbf{S}_{n-1} = \underbrace{\begin{bmatrix} \lambda_1 & * & * & \dots & * \\ 0 & \lambda_2 & * & \dots & * \\ & & \ddots & & \\ 0 & 0 & 0 & \dots & \lambda_n \end{bmatrix}}_{:=\mathbf{T} \text{ (upper triangular)}}.$$

Let $\mathbf{S} = \mathbf{S}_1 \dots \mathbf{S}_{n-1}$. This is a unitary matrix.

$$\mathbf{T}^H \mathbf{T} = \mathbf{S}^H \mathbf{A}^H \mathbf{S} \mathbf{S}^H \mathbf{A} \mathbf{S} = \mathbf{S}^H \mathbf{A}^H \mathbf{A} \mathbf{S},$$

$$\mathbf{T} \mathbf{T}^H = \mathbf{S}^H \mathbf{A} \mathbf{S} \mathbf{S}^H \mathbf{A}^H \mathbf{S} = \mathbf{S}^H \mathbf{A} \mathbf{A}^H \mathbf{S},$$

thus $\mathbf{T}$ is normal. $\mathbf{T}$ can be upper triangular only if it is diagonal. ■

Diagonalizability

Rmk. Every matrix can be written in the form $\mathbf{A} = \mathbf{S}\mathbf{T}\mathbf{S}^H$, where $\mathbf{S}$ is unitary and $\mathbf{T}$ is an upper triangular matrix. This is the so called Schur decomposition.

Rmk. Normal matrices can be diagonalized with a unitary matrix. Matrices that are diagonalizable with a unitary matrix are normal.

Rmk. Real normal matrices are e.g. symmetric, skew-symmetric and orthogonal matrices.

Thm. 93. A real matrix is diagonalizable with an orthogonal matrix if and only if it is symmetric.

Proof. $\Rightarrow$ Let $\mathbf{S}$ be orthogonal and $\mathbf{A} = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^T$. Then $\mathbf{A}^T = \mathbf{S}\boldsymbol{\Lambda}\mathbf{S}^T = \mathbf{A}$, which shows the symmetry.

Diagonalizability

$\Leftarrow$ Let $\overline{\mathbf{v}}_\lambda$ and $\overline{\mathbf{v}}_\mu$ be two eigenvalues corresponding to two different eigenvectors (λ and μ).

$$\overline{\mathbf{v}}_\lambda^T \mathbf{A} \overline{\mathbf{v}}_\mu = \overline{\mathbf{v}}_\lambda^T \mu \overline{\mathbf{v}}_\mu = \mu \overline{\mathbf{v}}_\lambda^T \overline{\mathbf{v}}_\mu,$$

$$\overline{\mathbf{v}}_\mu^T \mathbf{A} \overline{\mathbf{v}}_\lambda = \overline{\mathbf{v}}_\mu^T \lambda \overline{\mathbf{v}}_\lambda = \lambda \overline{\mathbf{v}}_\mu^T \overline{\mathbf{v}}_\lambda = \lambda \overline{\mathbf{v}}_\lambda^T \overline{\mathbf{v}}_\mu$$

These two values must be equal. This is possible only if $\overline{\mathbf{v}}_\lambda^T \overline{\mathbf{v}}_\mu = 0$. Thus the eigenvectors corresponding to different eigenvalues are orthogonal. Thus we can choose an orthonormal system of eigenvectors. The matrix can be diagonalized with the matrix that have the orthonormal eigenvectors in the columns. ■