Outline

Outline

Quadratic linear system of equations:

- a matrix $A \in \mathbb{R}^{n \times n}$ and a vector $b \in \mathbb{R}^n$ are a given,
- compute a solution $x \in \mathbb{R}^n$ with

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}.$$

Linear Algebra:

A system of lin. equations has a unique solution $\Leftrightarrow \quad \mathrm{rg}(A) = n$

$\Leftrightarrow \quad \det(A) \neq 0$

$\Leftrightarrow \quad A$ is regular.

Direct methods:

- compute the solution in a finite number of calculations,
- the number of calculations depends on the number of equations.

Iterative methods:

- theoretically an infinite number of iterations is necessary,
- generating a sequence $(x_0, x_1, x_2, \ldots)$ for a starting approximation x_0,
- the goal is convergence to the exact solution x^*,
- the algorithm must be stopped, e.g. if the iterate x_k is close enough to x^* or if the method fails.

Outline

Idea of Gaussian elimination:

- transform the system in a triangular system of equations,
- the set of solutions remains unchanged if
  (i) we multiply one equation by $\alpha \neq 0$,
  (ii) we add a multiple of one equation to another equation,
- the algorithm consists of two parts, a forward elimination and a backward substitution.

Forward elimination for a 3×3-example:

$$\begin{pmatrix} 2 & 4 & -2 \\ 1 & -1 & 5 \\ 4 & 1 & -2 \end{pmatrix} x = \begin{pmatrix} 6 \\ 0 \\ 2 \end{pmatrix},$$
$$\rightarrow \begin{pmatrix} 2 & 4 & -2 \\ 0 & -3 & 6 \\ 0 & -7 & 2 \end{pmatrix} x = \begin{pmatrix} 6 \\ -3 \\ -10 \end{pmatrix},$$
$$\rightarrow \begin{pmatrix} 2 & 4 & -2 \\ 0 & -3 & 6 \\ 0 & 0 & -12 \end{pmatrix} x = \begin{pmatrix} 6 \\ -3 \\ -3 \end{pmatrix}.$$

$\rightarrow$ triangular matrix.

Backward substitution:

3. equation:

$$-12x_3 = -3 \quad \Rightarrow \quad x_3 = 0.25,$$

2. equation:

$$-3x_2 + 6x_3 = -3 \quad \Rightarrow \quad x_2 = 1 + 2x_3 = 1.5,$$

1. equation:

$$2x_1 + 4x_2 - 2x_3 = 6 \quad \Rightarrow \quad x_1 = 3 - 2x_2 + x_3 = 0.25.$$

Solution:

$$x = \begin{pmatrix} 0.25 \\ 1.5 \\ 0.25 \end{pmatrix}.$$

Column pivoting:

- Select as a new output row the one with the largest leading absolute value,
- swap row i with row ℓ according to

$$|a_{\ell i}| \geq |a_{ji}| \quad \forall j = i, \ldots, n.$$

Possible results of the Gaussian elimination with pivot search:

- unique solution,
- appearance of an equation that cannot be solved,
  then the entire system of linear equations has no solutions,
- parameter dependent solution.

LU-decomposition:

- modification to solve several systems for the same A but different b,
- $A = LU$ with

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{21} & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ l_{n1} & & \dots & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{21} & \cdots & r_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix},$$

- computation of a solution

$$Ax = L \underbrace{Ux}_{u} = b \quad \Rightarrow \quad Lu = b,\ Ux = u,$$

- every step of Gaussian elimination can be represented as a multiplication by a matrix L_i, with $i = 1, 2, \dots, n-1$ of special structure.

Algorithm for solving a system of linear equations:

1. decompose the matrix A by Gaussian elimination in LU,
2. forward elimination to compute y as

$$y_i = \frac{1}{L_{ii}} \left(b_i - \sum_{j=1}^{i-1} L_{ij} y_j \right), \qquad i = 1, \ldots, n,$$

3. backward elimination to compute x as

$$x_i = \frac{1}{U_{ii}} \left(y_i - \sum_{j=i+1}^{n} U_{ji} x_j \right), \qquad i = n, n-1, \ldots, 2, 1.$$

Computational effort of Gaussian elimination:

- counting the floating point operations (flops),
- $n-j$ divisions to compute the elimination factors l_{ij}, $i=j+1,\ldots,n$,
- $2(n-j)^2$ multiplications and subtractions to compute a_{kl}, $k,l=j+1,\ldots,n$,
- $\frac{2}{3}n^3+\frac{1}{2}n^2-\frac{7}{6}n$ flops for forward elimination,
- n^2 flops for backward substitution,
- LU-decomposition together with forward and backward elimination takes

$$\frac{2}{3}n^3+\mathcal{O}(n^2)=\mathcal{O}(n^3)\text{flops.}$$

Definition 6.1

A <u>norm</u> (or vector norm) in $\mathbb{R}^n$ *is a map* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$*, that satisfies the following axions:*

(i) $\|x\| \geq 0 \; \forall x$, $\|x\| = 0 \Leftrightarrow x = (0, \ldots, 0)^T$ *(positivity, definiteness),*

(ii) $\|\lambda x\| = |\lambda| \|x\| \; \forall \lambda \in \mathbb{K}, \; x \in V$ *(homogeneity),*

(iii) $\|x + y\| \leq \|x\| + \|y\| \; \forall x, y \in V$ *(triangle inequality).*

Vector norms

Often used norms:

- Manhattan norm

$$\|x\|_1 = |x_1| + \cdots + |x_n|,$$

- Euclidean norm

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

- Chebyshev norm

$$\|x\|_\infty = \max_{i=1\ldots n} |x_i|,$$

- p-norm for $p \geq 1$

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}.$$

Matrix norms

Definition 6.2

A matrix norm in $\mathbb{R}^{n\times n}$ *is a map* $\|\cdot\|:\ \mathbb{R}^{n\times n}\to\mathbb{R}_+$*, which satisfies the following axioms:*

1. *Positivity, definiteness*

$$\|A\|\geq 0,\quad \|A\|=0\Leftrightarrow A=0\qquad \forall A\in\mathbb{R}^{n\times n},$$

2. *Homogeneity:*

$$\|cA\|=|c|\|A\|,\qquad \forall A\in\mathbb{R}^{n\times n},\ \forall c\in\mathbb{R},$$

3. *Triangle inequality:*

$$\|A+B\|\leq\|A\|+\|B\|,\qquad \forall A,B\in\mathbb{R}^{n\times n}.$$

Definition 6.3

A matrix norm $\|\cdot\|$ is called submultiplicative, if

$$\|AB\| \leq \|A\| \cdot \|B\| \qquad \forall A, B \in \mathbb{R}^{n\times n}.$$

Definition 6.4

A matrix norm $\|\cdot\|_M$ is called to be consistent (or compatible) with a given vector norm $\|\cdot\|_V$, if

$$\|Ax\|_V \leq \|A\|_M \cdot \|x\|_V \qquad \forall A \in \mathbb{R}^{n\times n},\ \forall x \in \mathbb{R}^n.$$

Example: the vector norm $\|\cdot\|_2$ induces the spectral norm

$$\|A\|_2 = \left(\rho(A^T A)\right)^{0.5} = \left(\lambda_{\max}(A^T A)\right)^{0.5}$$

with $\rho(B)$ being the spectral radius of B

$$\rho(B) = \max_j \|\lambda_j|,$$

where the λ_i are the eigenvalues of B.

Theorem 6.5

For any matrix norm $\|\cdot\| : \mathbb{R}^{n\times n} \to \mathbb{R}_+$ *holds the inequality*

$$\rho(A) \leq \|A\|.$$

Remarks:

1. For a symmetric A holds

$$\|A\|_2 = \sqrt{\rho(A^T A)} = \sqrt{\rho(A^2)} = \sqrt{(\rho(A))^2} = \rho(A).$$

2. For an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ holds

$$\|Q\|_2 = \sqrt{\rho(Q^T Q)} = \sqrt{\rho(I)} = 1.$$

Behaviour of errors during the Gaussian elimination:

- approximate solution and residual vector

$$\tilde{x} := x + \delta_x, \qquad r := b - A\tilde{x},$$

- if A is exact and b is a disturbed right-hand-side, then

$$\frac{\|\delta_x\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|r\|}{\|b\|}.$$

Definition 6.6

Let $\|\cdot\|_M$ be a matrix norm and A a regular matrix. Then

$$\mathrm{acond}_M(A) := \|A^{-1}\|_M$$

is called absolute condition number and

$$\mathrm{cond}_M(A) := \|A\|_M \cdot \|A^{-1}\|_M$$

is called relative condition number of A.

Theorem 6.7

It holds

1. $\mathrm{cond}(A) \geq 1$,
2. $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$,
3. $\mathrm{cond}(A) = \frac{\max_{\|x\|=1} \|Ax\|}{\min_{\|x\|=1} \|Ax\|}$.

Remarks:

1. A regular matrix with $\mathrm{cond}(A) \approx 1$ is called well-conditioned.
   If $\mathrm{cond}(A) \gg 1$, then A is called ill-conditioned.
2. Each orthogonal matrix $V \in \mathbb{R}^{n \times n}$ is of Euclidean condition 1, as

$$\|Vx\|_2^2 = (Vx, Vx)_2 = x^T V^T Vx = x^T x = \|x\|_2^2,$$
$$\mathrm{cond}_2(V) = \|V\|_2 \|V^{-1}\|_2 = 1.$$

Perturbation of the system matrix:

- let b exact and A a disturbed matrix,
- we assume $\|A^{-1}\| \cdot \|\delta_A\| \leq 1$,
- the relative error is

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\|A^{-1}\delta_A\|}{1 - \|A^{-1}\delta_A\|} = \frac{\operatorname{cond}(A)\frac{\|\delta_A\|}{\|A\|}}{1 - \operatorname{cond}(A)\frac{\|\delta_A\|}{\|A\|}}.$$

Perturbation of all data:

- let A and b are disturbed,
- the relative error is

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{\operatorname{cond}(A)\left(\frac{\|\delta_b\|}{\|b\|}\ \frac{\|\delta_A\|}{\|A\|}\right)}{1 - \operatorname{cond}(A)\frac{\|\delta_A\|}{\|A\|}}.$$

Outline

Theorem 6.8 (Cholesky decomposition)

Let $A \in \mathbb{R}^{n\times n}$ be a symmetric and positive definite matrix. There exists a unique decomposition of the form

$$A = CC^T$$

with C being an invertible lower triangle and positive diagonal entries.

Computation of C:

$$C_{ii} = \sqrt{A_{ii} - \sum_{j=1}^{i-1} C_{ij}^2}, \qquad i = 1, \dots, n,$$

$$C_{ji} = \frac{1}{C_{ii}}\left(A_{ji} - \sum_{k=1}^{i-1} C_{jk}C_{ik}\right), \qquad j = i+1, \dots, n.$$

Outline

Iterative methods to solve systems of linear equations

Gauss elimination:

- direct method to solve systems of linear equations,
- computation of the solution after a finite number of calculations.

Iterative methods:

- starting with an inital guess,
- sequence of iterations, try to improve the approximation in each iteration,
- the matrix A remains unchanged,
- theoretically an infinite number of iterations is necessary,
- hopefully the iterations converges to the exact solution fast.

Given:

- system of linear equations $Ax = b$,
- approximate inverse
$$B^{-1} \approx A^{-1},$$
- starting vector
$$x^{(0)} \in \mathbb{R}^n.$$

Iteration:

- iteration for $i = 0, 1, 2, \ldots$
$$x^{(i+1)} = x^{(i)} + B^{-1}(b - Ax^{(i)}).$$

Iterative methods to solve systems of linear equations

Jacobi's method:

- selection of $B = D$ results in the iteration

$$x^{(i+1)} = x^{(i)} + D^{-1}(b - Ax^{(i)}), \quad i = 0, 1, 2, \dots,$$

- element-wise form

$$x_k^{(i+1)} = \frac{1}{a_{kk}} \left(b_k - \sum_{\ell=1,\ \ell \neq k}^{n} a_{k\ell} x_\ell^{(i)} \right).$$

Gauss-Seidel's method:

- selection of $B = L + D$ results in the iteration

$$Dx^{(i+1)} = b - Lx^{(i+1)} - Rx^{(i)}, \quad i = 0, 1, 2, \dots,$$

- element-wise form

$$x_k^{(i+1)} = \frac{1}{a_{kk}} \left(b_k - \sum_{\ell=1}^{k-1} a_{k\ell} x_\ell^{(i+1)} - \sum_{\ell=k+1}^{n} a_{k\ell} x_\ell^{(i)} \right).$$

Convergence analysis:

- What means convergence?
- Under what conditions to the matrix does such an iteration work?

Definition 6.9

An iteration

$$x^{(i+1)} = x^{(i)} + B^{-1}(b - Ax^{(i)}), \quad i = 0, 1, 2, \ldots \tag{5}$$

is called convergent, if the series of iterations converges to the solution $x^* = A^{-1}b$ for an arbitrary (!) starting vector $x^{(0)}$.

Convergence of iterative methods

Definition 6.10

A matrix $A \in \mathbb{R}^{n \times n}$ is called <u>strictly diagonally dominant</u>, if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad \text{für alle } i = 1, \ldots, n. \tag{6}$$

Theorem 6.11 (sufficient but not necessary condition)

Jacobi's method and Gauss-Seidel's method converge in each case if the matrix A is strictly diagonally dominant.

Remark 5

It is possible to reduce the condition to a weakly diagonal dominant matrix. This means it only holds „$\geq$" in (6) instead of „$>$", as long as A is irreducible.
See, e. g. Bärwolff, Hämmerlin, Hoffmann, etc..

General analysis of convergence using error propagation matrix:

- exact solution x^*,
- error for the new iteration

$$x^{(i+1)} - x^* = \underbrace{(I - B^{-1}A)}_{\text{error propagation matrix}} (x^{(i)} - x^*).$$

Theorem 6.12 (Standard convergence condition)

The iteration from (5) converges if and only if the spectral radius of the error propagation matrix is smaller than one, i.e. all eigenvalues are smaller than 1 in magnitude.

Stopping criterion:

- check x-congergence and stop the iteration if

$$\left\|x^{(i)} - x^{(i-1)}\right\| < \varepsilon_{x,a} \quad \text{resp.} \quad \frac{\left\|x^{(i)} - x^{(i-1)}\right\|}{\left\|x^{(i-1)}\right\|} < \varepsilon_{x,r}$$

  for suitable control parameters $\varepsilon_{x,a}$, $\varepsilon_{x,r}$, e. g. $\varepsilon_{x,*} = 10^{-6}$,
- check the change in the residual and stop the iteration if

$$\left\|Ax^{(i)} - b\right\| < \varepsilon_{f,a} \quad \text{resp.} \quad \frac{\left\|Ax^{(i)} - b\right\|}{\left\|Ax^{(i-1)} - b\right\|} < \varepsilon_{f,r},$$

- stop the iteration if the number of steps is larger than a limit $i_{\max}$.

Relaxation

Relaxation in matrix notation:

$$\begin{aligned} x^{(i+1)} &= x^{(i)} + \omega B^{-1}(b - Ax^{(i)}) \\ &= (I - \omega B^{-1}A)x^{(i)} + \omega B^{-1}b. \end{aligned}$$

Types of relaxation:

- $\omega < 1$ under relaxation (usual choice $\omega = 2/3$),
- $\omega = 1$ original method,
- $\omega > 1$ over relaxation.

Definition 6.13

For an iteration with

$$\|x^{(i+1)} - x^*\| \leq C\|x^{(i)} - x^*\|^q$$

the integer q is called the order of convergence.

Linear convergence as

$$\underbrace{x^{(i+1)} - x^*}_{\delta^{(i)}} = (I - B^{-1}A)\underbrace{(x^{(i)} - x^*)}_{\delta^{(i)}}. \qquad (7)$$

A-priori error estimation:

$$\|x^{(i)} - x^*\| \leq \frac{C^k}{1-C}\|x^{(1)} - x^{(0)}\|,$$

A-posteriori estimation:

$$\|x^{(i)} - x^*\| \leq \frac{C}{1-C}\|x^{(i)} - x^{(i-1)}\|.$$

Outline

Conjugate gradient method

Problem:

- let $A \in \mathbb{R}^{n\times n}$ be symmeytric, positive definite, so

$$A^T = A, \qquad x^T Ax >= \quad \text{for all } x \neq 0,$$

- solve the system of linear equations $Ax = b$,
- common situation for (partial) differential equations.

Conjugate gradient method (1952):

- maximum of n iterations,
- all directions of search are orthogonal to each other and the step lengths are optimal in each case.

<u>Preliminary consideration 1</u>:

- as A is symm. positive definite $x^* = A^{-1}b$ is the uniquely determined minimum of

$$f : \mathbb{R}^n \to \mathbb{R}, \quad f(x) = \frac{1}{2}x^T Ax - x^T b,$$

- assume a direction of search $d \in \mathbb{R}^n$ and consider for the step size t

$$g(t) = f(x + t \cdot d),$$

- optimal step size t^* is

$$t^* = \frac{d^T(b - Ax)}{d^T Ad} \in \mathbb{R}.$$

Preliminary consideration 2:

- using the norm and the scalar product to A (this is possible as A is symm. positive definite)

$$\|\cdot\|_A \text{ with } \|x\|_A^2 = x^T A x \quad \text{and} \quad (x,y)_A = x^T A y,$$

- all direction of search should be orthogonal to each other with respect to $(\cdot,\cdot)_A$,
- if d is the last direction of search we use

$$\tilde{d} = \tilde{r} + s^* d$$

with

$$\tilde{r} = b - Ax^{(1)}, \quad \text{and} \quad s^* = -\frac{\tilde{r}^T A d}{d^T A d}.$$

Conjugate gradient method

Initial choices:

- starting vector $x^{(0)} \in \mathbb{R}^n$, with $r^{(0)} = b - Ax^{(0)}$,
- first direction of search, e. g. $d^{(0)} = r^{(0)} \in \mathbb{R}^n$.

Conjugate gradient method: step $i \to i+1$

$$
\begin{aligned}
t^* &= \frac{(d^{(i)})^T r^{(i)}}{(d^{(i)})^T A d^{(i)}}, \\
x^{(i+1)} &= x^{(i)} + t^* d^{(i)}, \\
r^{(i+1)} &= b - Ax^{(i+1)}, \\
s^* &= -\frac{(r^{(i+1)})^T A d^{(i)}}{(d^{(i)})^T A d^{(i)}}, \\
d^{(i+1)} &= r^{(i+1)} + s^* d^{(i)}.
\end{aligned}
$$

Theorem 6.14

The conjugate gradient method terminates with the exact solution after n iterations at the latest.

Example

Consider the boundary value problem:

$$\Delta u = u_{xx}(x,y) + u_{yy}(x,y) = -1, \quad \text{für } (x,y) \in \Omega = (0,1)^2,$$
$$u(x,y) = 0, \quad (x,y) \in \partial\Omega.$$

Discretization results in the system of linear equations

$$A = \frac{1}{h^2}\begin{pmatrix} T & -I & 0 & \cdots & 0 & 0 \\ -I & T & -I & \cdots & 0 & 0 \\ & \ddots & \ddots & \ddots & \vdots & \\ 0 & 0 & \cdots & -I & T & -I \\ 0 & 0 & 0 & \cdots & -I & T \end{pmatrix} = h^{-2}\text{tridiag}(-I, T, -I) \in \mathbb{R}^{m^2 \times m^2}$$

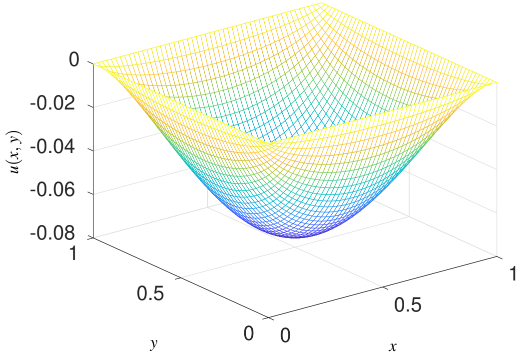
with $h = 1/(m+1)$ and I being the $m \times m$ unity matrix as well as

$$T = \begin{pmatrix} 4 & -1 & 0 & \cdots & 0 & \\ -1 & 4 & -1 & \cdots & 0 & \\ & \ddots & \ddots & \ddots & \vdots & \\ 0 & 0 & \cdots & -1 & 4 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 4 \end{pmatrix} = \text{tridiag}(-1, 4, -1).$$

Example

Solution for $m = 50$, thus $A \in \mathbb{R}^{2500\times 2500}$

Approximation to the solution $u(x, y)$ for $\Delta u = -1$.

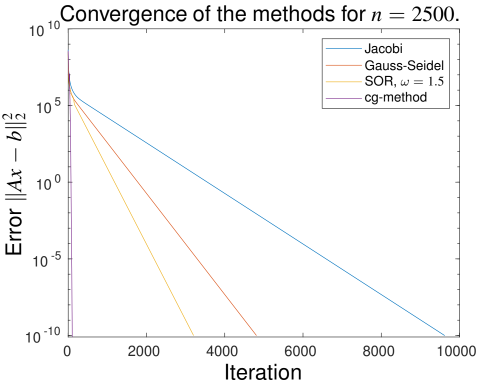


Example

Comparison of several methods:

- stopping criterion $\|Ax - b\|_2^2 < 10^{-10}$,
- simple implementation in MATLAB USING sparse,

Dimension		Iterations			Comp. time		
m	n	Jacobi	G-S	CG	Jacobi	G-S	CG
50	2500	9622	4812	104	0.29s	7.8s	0.1s
100	10 000	39166	19585	213	47.7s	579.7s	0.29



Outline

Outline

Least squares problems

General form:

- given data (x_i, y_i), $i = 1, \ldots, m$,
- set of functions $\varphi_1(x), \ldots, \varphi_n(x)$ with $n \leq m$,
- compute coefficients $\alpha_1, \ldots, \alpha_n$ such that the linear combination

$$f(x) = \sum_{i=1}^{n} \alpha_i \varphi_i(x)$$

  approximates the data as best as possible,
- evaluation via least-squares

$$\sum_{i=1}^{m} \left(y_i - f(x_i)\right)^2 \to \min!$$

Least squares problems

General matrix notation:

$$A = \begin{pmatrix} \varphi_1(x_1) & \cdots & \varphi_n(x_1) \\ \vdots & & \vdots \\ \varphi_1(x_m) & \cdots & \varphi_n(x_m) \end{pmatrix}, \qquad b = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

Simple example:

- regression line to $(x_1, y_1), \ldots, (x_m, y_m)$,
- approach $y(x) = \alpha_1 + \alpha_2 x$,
- resulting least-squares problem (matrix notation)

$$\left.\begin{aligned} \alpha_1 + \alpha_2 x_1 &= y_1 \\ \alpha_1 + \alpha_2 x_2 &= y_2 \\ &\vdots \\ \alpha_1 + \alpha_2 x_m &= y_m \end{aligned}\right\} \quad \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

Outline

Necessary condition for a minimum

Theorem 7.1

An x^* is a solution of the least-squares problem, if and only if x^* is a solution of

$$A^T A x = A^T b.$$

Remarks:

- if $\mathrm{rg}(A) = n$, then the solution is unique,
- if $\mathrm{rg}(A) < n$, then the set of solutions is a linear subspace,
- the approach to compute x^* as the solution of $A^T A x = A^T b$ is numerically susceptible to rounding errors.

Badly conditioned problem

<u>Example</u>:

- consider the matrix

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

  for $\varepsilon > 0$ and with $\mathrm{rg}(A) = 4$,
- the condition numbers for A and $A^T A$ are

$$\mathrm{cond}_\infty(A) = \left(1 + \frac{3}{|\varepsilon|}\right) \max(4, |\varepsilon|),$$
$$\mathrm{cond}_\infty(A^T A) = \left(\frac{3 + |\varepsilon^2 + 3|}{|\varepsilon|^2}\right) \max(4, 3 + |\varepsilon^2 + 1|)$$

  which are very different for small ε,
- rounding errors increase unnecessarily strongly with small ε for $A^T A$.

QR-decomposition to solve least squares problems

Theorem 7.2 (QR-decomposition)

Let $A \in \mathbb{R}^{m \times n}$. There exist an orthogonal matrix $Q \in \mathbb{R}^{m \times m}$ as well as an upper triangular matrix $R \in \mathbb{R}^{m \times n}$ with

$$A = QR.$$

Theorem 7.3 (Orthogonal invariance of the Euclidean norm)

If $Q \in \mathbb{R}^{n \times n}$ is an orthogonal matrix, then

$$\|Qx\|_2 = \|x\|_2.$$

Application to the least-squares problem:

- orthogonal invariance of the Euclidean norm results in

$$\|Ax - b\|_2 = \|Rx - Q^T b\|_2,$$

- the problem has the form

$$\left\| \begin{pmatrix} \tilde{R} \\ 0 \end{pmatrix} x - \begin{pmatrix} \tilde{b} \\ \hat{b} \end{pmatrix} \right\| \quad \text{with} \quad Q^T b = \begin{pmatrix} \tilde{b} \\ \hat{b} \end{pmatrix},$$

- the second part of the error $\|0x - \hat{b}\|_2$ is fixed, the first part is solvable and

$$Rx = \tilde{b}$$

  is the solution of the least-squares problem.

Definition 7.4

Let $u \in \mathbb{R}^n$ with $\|u\|_2 = 1$. A Householder matrix has the form

$$H = I - 2uu^T.$$

Computation of Q:

- sequence of Householder transformations,
- select $\alpha_1 = -\text{sign}(a_{11})\|a_1\|_2$ and compute

$$u_1 = \frac{a_1 - \alpha_1 e_1}{\|a_1 - \alpha_1 e_1\|}, \quad H_1 = I - 2u_1 u_1^T \qquad \Rightarrow \qquad H_1 A = \begin{pmatrix} \alpha_1 & * \\ 0 & A^{(1)} \end{pmatrix},$$

- do the same for $A^{(1)}$ and so on,
- the computation of n Householder transformations leads to

$$\underbrace{H_k H_{k-1} \cdot \ldots \cdot H_2 H_1}_{Q^T} A = R = \begin{pmatrix} \tilde{R} \\ 0_{(m-n)\times n} \end{pmatrix}.$$

Theorem 7.5 (Singular value decomposition)

Let $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, then there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ as well as a diagonal matrix $D \in \mathbb{R}^{m\times n}$, such that

$$A = UDV^T$$

with

$$D = \begin{pmatrix} \tilde{D} \\ 0 \end{pmatrix}, \qquad \tilde{D} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n).$$

A factorization of this form is called a singular value decomposition of A with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ being the singular values of A.

Singular value decomposition to solve least squares problems

Definition 7.6

Let $A = UDV^T$ for $A \in \mathbb{R}^{m\times n}$. The matrix $A^+ = VD^+U^T$ is called a Moore-Penrose inverse of A. Therein for

$$D = \begin{pmatrix} \tilde{D} \\ 0 \end{pmatrix}$$

the matrix D^+ is defined as

$$D^+ = (\tilde{D}^+, 0), \qquad (\tilde{D}^+)_{ij} = \begin{cases} \frac{1}{\sigma_i} & \text{for } i = j \text{ and } \sigma_i > 0, \\ 0 & \text{otherwise} \end{cases}.$$

Theorem 7.7

Let A^+ be a Moore-Penrose inverse of $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. Then

$$x^* = A^+ b$$

is a solution of the least-squares problem $\|Ax - b\|_2 \to \min$ and has the smallest Euclidean length of all solutions of the least-squares problem.

Outline

Example: eigenmodes of a guitare sound box.

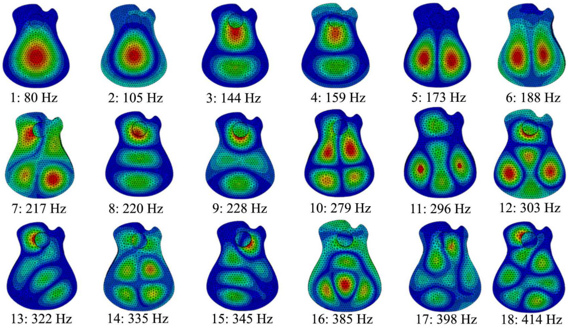


Figure 6. Mode shapes and natural frequency of guitar sound box.

J. Clinton, K. Wani, Open Journal of Acoustics 10(03): 41-50 (2020)

Eigenvalues

Definition 8.1

Let $A \in \mathbb{R}^{n \times n}$. A $\lambda \in \mathbb{C}$ is called eigenvalue of A, if there exists an $x \in \mathbb{C}^n$ with $x \neq 0$, so that

$$Ax = \lambda x.$$

Such an $x \neq 0$ is called a corresponding eigenvector. An eigenvalue and a corresponding eigenvector are called an eigenpair. The set of all eigenvalues of A is the spectrum $\sigma(A)$.

Remark 6

If $x \neq 0$ is an eigenvector to the corresponding eigenvalue λ and α is a scalar with $\alpha \neq 0$, then also αx is an eigenvector as

$$A(\alpha x) = \alpha Ax = \alpha\lambda x = \lambda(\alpha x).$$

Eigenvalue criterion (see linear algebra):

$$\lambda \text{ is an eigenvalue of } A \Leftrightarrow \det(A - \lambda I) = 0.$$

Remarks:

1. Main Theorem of algebra: a polynomial equation of degree n has exactly n roots, counting multiplicity. If all n roots are different, all eigenvalues are simple.
2. For a symmetric matrix A with real-valued entries, all eigenvalues are real and A is diagonalizable. In applications, discretization matrices are often symmetric.
3. If A is symmetric, eigenvectors to different eigenvalues are orthogonal with respect to the Euclidean scalar product. Hence one can pick n orthogonal eigenvectors of norm 1 and put them in an orthogonal matrix V with $VV^T = V^TV = I$.

Definition 8.2

Let $A \in \mathbb{R}^{n\times n}$. The Gershgorin's circles (disks, i.e. including the interior, not just the line) are

$$K_i = \left\{ z \in \mathbb{C} : \; |z - a_{ii}| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}| \right\}.$$

The circle K_i corresponding to the ith row of A has its center at the ith element of diagonal of A and the radius is the sum of the absolute values of all remaining elements in this row.

Theorem 8.3 (Gershgorin's circle Theorem)

Let $A \in \mathbb{R}^{n\times n}$. All eigenvalues of A are contained in the union of the n closed Gershgorin's disks

$$\sigma(A) \subset \bigcup_{i=1}^{n} K_i.$$

Outline

Idea for the power method:

- let $x \neq 0$ be arbitrary with the representation by the eigenvectors $v_1, \ldots, v_n$ as

$$x = \sum_{i=1}^{n} \alpha_i v_i,$$

- multiplication of A with x results in

$$Ax = A \sum_{i=1}^{n} \alpha_i v_i = \sum_{i=1}^{n} \alpha_i \underbrace{Av_i}_{=\lambda_i v_i} = \sum_{i=1}^{n} \alpha_i \lambda_i v_i,$$

- if $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$, then

$$A^k x = \lambda_1^k \Big(\alpha_1 v_1 + \alpha_2 \Big(\underbrace{\frac{\lambda_2}{\lambda_1}}_{|\cdot|<0} \Big)^k v_2 + \ldots + \alpha_n \Big(\underbrace{\frac{\lambda_n}{\lambda_1}}_{|\cdot|<0} \Big)^k v_n \Big) \xrightarrow{k\to\infty} \lambda_1^k \alpha_1 v_1.$$

- attention: Normalization after each step to avoid overflow.

Power method (survival of the largest)

Resulting algorithm of the power method:

- select a starting vector $x \neq 0$ with $\|x\|_2 = 1$, e. g. $x = n^{-0.5}(1, \ldots, 1)^T$,

$$\begin{aligned} y &:= Ax \\ r &:= y^T x \\ x &:= y/\|y\|_2 \end{aligned}$$

- r is an approximation for the eigenvalue with the largest magnitude and x is an approximation to its associated eigenvector,
- stopping criterion for example if $\|rx - y\|_2 < \varepsilon = 10^{-7}$,
- returns only one eigenvalue and the iteration can be computationally intensive.

Outline

Idea for the inverse power method:

- if λ is an eigenvalue of A, then $\frac{1}{\lambda}$ is an eigenvalue of A^{-1},
- for regular A holds

$$\begin{aligned} A: &\quad |\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n| \\ A^{-1}: &\quad \frac{1}{|\lambda_1|} \leq \frac{1}{|\lambda_2|} \leq \ldots \leq \frac{1}{|\lambda_n|}, \end{aligned}$$

- apply the power method to A^{-1} to determine the largest eigenvalue of A^{-1}, which is the smallest of A.

Resulting algorithm for inverse power method:

- select a starting vector $x \neq 0$ with $\|x\|_2 = 1$,

$$\text{solve } Ay = x$$
$$r := \frac{1}{y^T x}$$
$$x := y / \|y\|_2$$

Shifted inverse power method (survival of the most popular)

Lemma 8.4

If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then $\lambda_1 - \mu, \ldots, \lambda_n - \mu$ are the eigenvalues of $A - \mu I$.

Combination of statements:

- if μ is an approximation to λ_i, then $\lambda_i - \mu \approx 0$ is an eigenvalue of

$$A - \mu I,$$

- this eigenvalue can be determined by inverse vector iteration, since

$$1/(\lambda_i - \mu)$$

  is the largest eigenvalue of $(A - \mu I)^{-1}$,
- avoid to calculate an inverse by solving systems of linear equations.

Shifted inverse power method (survival of the most popular)

Resulting algorithm (inverse power method with spectral displacement):

- select a starting vector $x \neq 0$ with $\|x\|_2 = 1$ and a shift μ, e. g. an element on the diagonal of A, see Gershgorin,

$$\begin{aligned} &\text{solve } (A - \mu I)y = x \\ &r := \frac{1}{y^T x} + \mu \\ &x := y/\|y\|_2 \end{aligned}$$

Remarks:

- optionally, $\mu := r$ can also be used as an update in each iteration, although it is advisable not to do this in the initial steps steps,
- convergence rate of the iteration depends on the second largest eigenvalue of $(A - \mu I)^{-1}$.

Outline

Preliminary consideration:

- for an upper or lower triangular matrix, the eigenvalues are on the main diagonal,
- for a similarity transformation
$$T^{-1}AT$$
with a regular $T \in \mathbb{R}^{n\times n}$, the eigenvalues of A and $T^{-1}AT$ are equal.

Goal:

- achieving a triangular shape through a sequence of similarity transformations

$$\begin{aligned} A^{(0)} \quad &\rightarrow \quad A^{(1)} = (T^{(0)})^{-1}A^{(0)}T^{(0)} \\ &\rightarrow \quad A^{(2)} = (T^{(1)})^{-1}A^{(1)}T^{(1)} \\ &\qquad\qquad = \underbrace{(T^{(1)})^{-1}(T^{(0)})^{-1}}_{\widetilde{T}^{-1}} A^{(0)} \underbrace{T^{(0)}T^{(1)}}_{\widetilde{T}} \\ &\quad \dots \end{aligned}$$

QR-algorithm:

- set $A^{(0)} = A$,
- compute an QR-decomposition $A^{(0)} = Q^{(0)} R^{(0)}$,
- compute

$$A^{(1)} = (Q^{(0)})^T A^{(0)} Q^{(0)} = R^{(0)} Q^{(0)},$$

- iteration for $k \geq 1$

$$A^{(k)} = Q^{(k)} R^{(k)} \quad \to \quad A^{(k+1)} = R^{(k)} Q^{(k)}.$$

QR-algorithm

Theorem 8.5

Let $A \in \mathbb{R}^{n\times n}$ with the eigenvalues $\lambda_1, \ldots, \lambda_n$ in decreasing order, thus $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. The series $\{A^{(k)}\}_{k=0,1,2,\ldots}$ with

$$A^{(k+1)} = R^{(k)} Q^{(k)}$$

for $A^{(k)} = Q^{(k)} R^{(k)}$ convergences to an upper triangular matrix, thus

$$\lim_{k\to\infty} A^{(k)} = \begin{pmatrix} \lambda_1 & * & * & \cdots & * \\ 0 & \lambda_2 & * & \cdots & * \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & & \cdots & 0 & \lambda_n \end{pmatrix}.$$

Eigenvectors for symmetric matrix A:

- $Q = Q^{(0)} Q^{(1)} \cdot \ldots \cdot Q^{(k)}$ is an orthogonal matrix,
- thus $Q^T A Q = A^{(k)}$ is a similarity transformation and the eigenvetcors are in the columns of Q.

Outline

Idea of Jacobi- method:

- Transform the symmetric matrix A into the diagonal form by a sequence of similarity transformations,
- practically, the sequence is aborted when the off-diagonal component

$$\mathrm{off}(B) = \sum_{i,j=1, i\neq j}^{n} b_{ij}^2,$$

of the transformed matrix B is sufficiently small,
- the eigenvalues of B agree with the eigenvalues of A except for small deviations depending on $\mathrm{off}(B)$.

One step:

- similarity transformation

$$A \quad \rightarrow \quad B = J^T A J \tag{8}$$

- Jacobi rotation $J(k, l, \theta) \in \mathbb{R}^{n \times n}$ defined as

$$\big(J(k, l, \theta)\big)_{ij} = \begin{cases} 1 & \text{if } i = j \text{ but } i \notin \{k, l\}, \\ c & \text{for } (i,j) = (k,k) \text{ or } (i,j) = (l,l), \\ -s & \text{for } (i,j) = (k,l), \\ s & \text{for } (i,j) = (l,k), \\ 0 & \text{else.} \end{cases}$$

for $1 \leq k, l \leq n$, $c = \cos\theta$, $s = \sin\theta$.

Theorem 8.6

For the rotation from (8) holds

$$\operatorname{off}(B) = \operatorname{off}(A) - 2a_{kl}^2 + 2b_{kl}^2.$$

Goal:

- greatest possible reduction of $\operatorname{off}(A)$ if $b_{kl} = 0$,
- if $a_{kl} \neq 0$, then use

$$\rho = \frac{a_{ll} - a_{kk}}{2a_{kl}}$$

  as well as

$$t = \begin{cases} \frac{1}{\rho + \sqrt{1+\rho^2}} & \text{if } \rho \geq 0, \\ \frac{1}{\rho - \sqrt{1+\rho^2}} & \text{if } \rho < 0, \end{cases}$$

  (to avoid catastrophic cancellation),
- compute

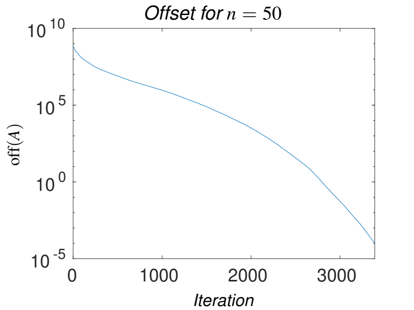
$$c := \frac{1}{1+t^2}, \qquad s = tc.$$

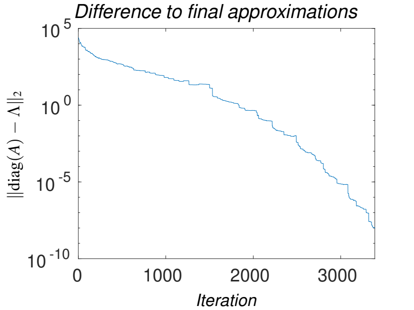
Example

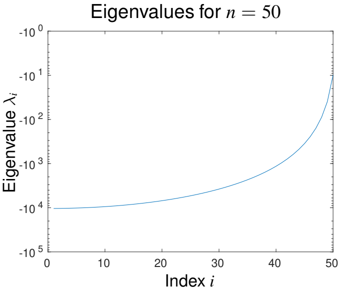
We apply Jacobi's method to compute the eigenvalues of the matrix

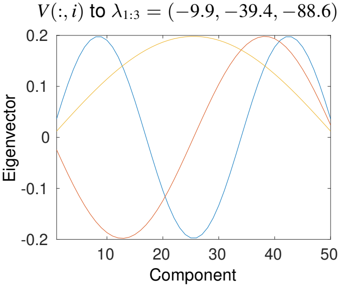
$$A = (n+1)^2 \text{tridiag}(1, -2, 1) \in \mathbb{R}^{n \times n}$$

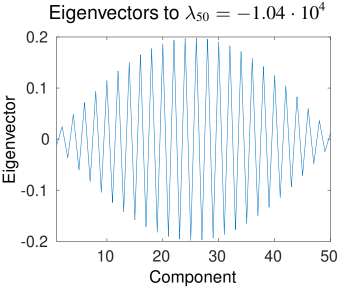
for $n = 50$.






Eigenvalues for $n = 50$
Eigenvalue λ_i
-10^0
-10^1
-10^2
-10^3
-10^4
-10^5
0 10 20 30 40 50
Index i


$V(:, i)$ to $\lambda_{1:3} = (-9.9, -39.4, -88.6)$
Eigenvector
0.2
0.1
0
-0.1
-0.2
10 20 30 40 50
Component


Eigenvectors to $\lambda_{50} = -1.04 \cdot 10^4$
Eigenvector
0.2
0.1
0
-0.1
-0.2
10 20 30 40 50
Component

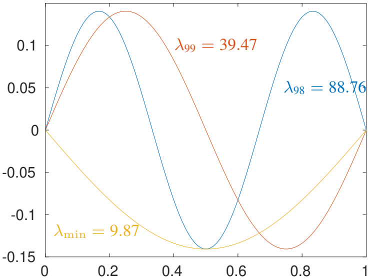
Application to 1D vibrating strings:

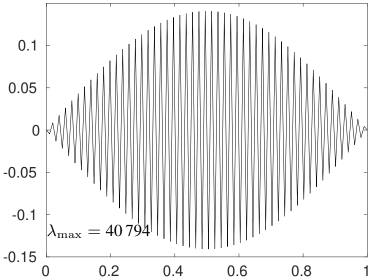
- solution of $u_{tt}(x,t) = c^2 u_{xx}(x,t)$ for $x \in \Omega = (0,1)$ by a factorial approach,
- search for eigenfunctions/eigenvibrations and eigenvalues of

$$u''(x) = -\lambda u(x) \quad \text{for } x \in (0,1), \qquad u(0) = u(1) = 0,$$

- discretization by n interior nodes for $h = 1/(n+1)$ results in the eigenvalue problem

$$Av = -\lambda v, \quad \text{with} \quad A = \frac{1}{h^2}\text{tridiag}(1,-2,1) \in \mathbb{R}^{n \times n}.$$





Approximation of eigenvalues - example 2

2D oscillation using the example of a circular area:

- Boundary value problem with a partial differential equation, see pde-lecture,

$$\Delta u(x,y) = c^2 u_{tt}(x,y), \qquad \text{differential equation,}$$

$$u(x,y) = 0 \text{ for } (x,y) \in \partial\Omega, \qquad \text{boundary condition,}$$

- domain $\Omega = \{(x,y) \in \mathbb{R}^2 : \ x^2 + y^2 < 1\}$ (circle),
- Numerical solution via finite element method.