

SCRIPTUM TO THE CLASS

ADVANCED TOPICS

IN NUMERICAL ANALYSIS

Part 1

WINTER SEMESTER 2006/2007

BY

PROF. DR. RUDOLF SCHERER

INSTITUT FÜR ANGEWANDTE UND NUMERISCHE MATHEMATIK
DER UNIVERSITÄT KARLSRUHE (TH)

Contents

I	Interpolation, Approximation and Quadrature	3
1	General Interpolation Problem	3
2	Trigonometric Interpolation	11
3	Fourier Transform	17
4	Cubic Spline Interpolation	27
5	Gaussian Quadrature Formulas	34
6	Extrapolation Methods	43
II	Eigenvalue Problems for Matrices	53
7	Bounds for the Eigenvalues	53
8	Eigenvalues of Symmetric Matrices	60
9	Reduction Method of Householder	66
10	Methods of Givens and Jacobi	71
11	Vector Iteration of Mises and Wielandt	77
12	LR and QR method	81
III	Numerical Treatment of Ordinary Differential Equations	91
13	Basic Ideas	91
14	Discretization Methods	96
15	Runge–Kutta Methods	100
16	Linear Multistep Methods of Adams	110
17	Asymptotic Stability and Convergence	114
18	Absolute Stability	122

Many thanks to Michael Lehn for preparing the figures and for valuable comments.

BIBLIOGRAPHY

- U.M. Ascher, L.R. Petzold:** *Computer Methods for Ordinary Differential Equations and Differential–Algebraic Equations*. SIAM, 1998.
- A. Björck:** *Numerical Methods for Least Squares Problems*. SIAM, 1996.
- J.C. Butcher:** *The Numerical Analysis of Ordinary Differential Equations*. John Wiley & Sons, 1987.
- P.J. Davis, P. Rabinowitz:** *Methods of Numerical Integration*. Academic Press, 1984.
- K. Dekker, J.G. Verwer:** *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations*. North Holland, 1984.
- W. Gautschi:** *Numerical Analysis*. Birkhäuser, 1997.
- G.H Golub, C. Van Loan:** *Matrix Computations (3rd edition)*. John Hopkins University Press, 1996.
- E. Hairer, S.P. Norsett, G. Wanner:** *Solving Ordinary Differential Equations I*. Springer, 1993/2000.
- G. Hämmerlin, K.-H. Hoffmann:** *Numerische Mathematik*. Springer, 1994.
- M. Hanke–Bourgeois:** *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*. Teubner, 2002.
- N.H. Higham:** *Accuracy and Stability of Numerical Algorithms (2nd edition)*. SIAM, 2002.
- A. Iserles:** *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.
- R. Kress:** *Numerical Analysis*. Springer, 1998.
- A. Quateroni, A. Valli:** *Numerische Mathematik*. Springer, 2002.
- H.R. Schwarz:** *Numerische Mathematik*. Teubner, 1986.
- J. Stoer, R. Bulirsch:** *Introduction to Numerical Analysis (2nd edition)*. Springer, 1996. [This English version contains the German version of Stoer (Part 1) and of Stoer & Bulirsch (Part 2)]
- L.N. Trefethen, D. Bau:** *Numerical Linear Algebra*. SIAM, 1997.

Chapter I

Interpolation, Approximation and Quadrature

1 General Interpolation Problem

$K \in \{\mathbb{R}, \mathbb{C}\}$, $D \subseteq K$ (mostly $K = \mathbb{R}$, D with at least n elements)

V n -dimensional *vector space* of real (or complex)-valued functions in D

Basis of V : $\{v_1, \dots, v_n\}$, $p = \sum_{\nu=1}^n \alpha_\nu v_\nu$

Given: *Nodes* $z_1, \dots, z_n \in D$ (distinct points)

Data $w_1, \dots, w_n \in K$ (function values $w_j := f(z_j)$)

Wanted: $p \in V$ satisfying $p(z_j) = w_j$, $j = 1, \dots, n$

Questions: Existence, uniqueness, construction of p ?

V **Haar space** respectively $\{v_1, \dots, v_n\}$ **Haar system:**

each $0 \neq p \in V$ has at most $n - 1$ zeros in D

Examples : Haar systems

$\{1, x, \dots, x^n\}$ in $\mathbb{C}$

$\{1, \cos x, \sin x, \dots, \cos nx, \sin nx\}$ in $[0, 2\pi)$

$\{e^{-inz}, \dots, 1, \dots, e^{inz}\}$ in $[0, 2\pi)$

$\{1, \cos x, \dots, \cos nx\}$ in $[0, \pi]$

$\{\sin x, \dots, \sin nx\}$ in $(0, \pi)$

Space $S_3(\Delta)$ of the cubic splines (see §4),

$\{1, x^2\}$ in $[-1, 1]$: *It is not a Haar system*

Theorem 1.1 (*Interpolation Criteria*)

1. The unique solvability of the interpolation problem is given if and only if the homogeneous problem $p(z_j) = 0$, $j = 1, \dots, n$, has only the trivial solution.
2. The unique solvability of the interpolation problem is given if and only if V is a Haar space.

Interpolation with algebraic polynomials

$$V = \mathcal{P}_n \text{ (dim} = n + 1) : p \in \mathcal{P}_n \text{ with } p(x_j) = y_j, j = 0, 1, \dots, n$$

$$\text{Lagrange representation } p = \sum_{j=0}^n y_j \ell_j, \ell_j(x_k) = \delta_{jk}$$

Newton representation

$$p = \sum_{j=0}^n a_j w_j, w_j(x) = \prod_{\nu=0}^{j-1} (x - x_\nu), w_0(x) = 1$$

Divided differences

$$a_j := [x_0, \dots, x_j] \text{ (recurrence definition, invariant under permutations)}$$

Scheme

$$\begin{array}{ccccccc} x_0 & [x_0] & & & & & \\ & & [x_0, x_1] & & & & \\ x_1 & [x_1] & & [x_0, x_1, x_2] & & & \\ & & [x_1, x_2] & & \ddots & & \\ x_2 & [x_2] & & & & [x_0, \dots, x_n] & \\ \vdots & \vdots & \vdots & & & & [x_0, \dots, x_{n+1}] \\ x_n & [x_n] & & & & & \\ & & & & & & \\ x_{n+1} & [x_{n+1}] & & & & & \end{array}$$

Gregory–Newton: Uniformly spaced nodes $x_j = x_0 + jh$ ($h > 0$)

Forward differences

$$\left. \begin{array}{l} \Delta^0 y_\nu := y_\nu \\ \Delta^k y_\nu := \Delta^{k-1} y_{\nu+1} - \Delta^{k-1} y_\nu \end{array} \right\} [x_\nu, \dots, x_{\nu+k}] = \frac{1}{h^k k!} \Delta^k y_\nu, k = 1, 2, \dots$$

Scheme

$$\begin{array}{ccccccc} \Delta^0 y_0 & & & & & & \\ & \Delta^1 y_0 & & & & & \\ \Delta^0 y_1 & & \Delta^2 y_0 & & & & \\ & \Delta^1 y_1 & & & & & \\ \Delta^0 y_2 & \vdots & \vdots & \ddots & & & \\ & \vdots & & & & & \end{array}$$

Representation of the interpolation polynomial

$$\boxed{p(x) = \sum_{k=0}^n \binom{t}{k} \Delta^k y_0, \quad t = \frac{x - x_0}{h}}$$

Backward differences (sequence of the nodes $x_n, x_{n-1}, \dots, x_0$)

$$\left. \begin{array}{l} \nabla^0 y_\nu := y_\nu \\ \nabla^k y_\nu := \nabla^{k-1} y_\nu - \nabla^{k-1} y_{\nu-1} \end{array} \right\} [x_\nu, \dots, x_{\nu-k}] = \frac{1}{h^k k!} \nabla^k y_\nu, k = 1, 2, \dots$$

Scheme

$$\begin{array}{ccccccc}
\nabla^0 y_0 & & & & & & \\
& \nabla^1 y_1 & & & & & \\
\nabla^0 y_1 & & \nabla^2 y_0 & & & & \\
& \nabla^1 y_2 & & \ddots & & & \\
\nabla^0 y_2 & \vdots & \nabla^2 y_n & & \nabla^n y_n & & \\
& \vdots & & & \nabla_{n+1}^n & & \\
& \nabla^1 y_n & & & & & \\
\nabla^0 y_n & & \nabla^1 y_{n+1} & & & & \\
\nabla^0 y_{n+1} & & & & & &
\end{array}$$

Representation of the interpolation polynomial

$$p(x) = \sum_{k=0}^n (-1)^k \binom{-t}{k} \nabla^k y_n, \quad t = \frac{x - x_n}{h}$$

Application: Construction of linear multi-step methods (Chap. III)

Algorithm of Aitken–Neville (see Numer. Math I)

Evaluation of the interpolation polynomial at a fixed point x by recurrence:

$$\begin{aligned}
p_j(x) &:= y_j, \quad j = 0, 1, \dots \\
p_{j, \dots, j+k}(x) &:= \frac{p_{j+1, \dots, j+k}(x)(x - x_j) + p_{j, \dots, j+k-1}(x)(x_{j+k} - x)}{x_{j+k} - x_j}
\end{aligned}$$

$$\begin{array}{c|ccccccc}
\text{Scheme} & x_0 & y_0 =: p_0(x) & & & & \\
& & & p_{01}(x) & & & \\
& x_1 & y_1 =: p_1(x) & & p_{012}(x) & & \\
& & & p_{12}(x) & & \ddots & \\
& \vdots & \vdots & & & & \boxed{p_{01\dots n}(x) = p(x)} \\
& x_n & y_n =: p_n(x) & & & &
\end{array}$$

Hermite Interpolation

Given: Nodes $x_0, x_1, \dots, x_n$

Data $y_0, y_1, \dots, y_n; y_0^{(1)}, y_1^{(1)}, \dots, y_n^{(1)}$

Wanted: $p \in \mathcal{P}_{2n+1}$ satisfying $p(x_j) = y_j, p'(x_j) = y_j^{(1)}, j = 0, 1, \dots, n$

Interpolation criterion: The homogeneous problem has only the trivial solution, because $0 \neq p \in \mathcal{P}_{2n+1}$ has at most $2n + 1$ zeros!

Lagrange representation

$$p = \sum_{\nu=0}^n U_{\nu} y_{\nu} + \sum_{\nu=0}^n V_{\nu} y_{\nu}^{(1)},$$

$$U_{\nu}(x) := \{1 - 2\ell'_{\nu}(x_{\nu})(x - x_{\nu})\}\ell_{\nu}^2(x), \quad U_{\nu}(x_j) = \delta_{\nu j}, \quad U'_{\nu}(x_j) = 0$$

$$V_{\nu}(x) := (x - x_{\nu})\ell_{\nu}^2(x), \quad V_{\nu}(x_j) = 0, \quad V'_{\nu}(x_j) = \delta_{\nu j}$$

Newton representation

$$p = \sum_{j=0}^{2n+1} c_j \hat{\omega}_j, \quad \hat{\omega}_j = \begin{cases} \omega_{j/2}^2, & j \text{ even} \\ \omega_{\frac{j-1}{2}} \omega_{\frac{j+1}{2}}, & j \text{ odd} \end{cases}$$

$$\omega_0(x) = 1, \quad \omega_j(x) = (x - x_0) \cdots (x - x_{j-1})$$

$$\hat{\omega}_0(x) = 1, \quad \hat{\omega}_1(x) = (x - x_0), \quad \hat{\omega}_2(x) = (x - x_0)^2, \dots$$

Confluent nodes

$$f \in C^1[a, b], f(x_j) =: y_j, \quad f[x_{\nu}, x_{\nu}] := \lim_{h \rightarrow 0} \frac{f(x_{\nu} + h) - f(x_{\nu})}{h} = f'(x_{\nu})$$

$$\text{Recurrence} \quad [x_{\nu}] := y_0, [x_{\nu}, x_{\nu}] := y_0^{(1)}, [x_{\nu}, x_{\nu}, x_{\nu+1}] = \frac{[x_{\nu}, x_{\nu+1}] - [x_{\nu}, x_{\nu}]}{x_{\nu+1} - x_{\nu}}, \dots$$

Scheme $x_0 \quad y_0$

$$\begin{array}{ccccccc} & & & & y_0^{(1)} & & \\ & & & & & & \\ x_0 & y_0 & & [x_0, x_0, x_1] & & & \\ & & [x_0, x_1] & & [x_0, x_0, x_1, x_1] & & \\ x_1 & y_1 & & [x_0, x_1, x_1] & & [x_0, x_0, x_1, x_1, x_2] & \\ & & y_1^{(1)} & & [x_0, x_1, x_1, x_2] & & \\ x_1 & y_1 & & [x_1, x_1, x_2] & & \vdots & \\ & & [x_1, x_2] & & \vdots & & \\ x_2 & y_2 & & & & & \end{array}$$

Example 1: Nodes $x_0 = -1, x_1 = 1$

Data $y_0 = -2, y_1 = 2, y_0^{(1)} = y_1^{(1)} = 6$

Scheme $-1 \quad -2$

$$\begin{array}{ccccccc} & & & & \boxed{6} & & \\ & & & & & & \\ -1 & -2 & & \boxed{-2} & & & \\ & & 2 & & \boxed{2} & & \\ 1 & 2 & & 2 & & & \\ & & 6 & & & & \\ 1 & 2 & & & & & \end{array}$$

$$p(x) = -2 + 6(x + 1) - 2(x + 1)^2 + 2(x + 1)^2(x - 1) = 2x^3$$

Example 2: Nodes $x_0 = -1, x_1 = 1$
 Data $y_0 = -2, y_1 = 2, y_1^{(1)} = 6, y_1^{(2)} = 12$
 $f[x_\nu, x_\nu, x_\nu] := \frac{f''(x_\nu)}{2!}$

Scheme

-1	-2		
		2	
1	2		2
		6	2
1	2		$\frac{12}{2}$
		6	
1	2		

$$p(x) = -2 + 2(x+1) + 2(x+1)(x-1) + 2(x+1)(x-1)^2 = 2x^3$$

The Peano–Sard representation

Integral representation of a linear functional R (see Peano: Numer. Math. I)

Functional $J : C^{m+1}[a, b] \rightarrow \mathbb{R}$

Functional $L : C[a, b] \rightarrow \mathbb{R}$

Functional $R := J - L$

Order m , i.e., $Rp = 0$ for all $p \in \mathcal{P}_m$

Examples: Interpolation, divided differences, differentiation, quadrature, ...

Theorem 1.2 (Peano–Sard)

The linear functional R of order m applied to $f \in C^{m+1}[a, b]$ satisfies

$$Rf = \int_a^b G(t) f^{(m+1)}(t) dt,$$

where

$$G(t) = Rv_t, \quad v_t(x) := \frac{(x-t)_+^m}{m!}$$

denotes the corresponding Peano kernel.

If $G(t)$ is definite in $[a, b]$, then it holds ($a < \xi < b$)

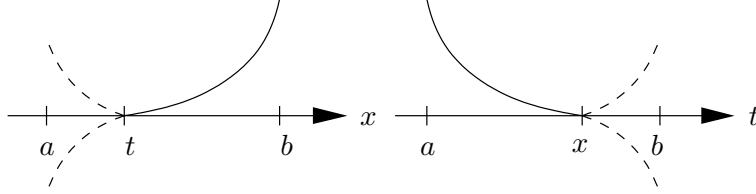
$$Rf = c_m f^{(m+1)}(\xi), \quad c_m = \int_a^b G(t) dt = \frac{1}{(m+1)!} Rh_{m+1}.$$

Proof is given later for the special case of divided differences. $\square$

Remarks

1. The kernel G is the *error function* corresponding to $v_t(x) = \frac{(x-t)_+^m}{m!}$,

$$\text{where } (x-t)_+^m := \begin{cases} (x-t)^m, & t \leq x \\ 0, & t > x \end{cases}$$



2. The kernel G is *definite*, i.e., $G(t) \geq 0$ or $G(t) \leq 0$ in $[a, b]$: mid-value theorem of integral theory
3. The constant c_m can easily be computed: $h_{m+1}(t) := t^{m+1}$

$$Rh_{m+1} = (m+1)! \int_a^b G(t) dt \quad \text{implies} \quad c_m = \frac{1}{(m+1)!} Rh_{m+1}.$$

4. Bounds of the type $|Rf| \leq c \|f^{(m+1)}\|$ (c independent of f)

$$\|\cdot\|_\infty : c = \int_a^b |G(t)| dt \quad (= \|G\|_1)$$

$$\|\cdot\|_1 : c = \max_{a \leq t \leq b} |G(t)| \quad (= \|G\|_\infty)$$

$$\|\cdot\|_2 : c = (\int_a^b G^2(t) dt)^{1/2} \quad (= \|G\|_2) \quad (\text{Schwarz' inequality})$$

5. Derivatives $f^{(\nu+1)}$, $\nu = 0, 1, \dots, m$

$$Rf = \int_a^b G_\nu(t) f^{(\nu+1)}(t) dt,$$

$$G_\nu(t) = Rv_t^\nu, \quad v_t^\nu(x) := \frac{(x-t)_+^\nu}{\nu!}$$

Special case: Divided differences

Corollary 1.3

$$L_k f = f[x_0, x_1, \dots, x_k] = \sum \mu_j f(x_j), \quad \mu_j = \prod_{\substack{\nu=0 \\ \nu \neq j}}^k (x_j - x_\nu)^{-1},$$

is of order $k-1$ and $L_k h_k = 1$.

Proof: Interpolation polynomial p_f of f with respect to $x_0, \dots, x_{k-1}$:

$$w(x) = (x - x_0) \cdots (x - x_{k-1}), \quad x_k \neq x_1, \dots, x_{k-1}$$

$$\text{Error (Numer. Math. I): } f(x_k) - p_f(x_k) = f[x_0, \dots, x_{k-1}, x_k] \underbrace{w(x_k)}_{\neq 0}$$

$$\ell \leq k-1 : f = h_\ell \implies p_f = h_\ell \implies L_k h_\ell = 0$$

$$\begin{aligned} f = h_k &\implies p_f = h_k - w \quad \text{since } h_k - w \in \mathcal{P}_{k-1} \\ &\quad \text{and } p_f(x_\nu) = h_k(x_\nu) - w(x_\nu) = h_k(x_\nu), \quad \nu = 0, \dots, k-1 \\ &\implies h_k(x_k) - p_f(x_k) = w(x_k) \implies L_k h_k = 1 \quad \square \end{aligned}$$

Corollary 1.4 L_k operating on the subspace $C^k[a, b]$ satisfies

$$L_k f = \int_a^b G(t) f^{(k)}(t) dt, \quad G(t) = L_k v_t, \quad v_t(x) = \frac{(x-t)_+^{k-1}}{(k-1)!}.$$

Proof: Taylor expansion in a using the integral remainder term: $f = p_{k-1} + R_k$

$$\begin{aligned} R_k(x) &= \int_a^b f^{(k)}(t) \frac{(x-t)_+^{k-1}}{(k-1)!} dt = \int_a^b f^{(k)}(t) v_t(x) dt \\ L_k f &= \underbrace{L_k p_{k-1}}_{=0} + L_k R_k = \sum_{j=0}^k \mu_j R_k(x_j) = \sum_{j=0}^k \mu_j \int_a^b f^{(k)}(t) v_t(x_j) dt \\ &= \int_a^b f^{(k)}(t) \underbrace{\sum_{j=0}^k \mu_j v_t(x_j)}_{=L_k v_t=:G(t)} dt. \quad \square \end{aligned}$$

Theorem 1.5 The Peano kernel of a divided difference L_k has the following properties:

- i) G is a spline of degree $k-1$ with respect to $x_0 < x_1 < \dots < x_k$,
- ii) G is identical to zero in $(-\infty, x_0)$ and (x_k, ∞) ,
- iii) G is strictly positive in (x_0, x_k) ,
- iv) $\int_{x_0}^{x_k} G(t) dt = \frac{1}{k!}$.

Proof: [E]

Corollary 1.6 The divided difference L_k satisfies

$$L_k f = \frac{1}{k!} f^{(k)}(\xi), \quad a < \xi < b.$$

SUPPLEMENTARY EXAMPLES – No. 1

INTERPOLATION, EXTRAPOLATION AND APPROXIMATION

Polynomials of higher order are not appropriate for extrapolation

Given: Number of the population in the USA from 1900 to 1990 every 10 years

Wanted: Prognosis for the year 1995 and 2000

Material of data: Number of population in Millions

1900	1910	1920	1930	1940	1950	1960	1970	1980	1990
75.995	91.972	105.711	123.203	131.696	150.697	179.323	203.212	226.505	248.7

Different possibilities:

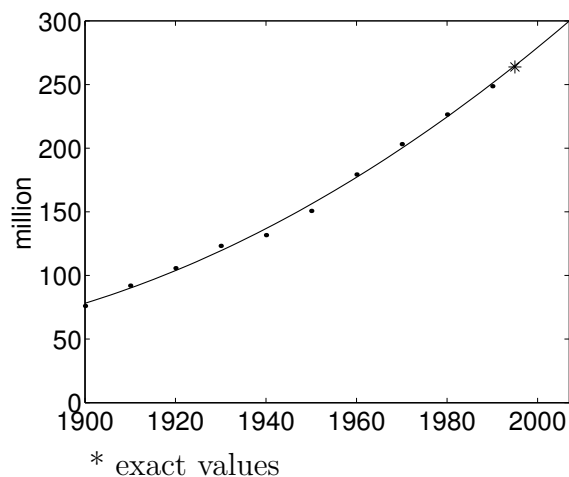
- The interpolation polynomial $p_9(x)$ with respect to the nodes 1900, 1910, ..., 1990.
- The construction of an interpolant cubic spline $s(x)$ with respect to these nodes and data.
- The construction of polynomials $p_j(x)$, $j = 1, \dots, 8$, by the principle of the least squares of errors.

For each case the values are extrapolated for 1995 and 2000, i.e., the functions $p_j(x)$, $j = 1, \dots, 9$, and $s(x)$ are evaluated at 1995 and 2000.

Numerical results

order	1995	2000
1	249.79	259.40
2	264.96	279.33
3	264.86	279.15
4	254.94	257.98
5	253.70	254.63
6	268.87	305.03
7	277.08	337.32
8	185.48	-77.12
9	243.33	218.13
cubic spline	258.39	266.62
exact value	263.8	273.8

order 3



Comments

$p_9(x)$ shows zigzag behaviour between 1980 and 2010;

$p_8(x)$ shows shortly before 2000 a zero point (i.e., the population of USA is zero);

$p_7(x)$ and $p_6(x)$ show an extreme increase after 1995;

$p_5(x)$ and $p_4(x)$ show a maximal point at 1995 and then falling down;

$p_3(x)$ and $p_2(x)$ show the realistic behaviour, as well as the cubic spline $s(x)$.

2 Trigonometric Interpolation

Periodic processes

All arguments are equivalent $\rightarrow$ *uniformly spaced nodes*

Haar system $\{1, \cos x, \sin x, \dots, \cos mx\}$ in $[0, 2\pi)$

Dimension $2m + 1$ (odd)

Trigonometric polynomial $T(x) = \frac{\alpha_0}{2} + \sum_{\nu=1}^m \{\alpha_\nu \cos \nu x + \beta_\nu \sin \nu x\}$

Degree m

Π_m : space of trigonometric polynomials of degree $\leq m$

Orthogonality $\int_0^{2\pi} \begin{matrix} \cos kx \\ \sin kx \end{matrix} \cdot \begin{matrix} \cos \ell x \\ \sin \ell x \end{matrix} dx = 0, k \neq \ell$

Fourier coefficients $\left. \begin{matrix} \hat{\alpha}_\nu \\ \hat{\beta}_\nu \end{matrix} \right\} = \frac{1}{\pi} \int_0^{2\pi} f(x) \begin{matrix} \cos \nu x \\ \sin \nu x \end{matrix} dx$

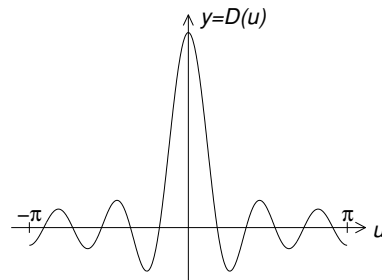
Dirichlet kernel $D(u) := \frac{\sin \frac{2m+1}{2}u}{\sin \frac{1}{2}u}, \quad u \in \mathbb{R} \quad (m \geq 1)$

$$D(0) = 2m + 1 \quad (\text{l'Hospital})$$

$$D\left(\frac{2(\nu-j)\pi}{2m+1}\right) = 0, \quad \nu \neq j \quad (\nu, j \in \mathbb{N})$$

$$D(u) = 1 + 2 \cos u + \dots + 2 \cos mu$$

Figure: $m = 5$



Interpolation problem

I. Odd number of uniformly spaced nodes

Given: Nodes $x_j = \frac{2j\pi}{2m+1}, j = 0, 1, \dots, 2m$

Data $y_j \in \mathbb{R}, j = 0, 1, \dots, 2m \quad (f \in C_{2\pi}, y_j := f(x_j))$

Wanted: $T \in \Pi_m$ satisfying $T(x_j) = y_j, \quad j = 0, 1, \dots, 2m$

The unique solvability for arbitrary nodes in $[0, 2\pi)$ (Haar space) is given.

The basic Lagrange functions (given by the Dirichlet kernel D):

$$t_j(x_\nu) = \begin{cases} 1, & \nu = j \\ 0, & \nu \neq j \end{cases} \quad \boxed{t_j(x) := \frac{1}{2m+1} D(x - x_j)}$$

Theorem 2.1 *The trigonometric interpolation polynomial with respect to the $2m + 1$ uniformly spaced nodes is given in Lagrange representation*

$$T(x) = \frac{1}{2m+1} \sum_{j=0}^{2m} y_j D(x - x_j).$$

Remark: Interesting representation, but not suitable for applications, better in the normal form!

Theorem 2.2 *The trigonometric interpolation polynomial $T(x)$ is given by*

$$\begin{aligned} T(x) &= \frac{\alpha_0}{2} + \sum_{\nu=1}^m \{\alpha_\nu \cos \nu x + \beta_\nu \sin \nu x\}, \\ \left. \begin{matrix} \alpha_\nu \\ \beta_\nu \end{matrix} \right\} &= \frac{2}{2m+1} \sum_{j=0}^{2m} y_j \begin{cases} \cos \nu x_j \\ \sin \nu x_j \end{cases}, \quad \nu = 0, 1, \dots, m. \end{aligned}$$

Remark: α_ν, β_ν are called **discrete Fourier coefficients**.

Proof: Transformation: Theorem 2.1 implies Theorem 2.2

$$D(x - x_j) = 1 + 2 \sum_{\nu=1}^m \cos \nu(x - x_j)$$

Trigonometric formulas: $\cos(\alpha \pm \beta) = \cos \alpha \cos \beta \mp \sin \alpha \sin \beta$
 $\sin(\alpha \pm \beta) = \pm \cos \alpha \sin \beta + \sin \alpha \cos \beta$

$$\begin{aligned} T(x) &= \frac{1}{2m+1} \sum_{j=0}^{2m} y_j \{1 + 2 \sum_{\nu=1}^m (\cos \nu x_j \cos \nu x + \sin \nu x_j \sin \nu x)\} \\ &= \underbrace{\frac{1}{2m+1} \sum_{j=0}^{2m} y_j}_{=:\frac{\alpha_0}{2}} + \underbrace{\frac{2}{2m+1}}_{\hookrightarrow} \sum_{\nu=1}^m \left\{ \underbrace{\sum_{j=0}^{2m} y_j \cos \nu x_j}_{=:\alpha_\nu} \cos \nu x + \underbrace{\sum_{j=0}^{2m} y_j \sin \nu x_j}_{=:\beta_\nu} \sin \nu x \right\} \quad \square \end{aligned}$$

Corollary 2.3 *The relations of discrete orthogonality are satisfied*

$$\begin{aligned} \frac{2}{2m+1} \sum_{j=0}^{2m} \cos kx_j \cos \ell x_j &= \begin{cases} 2, & k = \ell = 0 \\ 1, & k = \ell \neq 0 \\ 0, & k \neq \ell \end{cases} \quad (0 \leq k, \ell \leq m) \\ \frac{2}{2m+1} \sum_{j=0}^{2m} \sin kx_j \sin \ell x_j &= \begin{cases} 1, & k = \ell \\ 0, & k \neq \ell \end{cases} \quad (1 \leq k, \ell \leq m) \\ \frac{2}{2m+1} \sum_{j=0}^{2m} \cos kx_j \sin \ell x_j &= 0 \quad (0 \leq k, \ell \leq m) \end{aligned}$$

Proof: Interpolation of the basic functions $1, \cos x, \sin x, \dots, \cos mx, \sin mx$. $\square$

II. Even number of uniformly spaced nodes

Nodes $x_j = \frac{j\pi}{m}, \quad j = 0, 1, \dots, 2m-1$

Data $y_j \in \mathbb{R}, \quad j = 0, 1, \dots, 2m-1$

Space $\{1, \cos x, \sin x, \dots, \cos mx\}, \quad \dim = 2m$

($\sin mx$ is omitted, since $\sin mx_j = 0$ for $j = 0, 1, \dots, 2m-1$)

Haar system in $\{x_0, \dots, x_{2m-1}\}$

(solvability of the interpolation problem for these special nodes)

The *basic Lagrange functions* $t_j(x) = \frac{1}{2m} \frac{\sin m(x - x_j)}{\tan \frac{x - x_j}{2}}, \quad j = 0, 1, \dots, 2m-1$

Compared to the Dirichlet kernel: $\frac{\sin mu}{\tan \frac{1}{2}u} = D(u) - \cos mu$.

According to the Lagrange representation the normal representation follows.

Theorem 2.4 *The trigonometric interpolation polynomial with respect to the $2m$ uniformly spaced nodes is given by*

$$\begin{aligned} T(x) &= \frac{\alpha_0}{2} + \sum_{\nu=1}^m \{ \alpha_\nu \cos \nu x + \beta_\nu \sin \nu x \} \quad \left(\Sigma' : \frac{\alpha_m}{2} \right), \\ \left. \begin{matrix} \alpha_\nu \\ \beta_\nu \end{matrix} \right\} &= \frac{1}{m} \sum_{j=0}^{2m-1} y_j \begin{Bmatrix} \cos \nu x_j \\ \sin \nu x_j \end{Bmatrix}, \quad \nu = 0, 1, \dots, m \quad (\beta_0 = \beta_m = 0). \end{aligned}$$

Exponential representation (more compact, data $y_j \in \mathbb{C}$):

Space $\{e^{-imx}, \dots, 1, \dots, e^{imx}\}$ Haar system in $[0, 2\pi)$, $\dim = 2m + 1$

$$T(x) = \frac{\alpha_0}{2} + \sum_{\nu=1}^m \{\alpha_\nu \cos \nu x + \beta_\nu \sin \nu x\} = \sum_{\nu=-m}^m a_\nu e^{i\nu x}$$

$$\boxed{a_{\pm\nu} = \frac{1}{2}(\alpha_\nu \mp i\beta_\nu)} \quad (e^{ix} = \cos x + i \sin x)$$

Interpolation with $2m + 1$ nodes:

$$a_{\pm\nu} = \frac{1}{2m+1} \sum_{j=0}^{2m} y_j e^{\mp i\nu x_j}$$

$$z = e^{ix} : \quad T(x) = z^{-m} \sum_{\nu=0}^{2m} a_{\nu-m} z^\nu =: z^{-m} S(z)$$

$S(z)$ complex algebraic polynomial of degree $\leq 2m$.

Idea: It is easier to determine $S(z)$ than $T(x)$, i.e.,
the coefficients $a_{\pm\nu}$ are simpler than α_ν and β_ν .

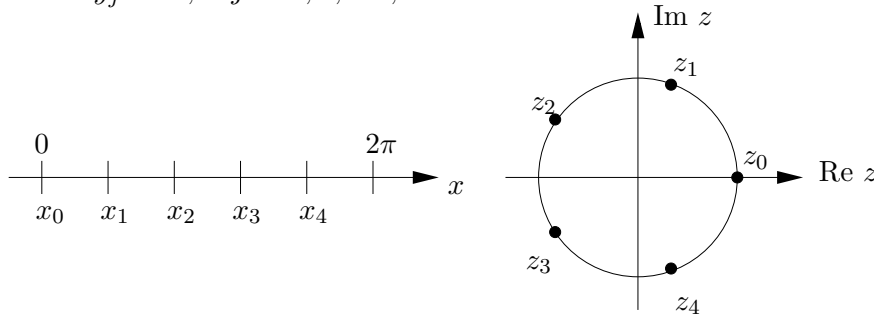
New interpolation problem

Space $\{1, e^{ix}, \dots, e^{i(n-1)x}\}$: Haar system in $[0, 2\pi)$, $\dim = n$

$$\left. \begin{array}{ll} \text{Complex trigonometric polynomial} & R(x) = \sum_{\nu=0}^{n-1} b_\nu e^{i\nu x} \\ \text{Complex algebraic polynomial} & S(z) = \sum_{\nu=0}^{n-1} b_\nu z^\nu \end{array} \right\} z = e^{ix}$$

Nodes $x_j = \frac{2j\pi}{n}$ respectively $z_j = e^{ix_j}$, $j = 0, 1, \dots, n-1$

Data $\hat{y}_j \in K$, $j = 0, 1, \dots, n-1$



Wanted: $R(x)$ respectively $S(z)$ satisfying $R(x_j) = S(z_j) = \hat{y}_j$, $j = 0, \dots, n-1$

$S(z)$: interpolation polynomial with respect to uniformly spaced nodes on the unit circle!

Relations of the nodes: $z_j = e^{\frac{2ij\pi}{n}}$, $z_j^\nu = e^{\frac{2ij\nu\pi}{n}}$

$$z_j^n = 1, z_j^0 = z_0^\nu = 1, z_j^\nu = z_\nu^j, \bar{z}_j = z_{-j}, z_j^{-k} = z_j^{n-k}$$

$$0 = z_j^n - 1 = \underbrace{(z_j - 1)}_{\neq 0, j \neq 0} \sum_{\substack{\nu=0 \\ =0, j \neq 0}}^{n-1} z_j^\nu, \sum_{\nu=0}^{n-1} z_j^\nu = \begin{cases} 0, & j \neq 0 \\ n, & j = 0 \end{cases}$$

Orthogonality $\frac{1}{n} \sum_{\nu=0}^{n-1} z_\nu^j z_\nu^{-k} = \begin{cases} 0, & j \neq k \\ 1, & j = k \end{cases}$

Basic functions $S_k(z) := \frac{1}{n} \sum_{\nu=0}^{n-1} (z_{-k} z)^\nu$ with $S_k(z_j) = \delta_{kj}$

Lagrange representation $S(z) = \sum_{k=0}^{n-1} \hat{y}_k S_k(z)$

Theorem 2.5 *The complex trigonometric interpolation polynomial $R(x)$ respectively the complex algebraic interpolation polynomial $S(z)$ with respect to the nodes $x_j = \frac{2j\pi}{n}$ respectively $z_j = e^{ix_j} \in K$ ($j = 0, 1, \dots, n-1$) and the data $\hat{y}_j$ ($j = 0, 1, \dots, n-1$) has the coefficients*

$$b_\nu = \frac{1}{n} \sum_{k=0}^{n-1} \hat{y}_k z_\nu^{-k}, \quad \nu = 0, 1, \dots, n-1.$$

Proof: The Lagrange representation implies the statement

$$S(z) = \sum_{k=0}^{n-1} \hat{y}_k S_k(z) = \sum_{\nu=0}^{n-1} \underbrace{\frac{1}{n} \sum_{k=0}^{n-1} \hat{y}_k z_{-k}^\nu}_{=b_\nu} z^\nu = \sum_{\nu=0}^{n-1} b_\nu z^\nu. \quad \square$$

Relation between the coefficients b_ν and α_ν, β_ν :

Assume that $n = 2m = 2^p$ (even number of uniformly spaced nodes)

$$\left. \begin{matrix} \alpha_\nu \\ \beta_\nu \end{matrix} \right\} = \frac{1}{m} \sum_{j=0}^{n-1} y_j \begin{cases} \cos \nu x_j \\ \sin \nu x_j \end{cases}, \quad \nu = 0, 1, \dots, m \quad (\beta_m = 0) \quad (\text{real})$$

$$b_\nu = \frac{1}{n} \sum_{k=0}^{n-1} \hat{y}_k z_\nu^{-k}, \quad \nu = 0, 1, \dots, n-1 \quad (\text{complex})$$

$$\hat{y}_k \rightarrow y_k$$

Backward transformation formulas

Consider the easier case of $2m + 1$ nodes (see above)

$$T(x) = z^{-m} S(z), \quad S(z) = \sum_{\nu=0}^{2m} a_{\nu-m} z^{\nu}$$

$$a_{\pm\nu} = \frac{1}{2}(\alpha_{\nu} \mp i\beta_{\nu})$$

$$a_{\nu} = \frac{1}{2}(\alpha_{\nu} - i\beta_{\nu}), \quad a_{-\nu} = \frac{1}{2}(\alpha_{\nu} + i\beta_{\nu})$$

hence we have

$$\alpha_{\nu} = a_{\nu} + a_{-\nu}, \quad \beta_{\nu} = i(a_{\nu} - a_{-\nu})$$

further

$$y_j = T(x_j) = z_j^{-m} S(z_j) = z_j^{-m} \hat{y}_j \rightarrow \hat{y}_j = y_j \cdot z_j^m$$

Summary: Trigonometric interpolation: $n = 2m$

$$\left. \begin{array}{l} \text{Nodes} \quad x_j = \frac{2j\pi}{n} \\ \text{Data} \quad y_j \in K \end{array} \right\} j = 0, 1, \dots, n-1$$

Trigonometric interpolation polynomial $T(x)$ with coefficients

$$\left. \begin{array}{l} \alpha_{\nu} \\ \beta_{\nu} \end{array} \right\} = \frac{1}{m} \sum_{j=0}^{n-1} y_j \left\{ \begin{array}{l} \cos \nu x_j \\ \sin \nu x_j \end{array} \right., \quad \nu = 0, 1, \dots, m$$

Computation by the coefficients of the complex interpolation polynomial:

$$\left. \begin{array}{l} \text{Nodes} \quad z_j = e^{i2j\pi/n} \\ \text{Data} \quad \hat{y}_j \in K \end{array} \right\} j = 0, 1, \dots, n-1$$

$$\text{Complex interpolation polynomial } S(z) = \sum_{\nu=0}^{n-1} b_{\nu} z^{\nu}, \quad b_{\nu} = \frac{1}{n} \sum \hat{y}_k z_{\nu}^{-k}$$

Assume an even number of nodes $n = 2m$

Choose the data $\hat{y}_k \rightarrow y_k$

Compute the coefficients b_{ν} with the FFT method (see below).

Then the coefficients $\alpha_{\nu}, \beta_{\nu}$ will follow for the real trigonometric interpolation polynomial $T(x)$ ($\rightarrow$ backward transformation formulas).

3 Fourier Transform

Space $C_{2\pi}$, $\langle f, g \rangle := \frac{1}{\pi} \int_0^{2\pi} f(x)g(x)dx$, L_2 -norm $\|f\|_2 := \langle f, f \rangle^{1/2}$

Subspace $\Pi_m : \left\{ \frac{1}{\sqrt{2}}, \cos x, \sin x, \dots, \cos mx, \sin mx \right\}$ **ONS**

Fourier series $f \sim \frac{\hat{\alpha}_0}{2} + \sum_{\nu=1}^{\infty} \{ \hat{\alpha}_{\nu} \cos \nu x + \hat{\beta}_{\nu} \sin \nu x \}$

Fourier coefficients $\left\{ \begin{smallmatrix} \hat{\alpha}_{\nu} \\ \hat{\beta}_{\nu} \end{smallmatrix} \right\} = \langle f, \begin{smallmatrix} \cos \nu x \\ \sin \nu x \end{smallmatrix} \rangle = \frac{1}{\pi} \int_0^{2\pi} f(x) \begin{Bmatrix} \cos \nu x \\ \sin \nu x \end{Bmatrix} dx$

Fourier polynomial S_m : the m^{th} partial sum of the series

Complex-valued functions

Scalar product $\langle f, g \rangle := \frac{1}{2\pi} \int_0^{2\pi} f(x)\bar{g}(x)dx$

ONS (orthogonal normal system) $\{1, e^{ix}, \dots, e^{i(n-1)x}\}$

Fourier series $\sum_{\nu=0}^{\infty} \hat{b}_{\nu} e^{i\nu x}$

Fourier coefficients $\hat{b}_{\nu} = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-i\nu x} dx$

L_2 -approximation

Space $C[a, b]$, $\|\cdot\|_2$

Finite dimensional subspace U with an

orthogonal normal basis $\{\varphi_0, \varphi_1, \dots, \varphi_{n-1}\}$

The *best approximation* $g^* \in U$ of $f \in C$, i.e.,

$$\|f - g^*\|_2 \leq \|f - g\|_2 \quad \text{for all } g \in U$$

Existence, uniqueness, construction of p^* ?

The *best approximation* $g^* = \sum_{\nu=0}^{n-1} \hat{c}_{\nu} \varphi_{\nu}$, $\hat{c}_{\nu} := \langle f, \varphi_{\nu} \rangle$ the *Fourier coefficients*

Theorem 3.1 (Minimal property of the Fourier coefficients)

The *Fourier polynomial* g^* is the best L_2 -approximation of $f \in C[a, b]$.

Proof: $g \in U$, $g = \sum_{\nu=0}^{n-1} c_\nu \varphi_\nu$

$$\begin{aligned}
\|f - g\|_2^2 &= \langle f - g, f - g \rangle = \langle f, f \rangle - 2\langle f, g \rangle + \langle g, g \rangle \\
&= \|f\|_2^2 - 2 \sum_j c_j \langle f, \varphi_j \rangle + \sum_j \sum_k c_j c_k \underbrace{\langle \varphi_j, \varphi_k \rangle}_{=\delta_{jk}} \\
&= \|f\|^2 + \sum (c_j^2 - 2c_j \hat{c}_j + \hat{c}_j^2) - \sum \hat{c}_j^2 \\
&= \underbrace{\|f\|^2 - \sum \hat{c}_j^2}_{=0} + \sum (c_j - \hat{c}_j)^2 \\
&= \|f - g^*\|_2^2 \quad (\text{insert } g = g^*). \quad \square
\end{aligned}$$

Corollary 3.2 S_m is the best L_2 -approximation of $f \in C_{2\pi}$, i.e.,

$$\|f - s_m\|_2 \leq \|f - t_m\|_2 \quad \text{for all } t_m \in \Pi_m.$$

Computation of the Fourier coefficients

$$\left. \begin{matrix} \hat{\alpha}_\nu \\ \hat{\beta}_\nu \end{matrix} \right\} = \frac{1}{\pi} \int_0^{2\pi} f(x) \begin{cases} \cos \nu x \\ \sin \nu x \end{cases} dx \quad \text{resp.} \quad \hat{b}_\nu = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-i\nu x} dx$$

Quadrature formula: Periodic integrand, hence we choose uniformly spaced nodes

Trapezoidal rule: Nodes $x_j = \frac{2j\pi}{n}, j = 0, 1, \dots, n$, $h = \frac{2\pi}{n}$

$$\begin{aligned}
\frac{h}{2} (g(x_0) + g(x_1)) + \dots + \frac{h}{2} (g(x_{n-1}) + g(x_n)) &= \frac{2\pi}{n} \sum_{j=0}^{n-1} g(x_j), \\
\left. \begin{matrix} \hat{\alpha}_\nu \\ \hat{\beta}_\nu \end{matrix} \right\} &= \frac{2}{n} \sum_{j=0}^{n-1} f(x_j) \begin{cases} \cos \nu x_j \\ \sin \nu x_j \end{cases} + \text{error} = \begin{cases} \alpha_\nu \\ \beta_\nu \end{cases} + \text{error} \\
\hat{b}_\nu &= \frac{1}{2} \sum_{j=0}^{n-1} f(x_j) e^{-i\nu x_j} + \text{error} = b_\nu + \text{error}
\end{aligned}$$

Discrete Fourier coefficients $\alpha_\nu, \beta_\nu, b_\nu$

Trapezoidal Σ = mid-point Σ = arithmetic Σ

Remark: The trapezoidal rule is very convenient for periodic functions (because of high order of approximation) (Euler–MacLaurin §6).

Result 3.3 The Fourier coefficients $\hat{\alpha}_\nu, \hat{\beta}_\nu$ and $\hat{b}_\nu$ (see above) are very good approximated by the discrete Fourier coefficients α_ν, β_ν and b_ν .

Remark: The discrete Fourier coefficients are the coefficients of the trigonometric respectively complex algebraic interpolation polynomial $T(x)$ (Theorem 2.2, 2.4) resp. $S(z)$ (Theorem 2.5).

The discrete Fourier evaluation

Space $C_{2\pi}$: nodes $x_j = \frac{2j\pi}{n}$, $j = 0, 1, \dots, n-1$ ($n = 2m+1$ or $n = 2m$)

$$\langle f, g \rangle_n := \frac{2}{n} \sum_{j=0}^{n-1} f(x_j)g(x_j) \quad (\text{arithmetic } \Sigma \text{ of } \langle f, g \rangle)$$

discrete analogon to $\langle f, g \rangle$, not a scalar product in $C_{2\pi}$,

because of the missing definiteness!

Seminorm $\|f\|_{2,n} := \langle f, f \rangle_n^{1/2}$.

Space Π_k ($0 \leq k \leq m$, where $m := \left\lfloor \frac{n}{2} \right\rfloor$): $\langle f, g \rangle_n$ is a semi scalar product,

hence the **discrete Fourier polynomial** $T_k(x)$ with α_ν and β_ν (Th. 3.1).

Analogously: *ONS* $\{1, e^{ix}, \dots, e^{i(n-1)x}\}$

$$\langle f, g \rangle_n := \frac{1}{n} \sum_{j=0}^{n-1} f(x_j) \overline{g(x_j)}$$

Least square problem

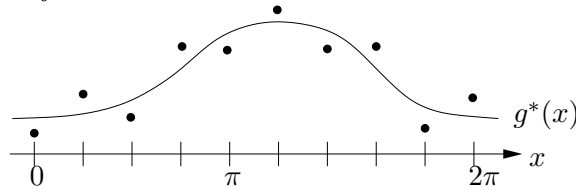
Given: Nodes $x_j = \frac{2j\pi}{n}$, $j = 0, 1, \dots, n-1$ ($n = 2m+1$ or $n = 2m$)

Data y_j ($y_j := f(x_j)$, $f \in C_{2\pi}$), $j = 0, 1, \dots, n-1$

Wanted: $g^* \in \Pi_k$ ($k \leq m$)

by the *principle of the smallest squares of errors*

$$\frac{2}{n} \sum_{j=0}^{n-1} (y_j - g^*(x_j))^2 = \text{minimum}$$



Approximation problem: $f \in C_{2\pi}$ approximated by $g^* \in \Pi_k$ ($0 \leq k \leq m$)

First step: approximate f by the interpolation polynomial $T \in \Pi_m$
i.e., $T(x_j) = f(x_j) =: y_j$ for $j = 0, 1, \dots, n-1$

Second step: approximate T by $g \in \Pi_k$ best possible; $\|\cdot\|_{2,n}$, semi-norm in Π_k ,
then the *best approximation* $g^* \in \Pi_k$ satisfies
 $\|T - g^*\|_{2,n} \leq \|T - g\|_{2,n}$ for all $g \in \Pi_k$,
solution $g^* = T_k =$ discrete Fourier polynomial (see Theorem 3.1).

Theorem 3.4 *The discrete Fourier polynomial T_k ($k < m$) equal to the k – th partial polynomial of the interpolation polynomial T_m solves the given least squares problem.*

Direct proof (without approximation theory):

$$F(\alpha_0, \dots, \alpha_k, \beta_1, \dots, \beta_k) := \sum_{j=0}^{n-1} \left\{ \frac{1}{2} \alpha_0 + \sum_{\nu=1}^k (\alpha_\nu \cos \nu x_j + \beta_\nu \sin \nu x_j) - f(x_j) \right\}^2$$

Necessary conditions:

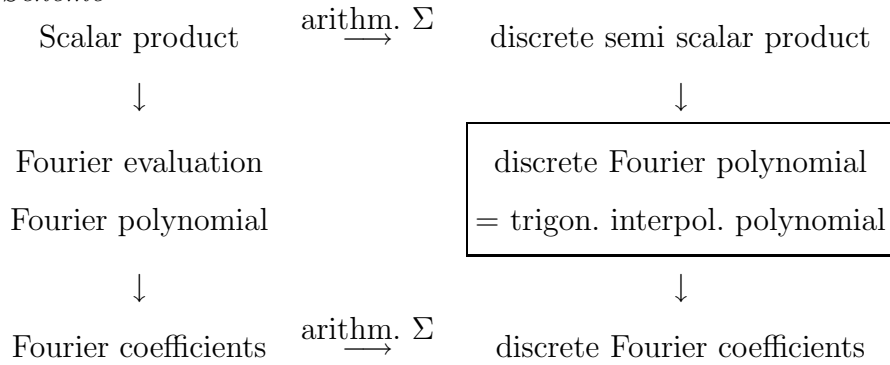
the discrete Fourier coefficients α_ν can be computed by $\frac{\partial F}{\partial \alpha_\nu} = 0, \nu = 0, 1, \dots, k$

the discrete Fourier coefficients β_ν can be computed by $\frac{\partial F}{\partial \beta_\nu} = 0, \nu = 1, \dots, k$

Remark: This is a very elegant solution!

Compare the least squares problem for algebraic polynomials $\in \mathcal{P}_k$, then the normal equations $A^T A x = A^T b$ (linear system) has to be solved.

Scheme



SUPPLEMENTARY EXAMPLES – NO. 2

FOURIER SERIES (Schwarz*)

Example 1

“Roof” function $f(x)$, 2π –periodic, $f(x) = |x|$ in the basic interval $[-\pi, \pi]$.

a) *Wanted:* Fourier series of $f(x)$: $f(x)$ even, i.e., all $\hat{\beta}_k = 0$;

$$\begin{aligned} \hat{\alpha}_0 &= \frac{1}{\pi} \int_{-\pi}^{\pi} |x| dx = \frac{2}{\pi} \int_0^{\pi} x dx = \pi, \\ \hat{\alpha}_k &= \frac{2}{\pi} \int_0^{\pi} x \cos kx dx = \frac{2}{\pi} \left\{ \frac{1}{k} x \sin kx \right\} \Big|_0^{\pi} - \frac{1}{k} \int_0^{\pi} \sin kx dx \\ &= \frac{2}{\pi k^2} \cos kx \Big|_0^{\pi} = \frac{2}{\pi k^2} \{(-1)^k - 1\}, \quad k > 0. \end{aligned}$$

The Fourier series reads

$$f(x) \sim \frac{1}{2}\pi - \frac{4}{\pi} \left\{ \frac{\cos x}{1^2} + \frac{\cos 3x}{3^2} + \frac{\cos 5x}{5^2} + \dots \right\},$$

and hence it follows

$$\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \dots = \frac{\pi^2}{8}.$$

Error estimation:

$$|f(x) - S_{25}(x)| \leq \frac{4}{\pi} \sum_{\nu=13}^{\infty} \frac{1}{(2\nu+1)^2} \leq 0.025.$$

b) *Wanted:* Trigon. interpolation polynomial (discrete Fourier polynomial)

$$T(x) = \frac{\alpha_0}{2} + \sum_{\nu=0}^3 (\alpha_{\nu} \cos \nu x + \beta_{\nu} \sin \nu x)$$

with respect to the nodes $x_j = \frac{j\pi}{4}$ and data $y_j = f(x_j)$, $j = 0, 1, \dots, 7$.

The discrete Fourier coefficients read

$$\alpha_0 = \pi, \alpha_1 = -1.34, \alpha_2 = 0, \alpha_3 = -0.23, \alpha_4 = 0, \beta_1 = \beta_2 = \beta_3 = 0.$$

Compare with the classical Fourier coefficients

$$\hat{\alpha}_0 = \pi, \hat{\alpha}_1 = -1.27, \hat{\alpha}_2 = 0, \hat{\alpha}_3 = -0.14, \hat{\alpha}_4 = 0, \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0.$$

Notice that $T(0) = 0$ (interpolation) and $S_{25}(0) = 0.0244$ (trapezoidal sum) hold.

Example 2

The function $x \rightarrow x^2$, $0 \leq x < 2\pi$, will be continued periodically by $f(x)$.

Wanted: Fourier series of $f(x)$: Integration by parts

$$\begin{aligned} \hat{\alpha}_0 &= \frac{1}{\pi} \int_0^{2\pi} x^2 dx = \frac{8\pi^2}{3}, \\ \hat{\alpha}_k &= \frac{1}{\pi} \int_0^{2\pi} x^2 \cos kx dx = \frac{4}{k^2}, \quad k = 1, 2, \dots, \\ \hat{\beta}_k &= \frac{1}{\pi} \int_0^{2\pi} x^2 \sin kx dx = -\frac{4\pi}{k}, \quad k = 1, 2, \dots \end{aligned}$$

The Fourier series reads

$$\frac{4\pi^2}{3} + \sum_{k=1}^{\infty} \left\{ \frac{4}{k^2} \cos kx - \frac{4\pi}{k} \sin kx \right\}.$$

Computation of the discrete Fourier coefficients

$$\left. \begin{matrix} \alpha_\nu \\ \beta_\nu \end{matrix} \right\} = \frac{2}{n} \sum_{j=0}^{n-1} y_j \begin{cases} \cos \nu x_j \\ \sin \nu x_j \end{cases} \quad \nu = 0, 1, \dots, \left[\frac{n}{2} \right]$$

In general (Theorem 2.5):

$b_k = \frac{1}{n} \sum_{j=0}^{n-1} \hat{y}_j z_j^{-k}$	Fourier analysis $(k = 0, 1, \dots, n-1)$
$\hat{y}_k = \sum_{j=0}^{n-1} b_j z_j^k$	Fourier synthesis

Fast Fourier Transform (FFT)

Computational costs for the n complex algebraic sums: $n = 2^p$ ($p \in \mathbb{N}$)

Horner $\sim 2n^2$ arithmetic operations ($n \cdot 2^{p+1}$)

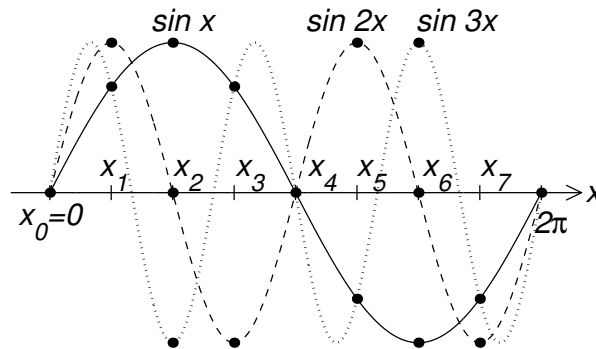
FFT $\sim n \log_2 n$ arithmetic operations ($n \cdot p$)

FFT: Algorithm of **Cooley & Tukey** 1965
(see Stoer I; Schwarz)

The **basic idea** is **convolution** (Runge 1903):

for example, consider the real coefficients α_ν and β_ν for $n = 2^p$

$$n = 2^3 : \beta_\nu = \frac{1}{4} \sum_{j=0}^7 y_j \sin \nu x_j, \quad \nu = 1, 2, 3 \quad (\beta_4 = 0), \quad x_j = \frac{j\pi}{4}, \quad j = 0, 1, \dots, 7$$



$\beta_1 : \sin x_j$ 3 distinct absolute values

$\beta_2 : \sin 2x_j$ 2 distinct absolute values

$\beta_3 : \sin 3x_j$ 3 distinct absolute values

Totally 3 distinct values: $\sin x_0, \sin x_1, \sin x_2$

Convolution: Gathering together all the y_j , which have to be multiplied with the same absolute value of the sine-function S_k .

FFT applied to $b_k = \frac{1}{n} \sum_{j=0}^{n-1} \hat{y}_j z_j^{-k}$, $\hat{y}_k = \sum_{j=0}^{n-1} b_j z_j^k$, $k = 0, 1, \dots, n-1$, $n = 2^p$:

The principle of convolution leads to a more efficient algorithm on the computer!

Linear transformation $\mathbb{C}^n \rightarrow \mathbb{C}^n$ with symmetric matrix T

$$\boxed{y = Tb} \quad y := \begin{pmatrix} \hat{y}_0 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_{n-1} \end{pmatrix}, \quad b := \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{n-1} \end{pmatrix}, \quad T := \begin{pmatrix} z_0^0 & \dots & z_{n-1}^0 \\ z_0^1 & \dots & z_{n-1}^1 \\ \vdots & & \vdots \\ z_0^{n-1} & \dots & z_{n-1}^{n-1} \end{pmatrix}$$

$$\boxed{b = \frac{1}{n} P T y} \quad P := \left(\begin{array}{c|c} 1 & \\ \hline & 1 \\ \hline & 1 \end{array} \right) \quad \text{permutation matrix (orthogonal)}$$

Using the relations of the powers z_k^j (see §2): $z_k^j = e^{2ijk\pi/n}$

$$z_j^{-k} = z_j^{n-k}, \quad z_j^{-k} = z_k^{-j}, \quad \sum_{\nu=0}^{n-1} z_j^\nu = \begin{cases} 0, & j \neq 0 \\ n, & j = 0 \end{cases}$$

Convolution: Factorization of the matrix T into a product of sparse matrices

Result 3.5 (see Stoer I)

$$n = 2^p : T = (QSP)(D_{p-1}SP) \cdots (D_1SP) =: T_p T_{p-1} \dots T_1$$

$$S = \begin{pmatrix} \boxed{\begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix}} & & & 0 \\ & \boxed{\begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix}} & & \\ & & \ddots & \\ 0 & & & \boxed{\begin{matrix} 1 & 1 \\ 1 & -1 \end{matrix}} \end{pmatrix},$$

Q and P permutation matrices, $D_1, \dots, D_{p-1}$ diagonal matrices;

matrices T_j : In each row and column there are exactly two elements $\neq 0$;

more precise: $D_\ell = \text{diag}(\delta_0, \dots, \delta_{n-1})$, $\delta_r = \exp \frac{2ir_0 r_\ell^* \pi}{2^{p-\ell+1}}$

$$\begin{aligned} \text{for} \quad & r = r_0 + r_1 \cdot 2 + \dots + r_{\ell-1} 2^{\ell-1} + r_\ell^* 2^\ell = 0, 1, \dots, n-1 \\ \text{with} \quad & r_0, \dots, r_{\ell-1} \in \{0, 1\}, r_\ell^* \in \{0, 1, \dots, 2^{p-1} - 1\}. \end{aligned}$$

P defined by $\hat{x} = Px$ with

$$\hat{x}_{k+j \cdot 2} = x_{j+k \cdot 2^{p-1}} \quad \text{for } k = 0, 1 \quad \text{and } j = 0, 1, \dots, 2^{p-1} - 1;$$

Q defined by $\hat{x} = Qx$ with

$$\hat{x}_{j_0+j_1 \cdot 2+\dots+j_{p-1} \cdot 2^{p-1}} = x_{j_{p-1}+j_{p-2} \cdot 2+\dots+j_0 \cdot 2^{p-1}} \quad \text{for } j_0, \dots, j_{p-1} = 0, 1.$$

Without proof $\square$

FFT (fast computation): $y = T_p(T_{p-1} \dots (T_2(T_1 b)) \dots)$

$$\begin{array}{ccccccc} b =: & v_1 & \rightarrow & T_1 v_1 =: v_2 & \rightarrow & T_2 v_2 =: v_3 & \rightarrow \dots \rightarrow T_p v_p = y \\ & \uparrow & & \uparrow & & \uparrow & \uparrow \\ & & & \text{convolution} & & & \end{array}$$

Realization: $T_j v_j = D_j S P v_j = D_j (S(P v_j))$

Inversion: $b = T^{-1} y$ with $T^{-1} = T_1^{-1} \dots T_p^{-1}$, $T_j^{-1} = \frac{1}{2} P S D_j^{-1}$

SUPPLEMENTARY EXAMPLES – NO. 3
FAST FOURIER TRANSFORM (FFT)

Example 1: $n = 4, p = 2$: $y = Tb, b = T^{-1}y$

Nodes $z_j^k = \exp(\frac{1}{2}ijk\pi) = \cos \frac{1}{2}jk\pi + i \sin \frac{1}{2}jk\pi \quad (j = 0, 1, 2, 3)$

Matrix
$$T = \begin{pmatrix} z_0^0 & z_1^0 & z_2^0 & z_3^0 \\ z_0^1 & z_1^1 & z_2^1 & z_3^1 \\ z_0^2 & z_1^2 & z_2^2 & z_3^2 \\ z_0^3 & z_1^3 & z_2^3 & z_3^3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}$$

Factorization $T = (QSP)(D_1SP) =: T_2T_1, \quad T^{-1} = T_1^{-1}T_2^{-1}$

$$S = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad D_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & i \end{pmatrix}, \quad P = Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$SP = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & -1 \end{pmatrix}, \quad T_1 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & i & 0 & -i \end{pmatrix}, \quad T_2 = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{pmatrix}$$

$$S^{-1} = \frac{1}{2}S, \quad D_1^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -i \end{pmatrix}, \quad PS = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix},$$

$$T_1^{-1} = \frac{1}{2}PSD_1^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -i \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & i \end{pmatrix}, \quad T_2^{-1} = \frac{1}{2}PSQ = \frac{1}{2}T_2$$

Convolutions

$$\begin{array}{llll}
\mathbf{Y} = \mathbf{T}\mathbf{b} : & b & \longrightarrow & v = T_1 b \qquad \longrightarrow \quad y = T_2 v \\
\\
& b_0 & & v_0 = b_0 + b_2 \qquad \qquad y_0 = v_0 + v_2 = b_0 + b_1 + b_2 + b_3 \\
& b_1 & & v_1 = b_0 - b_2 \qquad \qquad y_1 = v_1 + v_3 = b_0 + ib_1 - b_2 + ib_3 \\
& b_2 & & v_2 = b_1 + b_3 \qquad \qquad y_2 = v_0 - v_2 = b_0 - b_1 + b_2 - b_3 \\
& b_3 & & v_3 = i(b_1 - b_3) \qquad \qquad y_3 = v_1 - v_3 = b_0 - ib_1 - b_2 + ib_3
\end{array}$$

$$\begin{array}{llll}
\mathbf{b} = \mathbf{T}^{-1}\mathbf{y} : & y & \longrightarrow & w = T_2^{-1}y \qquad \longrightarrow \quad b = T_1^{-1}w \\
\\
& y_0 & & w_0 = \frac{1}{2}(y_0 + y_2) \qquad \qquad b_0 = \frac{1}{2}(w_0 + w_1) = \frac{1}{4}(y_0 + y_1 + y_2 + y_3) \\
& y_1 & & w_1 = \frac{1}{2}(y_1 + y_3) \qquad \qquad b_1 = \frac{1}{2}(w_2 - iw_3) = \frac{1}{4}(y_0 - iy_1 - y_2 + iy_3) \\
& y_2 & & w_2 = \frac{1}{2}(y_0 - y_2) \qquad \qquad b_2 = \frac{1}{2}(w_0 - w_1) = \frac{1}{4}(y_0 - y_1 + y_2 - y_3) \\
& y_3 & & w_3 = \frac{1}{2}(y_1 - y_3) \qquad \qquad b_3 = \frac{1}{2}(w_2 + iw_3) = \frac{1}{4}(y_0 + iy_1 - y_2 - iy_3)
\end{array}$$

Costs: 9 operations ($np = 8$) comp. to 28 operations ($n2^{p+1} = 32$) for Tb .

Backward transformation to the α_ν and β_ν is possible.

Example 2: Trigonometric interpolation of $f(x) = \cos 4x$ with respect to the nodes $x_j = \frac{j\pi}{4}, j = 0, 1, \dots, 7$.

Wanted: The discrete Fourier coefficients in the representation

$$T(x) = \frac{\alpha_0}{2} + \sum_{\nu=1}^3 (\alpha_\nu \cos \nu x + \beta_\nu \sin \nu x) + \frac{\alpha_4}{2} \cos 4x.$$

The function values are $f(x_j) = (-1)^j, j = 0, 1, \dots, 4$.

Choose $y_k := f(x_{2k}) + if(x_{2k+1}) = 1 - i$ for $k = 0, 1, 2, 3$, then the complex polynomial $S(z)$ has the coefficients

$$b_\nu = \frac{1}{4} \sum_{k=0}^3 y_k z_\nu^{-k}, \quad \nu = 0, 1, 2, 3.$$

FFT yields

$$b_0 = 1 - i, b_1 = b_2 = b_3 = 0, \quad \text{i.e.,} \quad S(z) = b_0 = y_k.$$

Backward transformation yields the discrete Fourier coefficients

$$\alpha_0 = \alpha_1 = \alpha_2 = \alpha_3 = \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{and} \quad \alpha_4 = 2, \quad \text{i.e.,} \quad T(x) = \cos 4x.$$

4 Cubic Spline Interpolation

Given: Nodes $a \leq x_0 < \cdots < x_n \leq b$
 Data $y_0, y_1, \dots, y_n$ ($y_j := f(x_j), f \in C[a, b]$)

Wanted: The most simple and smooth function g in $[a, b]$ satisfying
 $g(x_j) = y_j, j = 0, 1, \dots, n$ with a small error $f - g$.

$g \in \mathcal{P}_n$: Unique solvability, simple computation, error estimates
 but the error may become large using many nodes and high
 polynomial degree $\rightarrow$ divergence

$g \in \Pi_n$: Unique solvability, fast computation FFT $\rightarrow$ periodic function

g rational function: nonlinear problem, no Haar system
 singular points/poles $\rightarrow$ “unreachable” points

g spline function: Polynomial on subintervals, smooth transitions

Grid Δ : $a = x_0 < x_1 < \cdots < x_n = b$ called the *spline nodes*

Definition

$s_\Delta : [a, b] \rightarrow \mathbb{R}$ is called **spline of degree ℓ** ($\ell \geq 0$) with respect to
 the grid Δ , if

$$s_\Delta \Big|_{[x_{j-1}, x_j]} \in \mathcal{P}_\ell \text{ for } j = 1, \dots, n \text{ and } s_\Delta \in C^{\ell-1}[a, b] \text{ hold.}$$

Spline of degree 0: step function (piecewise continuous)

Spline of degree 1: polygon (continuous)

Spline of degree 2: piecewise parabola (continuously differentiable)

Spline of degree 3: cubic spline

Cubic spline: The total curvature is minimal, i.e., $\|s''\|_2^2 = \text{minimal}$

$S_3(\Delta)$: Space of the cubic splines with respect to the grid Δ

$$\boxed{\dim S_3(\Delta) = n + 3}$$

Representation by truncated power function (see Numer. Math. I):

$$s(x) = \sum_{j=0}^2 b_j (x - x_0)_+^j + \sum_{k=0}^{n-1} c_k (x - x_k)_+^3$$

Representation interval by interval

$$s(x) = \begin{cases} p_0(x) & \text{in } [x_0, x_1] \\ p_1(x) & \text{in } [x_1, x_2] \\ \vdots & \\ p_{n-1}(x) & \text{in } [x_{n-1}, x_n] \end{cases} \quad p_0, p_1, \dots, p_{n-1} \in \mathcal{P}_3.$$

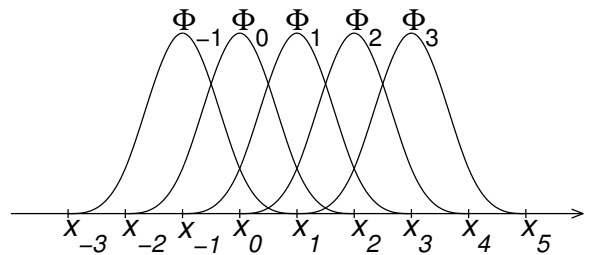
Representation by B-splines $s(x) = \sum_{\nu=-1}^{n+1} c_\nu \phi_\nu(x)$

$$\phi_\nu(t) := v_t[x_{\nu-2}, \dots, x_{\nu+2}], \quad \nu = -1, \dots, n+1$$

divided differences for $v_t(x) = \frac{1}{3!} (x-t)_+^3$ (see Corollary 1.3).

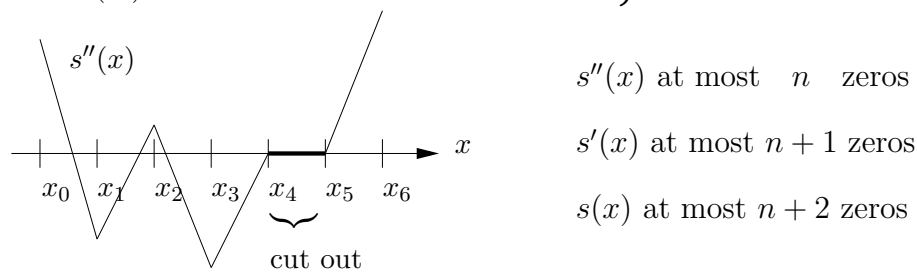
Additional nodes:

$$x_{-3} < x_{-2} < x_{-1} < x_0 < \dots < x_n < x_{n+1} < x_{n+2} < x_{n+3}$$



Interpolation in the space $S_3(\Delta)$

$$\left. \begin{array}{l} \dim = n+3, \text{ i.e., } n+3 \text{ conditions} \\ s \in S_3(\Delta) \text{ has at most } n+2 \text{ essential zeros} \end{array} \right\} \text{ Haar space}$$



Interpolation nodes = spline nodes (i.e., $n+1$ nodes)
(suitable for the construction) and additionally 2 further conditions!

Boundary conditions:

- a) Hermite condition $s'(x_0) = y_0^{(1)} (= f'(x_0)), s'(x_n) = y_n^{(1)} (= f'(x_n))$
- b) Curvature condition $s''(x_0) = s''(x_n) = 0$
(i.e., the spline runs straightly outside the interval $[a, b]$)

Definition $s_f \in S_3(\Delta)$ with $s_f(x_j) = y_j$ ($= f(x_j)$), $j = 0, 1, \dots, n$,

is called **interpolating spline** of f with respect to Δ .

Theorem 4.1 *The interpolating cubic spline s_f with boundary condition (a) or (b) has the following representation in the subinterval $[x_{j-1}, x_j]$ of length h_j ($j = 1, \dots, n$):*

$$s_f(x) = \frac{1}{6h_j} \{M_j(x - x_{j-1})^3 + M_{j-1}(x_j - x)^3\} + b_j(x - \frac{x_j + x_{j-1}}{2}) + a_j$$

with the coefficients

$$\begin{aligned} a_j &= \frac{y_j + y_{j-1}}{2} - \frac{1}{12}h_j^2(M_j + M_{j-1}), \\ b_j &= \frac{y_j - y_{j-1}}{h_j} - \frac{1}{6}h_j(M_j - M_{j-1}), \end{aligned}$$

and with the moments M_j given by the linear system (of $\dim n + 1$ resp. $n - 1$)

$$\left(\begin{array}{c|cc|c} \frac{h_1}{3} & \frac{h_1}{6} & & 0 \\ \frac{h_1}{6} & \frac{h_1+h_2}{3} & \frac{h_2}{6} & \\ \hline & \ddots & & \ddots \\ & & \frac{h_{n-1}}{6} & \frac{h_{n-1}+h_n}{3} & \frac{h_n}{6} \\ \hline 0 & & & \frac{h_n}{6} & \frac{h_n}{3} \end{array} \right) \begin{pmatrix} M_0 \\ M_1 \\ \vdots \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} m_0 \\ m_1 \\ \vdots \\ \vdots \\ m_{n-1} \\ m_n \end{pmatrix}$$

and with

$$\begin{aligned} m_0 &= \frac{y_1 - y_0}{h_1} - y_0^{(1)}, \quad m_n = -\frac{y_n - y_{n-1}}{h_n} + y_n^{(1)}, \\ m_j &= \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}, \quad j = 1, \dots, n-1. \end{aligned}$$

Proof: Construction of the interpolating cubic spline

s'' : polygon: linear interpolation between the nodes, $h_j := x_j - x_{j-1}$
moments $M_j := s''(x_j)$, $j = 0, 1, \dots, n$

$\downarrow$ integration

s' : continuity in the nodes

$\downarrow$ integration

s : $\left. \begin{array}{l} \text{continuity in the nodes} \\ \text{and interpolation conditions} \end{array} \right\} \implies \text{linear system for the moments}$

$$[x_{j-1}, x_j] : \quad s''(x) = \frac{1}{h_j} \{M_j(x - x_{j-1}) + M_{j-1}(x_j - x)\}$$

$$\text{Primitive function} \quad s'(x) = \frac{1}{2h_j} \{M_j(x - x_{j-1})^2 - M_{j-1}(x_j - x)^2\} + b_j$$

Primitive function

$$s(x) = \frac{1}{6h_j} \{M_j(x - x_{j-1})^3 + M_{j-1}(x_j - x)^3\} + b_j(x - \frac{x_j + x_{j-1}}{2}) + a_j$$

The unknowns: M_j, b_j, a_j ; number = $(n+1) + n + n = 3n+1$

s' continuous

$$\left. \begin{array}{ll} [x_{j-1}, \underline{\underline{x_j}}] & : \quad s'(x) = \frac{1}{2h_j} \{M_j(x - x_{j-1})^2 - M_{j-1}(x_j - x)^2\} + b_j \\ [\underline{\underline{x_j}}, x_{j+1}] & : \quad s'(x) = \frac{1}{2h_{j+1}} \{M_{j+1}(x - x_j)^2 - M_j(x_{j+1} - x)^2\} + b_{j+1} \end{array} \right\} \underline{\underline{x = x_j}}$$

$$\boxed{\frac{h_j}{2} M_j + b_j = -\frac{h_{j+1}}{2} M_j + b_{j+1}, \quad j = 1, \dots, n-1} \quad \underline{n-1 \text{ equations}}$$

s continuous and s interpolating (i.e., interpolation conditions in x_{j-1} and x_j)

$$\left. \begin{array}{ll} s(x_{j-1}) & = y_{j-1} : \quad \frac{1}{6}h_j^2 M_{j-1} - \frac{1}{2}h_j b_j + a_j = y_{j-1} \\ s(x_j) & = y_j : \quad \frac{1}{6}h_j^2 M_j + \frac{1}{2}h_j b_j + a_j = y_j \end{array} \right\} j = 1, \dots, n$$

$2n$ equations for $2n$ unknowns $b_1, \dots, b_n, a_1, \dots, a_n$,

hence the constants a_j and b_j , $j = 1, \dots, n$ follow.

Furthermore the linear system for the moments M_j ($j = 0, 1, \dots, n$) follows

$$\frac{h_j}{6} M_{j-1} + \frac{h_{j+1} + h_j}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \underbrace{\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}}_{=: m_j} \quad (j = 1, \dots, n-1)$$

$$\begin{aligned} \text{boundary condition (a): } s'(x_0) &= y_0^{(1)}, s'(x_n) = y_n^{(1)}, \\ \frac{h_1}{3} M_0 + \frac{h_1}{6} M_1 &= \frac{y_1 - y_0}{h_1} - y_0^{(1)} =: m_0, \\ \frac{h_n}{6} M_{n-1} + \frac{h_n}{3} M_n &= -\frac{y_n - y_{n-1}}{h_n} + y_n^{(1)} =: m_1. \end{aligned}$$

boundary condition (b): curvature $M_0 = M_n = 0$.

Solvability of the linear system: symmetric triangular matrix, diagonal dominant

$$\text{Gerschgorin} \quad \left| \lambda - \frac{h_{j+1} + h_j}{3} \right| \leq \frac{h_{j+1} + h_j}{6} \rightarrow \lambda \neq 0$$

Hence the interpolating cubic spline is known in each subinterval. $\square$

Uniformly spaced nodes

$$bc (a): \begin{pmatrix} 2 & 1 & & & 0 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & 1 & 4 & 1 \\ 0 & & & & 1 & 2 \end{pmatrix} \quad bc (b): \begin{pmatrix} 4 & 1 & & & 0 \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & & & 1 \\ 0 & & & & 1 & 4 \end{pmatrix}$$

Stability of the cubic interpolation

The linear system is well conditioned: (a): $\text{cond} \leq 6$ (b): $\text{cond} \leq 3$

Representation by B-splines

$$s(x) = \sum_{\nu=-1}^{n+1} c_{\nu} \phi_{\nu}(x) \quad (\text{uniformly spaced nodes})$$

$$s(x_{\nu}) = c_{\nu-1} \underbrace{\phi_{\nu-1}(x_{\nu})}_{\frac{1}{144h}} + c_{\nu} \underbrace{\phi_{\nu}(x_{\nu})}_{\frac{4}{144h}} + c_{\nu+1} \underbrace{\phi_{\nu+1}(x_{\nu})}_{\frac{1}{144h}} = y_{\nu}$$

linear system : the ν -th row 1 4 1

additionally : 2 boundary conditions

Error theory

Choose a uniformly spaced grid Δ with grid length h (for simplicity).

Theorem 4.2 *The interpolating cubic spline s of $f \in C^4[a, b]$ with respect to the uniformly spaced grid Δ with boundary condition (a) or (b) satisfies*

$$\|f - s\|_{\infty} \leq c \cdot h^4 \|f^{(4)}\|_{\infty}$$

with $c = \frac{7}{8}$ for (a) and $c = 2$ for (b).

Proof: Peano–Sard (Theorem 1.2) $\square$

Remark: Fast convergence!

Error $\|f - s\|_{\infty} = O(h^4) \quad (h \rightarrow 0)$

$$\|f' - s'\|_{\infty} = O(h^3) \quad (h \rightarrow 0)$$

Compare with the Cauchy remainder term in case of algebraic interpolation

$$\|f - p_n\|_\infty \leq \frac{\|\omega_{n+1}\|_\infty}{(n+1)!} \|f^{(n+1)}\|_\infty$$

Tschebyscheff nodes

$$\|\omega_{n+1}\|_\infty = \frac{1}{2^n} \rightarrow 0$$

$$\text{but maybe } \|f^{(n+1)}\|_\infty \rightarrow \infty \text{ for } n \rightarrow \infty$$

and divergence of algebraic interpolation processes possible!

Interpolation processes: convergence of the series $\{s_n f\}_{n \geq 0}$

$s_n f$ interpolating spline with respect to $\{x_0^{(n)}, x_1^{(n)}, \dots, x_n^{(n)}\}$

(the grid becomes denser uniformly)

$$\text{Matrix of the nodes} \quad \begin{pmatrix} x_0^{(0)} & & & \\ x_0^{(1)} & x_1^{(1)} & & \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Theorem 4.3 *The series $\{s_n f\}_{n \geq 0}$ corresponding to $f \in C^2[a, b]$ converges uniformly to f and the series $\{s'_n f\}_{n \geq 0}$ uniformly to f' .*

Theorem 4.4 *The interpolating cubic spline s of $f \in C^2[a, b]$ with boundary condition (a) satisfies*

$$\|f'' - s''\|_2 \leq \|f'' - g''\|_2 \text{ for all } g \in G,$$

where

$$G = \{g \in C^2[a, b], g(x_j) = f(x_j), j = 0, 1, \dots, n, g'(x_j) = f'(x_j), j = 0, n\}.$$

Proof: [E]

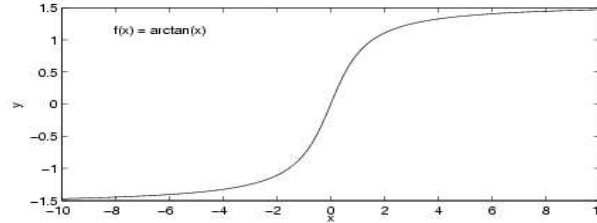
Summary: Approximation results with the interpolating cubic spline are very satisfactory.

SUPPLEMENTARY EXAMPLES – No. 4

POLYNOMIAL AND CUBIC SPLINE INTERPOLATION

1. Interpolation of $f(x) = \arctan x$ in $[-10, 10]$ with 21 nodes:

$$-10, -9, \dots, -1, 0, +1, \dots, +9, +10$$



error $e_p(x) := |\arctan x - p_{20}(x)|$ for polynomial interpolation

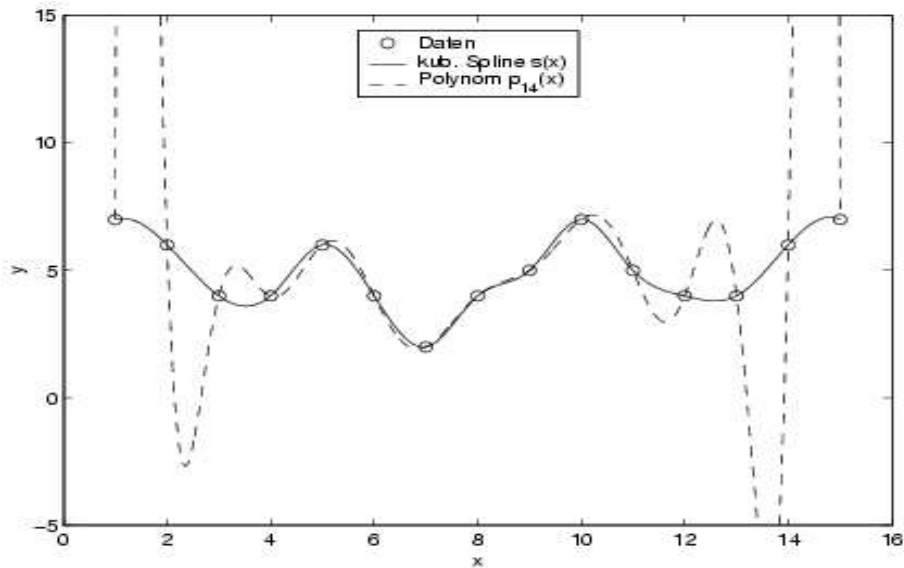
error $e_s(x) := |\arctan x - s_f(x)|$ for spline interpolation with $M_0 = M_n = 0$

x	± 0.5	± 1.5	± 2.5	± 3.5	± 4.5	± 5.5	± 6.5	± 7.5	± 8.5	± 9.5
$e_p(x)$	0.02	0.02	0.01	0.02	0.01	0.02	0.06	0.3	1.5	16.1
$e_s(x)$	0.03	0.01	10^{-3}	10^{-3}	10^{-4}	10^{-4}	10^{-5}	10^{-6}	10^{-5}	10^{-4}

2. Interpolation of data

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y	7	6	4	4	5	4	2	3	5	7	6	4	4	5	7

Graph of the interpolation polynomial $p_{14}(x)$ and of the interpolating cubic spline $s(x)$ with $M_0 = M_n = 0$:



5 Gaussian Quadrature Formulas

Interval $[-1, 1]$, *grid* $-1 \leq x_1 < \dots < x_s \leq 1$

Nodes polynomial $\omega_s(x) = (x - x_1) \dots (x - x_s)$

Weight function $w(x) \geq 0$, $\mu_0 := \int_{-1}^1 w(x)dx < \infty$

Moments $\mu_k := \int_{-1}^1 w(x)x^k dx$, $k = 1, 2, \dots$

Linear functionals $I, Q, R: C[-1, 1] \rightarrow \mathbb{R}$

$$If := \int_{-1}^1 w(x)f(x)dx, \quad Qf := \sum_{j=1}^s w_j f(x_j), \quad Rf := If - Qf$$

Quadrature formula $If = Qf + Rf$

Degree of exactness m , i.e., $Rf = 0$ for all $f \in \mathcal{P}_m$

Interpolatory type quadrature formula

$$w_j = \int_{-1}^1 w(x)\ell_j(x)dx \iff \text{degree of exactness } s-1$$

Problem: Nodes and weights with maximal degree of exactness?

Theorem 5.1 *The degree of exactness of a quadrature formula with s nodes is at most $2s - 1$.*

Proof: Test function $f(x) = \{w_s(x)\}^2$, $f \in \mathcal{P}_{2s}$ satisfies

$$If = \int_{-1}^1 w(x)\{w_s(x)\}^2 dx > 0 \text{ and } Qf = \sum w_j \{w_s(x_j)\}^2 = 0. \quad \square$$

Scalar product $\langle f, g \rangle_w := \int_{-1}^1 w(x)f(x)g(x)dx$

Orthogonal polynomials p_n , $n = 0, 1, 2, \dots$ (with the exact degree n)

Theorem 5.2 *The zeros of orthogonal polynomials p_n are real and simple and are lying in $(-1, 1)$.*

Proof: Assume that p_n has in $(-1, 1)$ exactly $k < n$ distinct zeros $\xi_1, \dots, \xi_k$ of odd multiplicity.

Choose $p(x) := (x - \xi_1) \dots (x - \xi_k)$, $p \in \mathcal{P}_k$,
then $p_n(x)p(x)$ has no change of sign in $(-1, 1)$;

and hence $\langle p_n, p \rangle_w = \int_{-1}^1 \omega(x) p_n(x) p(x) dx \neq 0$,

which is a contradiction to $\langle p_n, p \rangle_w = 0$ for $p \in \mathcal{P}_{n-1}$,

hence $k = n$. $\square$

Definition The quadrature formula $If = Qf + Rf$ is called Gaussian quadrature formula corresponding to w , if the nodes

$$\{x_1, \dots, x_s\} = \{\text{zeros of the orthogonal polynomial } p_s\}$$

and the weights $w_1, \dots, w_s$ are chosen of interpolatory type,

$$\text{i.e., } w_j = \int_{-1}^1 w(x) \frac{p_s(x)}{(x - x_j) p'_s(x_j)} dx, \quad j = 1, \dots, s.$$

Theorem 5.3 The Gaussian quadrature formula with s nodes has the degree of exactness $2s - 1$.

Proof: The quadrature formula is of interpolatory type,

i.e., the degree of exactness is at least $s - 1$.

Euklid's algorithm delivers for $f \in \mathcal{P}_{2s-1}$:

$$f = p_s q + r \quad \text{with } q, r \in \mathcal{P}_{s-1}$$

$$\begin{array}{ccccccc} Rf & = & R p_s q & + & \underbrace{Rr}_{=0} & = & \underbrace{I p_s q}_{=0} - \underbrace{Q p_s q}_{=0} = 0. \\ \uparrow & & \uparrow & & \uparrow & & \uparrow \\ \text{linear funct.} & & \text{interpol.} & & \text{orthog.} & & x_j \text{ zeros of } p_s \end{array} \quad \square$$

Theorem 5.4 The weights $w_1, \dots, w_s$ of the Gaussian quadrature formula are positive.

Proof: Test function $f_k(x) := \left\{ \frac{p_s(x)}{x - x_k} \right\}^2$, $f_k \in \mathcal{P}_{2s-2}$ ($1 \leq k \leq s$)

$$f_k(x_j) = \begin{cases} 0, & j \neq k \\ > 0, & j = k \end{cases}$$

$$0 < If_k = Qf_k = \underbrace{w_k f_k(x_k)}_{>0}, \quad \text{hence it follows } w_k > 0. \quad \square$$

Theorem 5.5 The quadrature error of a Gaussian quadrature formula satisfies

$$Rf = c_{2s} f^{(2s)}(\xi), \quad -1 < \xi < 1, \quad c_{2s} = \frac{1}{(2s)!} \int_{-1}^1 w(x) \tilde{p}_s^2(x) dx,$$

where $\tilde{p}_s$ is the orthogonal polynomial with leading coefficient 1.

Proof: Hermite interpolation (§1) $H_f = \sum_{j=1}^s f(x_j) U_j + \sum_{j=1}^s f'(x_j) V_j \in \mathcal{P}_{2s-1}$

$$V_j(x) = (x - x_j) \ell_j^2(x) = (x - x_j) \left\{ \frac{p_s(x)}{(x - x_j) p'_s(x_j)} \right\}^2 = \frac{1}{(p'_s(x_j))^2} \underbrace{\frac{p_s(x)}{x - x_j}}_{\in \mathcal{P}_{s-1}} p_s(x)$$

i) The ansatz

$$\int_{-1}^1 w(x)f(x)dx = \int_{-1}^1 w(x)H_f(x)dx + \underbrace{\int_{-1}^1 w(x)(f(x) - H_f(x))dx}_{\text{error}}$$

also implies the Gaussian quadrature formula:

$$\begin{aligned} \int_{-1}^1 w(x)H_f(x)dx &= \sum_{j=1}^s f(x_j) \underbrace{\int_{-1}^1 w(x)U_j(x)dx}_{=: \tilde{w}_j} + \sum_{j=1}^s f'(x_j) \underbrace{\int_{-1}^1 w(x)V_j(x)dx}_{=0 \text{ orthogonal}} \\ 0 = RH_f &= IH_f - Q_s H_f = \sum_{j=1}^s f(x_j) \tilde{w}_j - \sum_{j=1}^s w_j \underbrace{H_f(x_j)}_{=f(x_j)} \end{aligned}$$

inserting the test function $f_k \in \mathcal{P}_{2s-2}$ implies $\tilde{w}_j = w_j$.

ii) Interpolation error $f(x) - H_f(x) = \frac{1}{(2s)!} f^{(2s)}(\xi) w_s^2(x)$
 Quadrature error

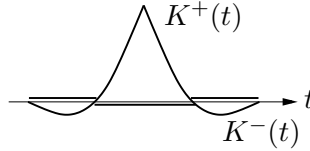
$$\begin{aligned} Rf &= R(f - H_f) = I(f - H_f) - \underbrace{Q_s(f - H_f)}_{=0 \text{ interpolatory}} \\ &= \frac{1}{(2s)!} \int_{-1}^1 w(x) f^{(2s)}(x) w_s^2(x) dx \\ &= \underbrace{\frac{1}{(2s)!} \int_{-1}^1 w(x) w_s^2(x) dx}_{=: c_{2s} > 0} \cdot f^{(2s)}(\xi) \quad (w_s = \tilde{p}_s). \quad \square \end{aligned}$$

Theorem 5.6 *The Peano kernel K in*

$$Rf = \int_{-1}^1 K(t) f^{(2s)}(t) dt$$

of a Gaussian quadrature formula is positive definite.

Proof: $K = K^+ + K^-$
 K^+, K^- continuous



Choose f such that $f^{(2s)}(t) = K^-(t)$, then it follows (Theorem 5.5)

$$Rf = c_{2s} f^{(2s)}(\xi) = c_{2s} K^-(\xi) \leq 0 \quad \text{and}$$

$$Rf = \int_{-1}^1 K(t) f^{(2s)}(t) dt = \int_{-1}^1 K(t) K^-(t) dt = \int_{-1}^1 (K^-(t))^2 dt \geq 0,$$

$Rf \neq 0$, hence $K^-(t) \equiv 0$. $\square$

Theorem 5.7 *The Gaussian quadrature formulas are stable.*

Proof: Stability $\rightarrow$ error propagation $\tilde{f}(x_j) = \begin{cases} f(x_j) + \varepsilon_j, |\varepsilon_j| \leq \varepsilon \\ f(x_j)(1 + \delta_j), |\delta_j| \leq \delta \end{cases}$

the true value $Q_s f = \sum w_j f(x_j)$
 perturbed value $Q_s \tilde{f} = \sum w_j \tilde{f}(x_j)$ } what is the effect?

Absolute error $|Q_s f - Q_s \tilde{f}| \leq \|Q_s\| \cdot \varepsilon$

Relative error $\left| \frac{Q_s f - Q_s \tilde{f}}{Q_s f} \right| \leq \underbrace{\frac{\max |f(x_j)|}{|Q_s f|}}_{\text{natural}} \cdot \underbrace{\|Q_s\|}_{\text{numerical}} \cdot \delta$

condition number

Positive weights $\|Q_s\| = \sum_{j=1}^s |w_j| = \int_{-1}^1 w(x) dx = \mu_0$

Natural stability

$$|If - I\tilde{f}| \leq \mu_0 \cdot \varepsilon, \quad \|I\| = \int_{-1}^1 \omega(x) dx = \mu_0,$$

$$\left| \frac{If - I\tilde{f}}{If} \right| \leq \frac{\|f\|_\infty}{|If|} \mu_0 \cdot \delta. \quad \square$$

Theorem 5.8 *The Gaussian quadrature formula $\{Q_s f\}_{s \geq 1}$ is convergent, i.e., $\lim_{s \rightarrow \infty} Q_s f = If$.*

Proof

1. $\{Q_s\}_{s \geq 1}$ series of bounded functionals ($\|Q_s\| \leq \mu_0$)
2. $\{Q_s f\}_{s \geq 1}$ convergent for each algebraic polynomial

Hence, the convergence follows (Theorem of Banach–Steinhaus). $\square$

Classical orthogonal polynomials

Jacobian polynomials $P_n^{(\alpha,\beta)}$

Interval $[-1, 1]$: $w(x) = (1-x)^\alpha(1+x)^\beta$, $\alpha, \beta > -1$ (singularity ± 1)

Special cases

$w(x) = 1$: **Legendre polynomials** $P_n(\mu_0 = 2)$

$w(x) = \frac{1}{\sqrt{1-x^2}}$: **Tschebyscheff polynomials** of the first kind T_n ($\mu_0 = \pi$)

$w(x) = \sqrt{1-x^2}$: **Tschebyscheff polynomials** of the second kind U_n ($\mu_0 = \frac{\pi}{2}$)

Recurrence formulas

$$P_0(x) = 1, P_1(x) = x, \quad (n+1) P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$$

$$T_0(x) = 1, T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

$$U_0(x) = 1, U_1(x) = 2x, \quad U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x)$$

Infinite intervals (Infinite integrals)

$[0, \infty)$: $w(x) = e^{-x}$: *Laguerre polynomials* L_n ($\mu_0 = 1$)

$$L_0(x) = 1, L_1(x) = -x + 1, L_{n+1}(x) = (1 + 2n - x)L_n(x) - n^2L_{n-1}(x)$$

$(-\infty, \infty)$: $w(x) = e^{-x^2}$: *Hermite polynomials* H_n ($\mu_0 = \sqrt{\pi}$)

$$H_0(x) = 1, H_1(x) = 2x, H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$$

$$\text{Gauss-Legendre} \quad \int_{-1}^1 f(x)dx = \sum_{j=1}^s w_j f(x_j) + Rf$$

$$\text{Gauss-Tschebyscheff} \quad \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} f(x)dx = \frac{\pi}{s} \sum_{j=1}^s f(x_j) + Rf$$

All the weights are equal and the nodes $x_j = \cos \frac{2j-1}{2s}\pi$, $j = 1, \dots, s$.

$$\text{Gauss-Laguerre} \quad \int_0^\infty e^{-x} f(x)dx = \sum_{j=1}^s w_j f(x_j) + Rf$$

$$\text{Gauss-Hermite} \quad \int_{-\infty}^\infty e^{-x^2} f(x)dx = \sum_{j=1}^s w_j f(x_j) + Rf$$

The main problem: Computation of the nodes and weights

Eigenvalue problem: symmetric tridiagonal matrix
the eigenvalues are the nodes x_j
the first coefficient of the normed eigenvector
is equal to the weight w_j
QR-Algorithm $\rightarrow$ Method of Golub & van Loan

SUPPLEMENTARY EXAMPLES – NO. 5

GAUSSIAN QUADRATURE

P.J. Davis & P. Rabinowitz: Methods of Numerical Integration (1984).

Examples

1. $\int_{-1}^1 \frac{\cos x}{\sqrt{1-x^2}} dx = 2.403\ 94\dots$ (Bessel function)
Singularities in ± 1

Mid-point rule $Q_1 f = 2$, $Q_2 f = 2.02\dots$, $Q_4 f = 2.106\dots$

Gauss-Legendre $Q_2 f = 2.05\dots$

Gauss-Tschebyscheff $Q_2 f = 2.388\dots$

2. $\int_0^\infty e^{-x} x^{t-1} dx = \Gamma(t)$ ($t > 0$) (Gamma function)

$$\Gamma(1.5) = \frac{1}{2}\sqrt{\pi} = 0.8862\dots$$

Gauss-Laguerre $Q_4 f = 0.8992\dots$, $Q_8 f = 0.8910\dots$, $Q_{12} f = 0.8887\dots$

3. $e^{-2} \int_2^\infty \frac{\sin x}{x} dx = 0.004\ 683$ (oscillating)

Infinite integral $\int_2^\infty \dots = \int_2^M \dots + \int_M^\infty \dots$

Quadrature formula $\int_2^M \dots = Qf + Rf$ and estimation $|\int_M^\infty \dots| \leq \varepsilon :$

$$\int_{2k\pi}^\infty \frac{\sin x}{x} dx = \sum_{j=k}^\infty \left\{ \int_{2j\pi}^{(2(j+1))\pi} + \dots \int_{(2j+1)\pi}^{(2(j+1))\pi} \dots \right\}$$

$$\leq \sum_{j=k}^\infty \left\{ \frac{2}{2j\pi} - \frac{2}{2(j+1)\pi} \right\} = \frac{1}{k\pi}$$

Transformation $e^{-2} \int_0^\infty e^{-t} e^t \frac{\sin(t+2)}{t+2} dt$

Gauss-Laguerre $Q_8 f = -0.04\dots$, $Q_{16} f = -0.039\dots$, $Q_{32} f = -0.000\ 23\dots$

4. $e^{-2} \int_2^\infty \frac{\sin(x-1)}{\sqrt{x(x-2)}} dx = 0.162\ 668 \dots$ (Bessel function)

Transformation $e^{-2} \int_0^\infty e^{-t} e^t \frac{\sin(t+1)}{\sqrt{t(t+2)}} dt$

Gauss-Laguerre

$Q_8 f = 0.039 \dots$, $Q_{16} f = -0.097 \dots$, $Q_{32} f = 0.1007 \dots$

5. $e^{-2} \int_2^\infty e^{-x^2} dx = 0.000\ 561\ 0371$

Gauss-Laguerre

$Q_4 f = 0.000\ 512 \dots$, $Q_8 f = 0.000\ 5638 \dots$, $Q_{16} f = 0.000\ 561\ 007 \dots$

6. $e^{-2} \int_2^\infty \frac{x}{e^x - 1} dx = 0.058\ 3349 \dots$ (Debye function)

Gauss-Laguerre $Q_4 f = 0.058\ 3351 \dots$, $Q_8 f = 0.058\ 3348 \dots$

7. $e^{-2} \int_2^\infty \frac{1}{x(\log x)^2} dx = 0.195 \dots$

Gauss-Laguerre

$Q_4 f = 0.145 \dots$, $Q_8 f = 0.155 \dots$, $Q_{16} f = 0.166 \dots$, $Q_{32} f = 0.167 \dots$

8. $e^{-2} \int_2^\infty \frac{1}{x(\log x)^{3/2}} dx = 0.325 \dots$

Gauss-Laguerre

$Q_4 f = 0.16 \dots$, $Q_8 f = 0.17 \dots$, $Q_{16} f = 0.19 \dots$, $Q_{32} f = 0.20 \dots$

9. $e^{-2} \int_2^\infty \frac{dx}{x^{1.01}} = 13.6 \dots$

Gauss-Laguerre

$Q_4 f = 0.2 \dots$, $Q_8 f = 0.3 \dots$, $Q_{16} f = 0.4 \dots$, $Q_{32} f = 0.5 \dots$

SUPPLEMENTARY EXAMPLES – NO. 6

GAUSS–LAGUERRE FORMULAS

A.H. Stroud & D.H. Secrest

Gaussian Quadrature Formulas, Prentice Hall, 1966

Nodes and weights of the Gauss–Laguerre quadrature formulas(nodes x_j = zeros of the Laguerre polynomial $L_s(x)$)

$$\int_0^\infty e^{-x} f(x) dx = \sum_{j=1}^n w_j f(x_j) + Rf$$

$s = 2$	x_j	w_j
	0.58578 64376 27	8.53553 390593 (-1)
	3.41421 35623 73	1.46446 609407 (-1)
$s = 4$	x_j	w_j
	0.32254 76896 19	6.03154 104342 (-1)
	1.74576 11011 58	3.57418 692438 (-1)
	4.53662 02969 21	3.88879 085150 (-2)
	9.39507 09123 01	5.39294 705561 (-4)
$s = 6$	x_j	w_j
	0.22284 66041 79	4.58964 673950 (-1)
	1.18893 21016 73	4.17000 830772 (-1)
	2.99273 63260 59	1.13373 382074 (-1)
	5.77514 35691 05	1.03991 974531 (-2)
	9.83746 74183 83	2.61017 202815 (-4)
	15.98287 39806 02	8.98547 906430 (-7)
$s = 8$	x_j	w_j
	0.17027 96323 05	3.69188 589342 (-1)
	0.90370 17767 99	4.18786 780814 (-1)
	2.25108 66298 66	1.75794 986637 (-1)
	4.26670 01702 88	3.33434 922612 (-2)
	7.04590 54023 93	2.79453 623523 (-3)
	10.75851 60101 81	9.07650 877336 (-5)
	15.74067 86412 78	8.48574 671627 (-7)
	22.86313 17368 89	1.04800 117487 (-9)

$s = 10$

x_j	w_j
0.13779 34705 40	3.08441 115765 (-1)
0.72945 45495 03	4.01119 929155 (-1)
1.80834 29017 40	2.18068 287612 (-1)
3.40143 36978 55	6.20874 560987 (-2)
5.55249 61400 64	9.50151 697518 (-3)
8.33015 27467 64	7.53008 388588 (-4)
11.84378 58379 00	2.82592 334960 (-5)
16.27925 78313 78	4.24931 398496 (-7)
21.99658 58119 81	1.83956 482398 (-9)
29.92069 70122 74	9.91182 721961 (-13)

 $s = 12$

x_j	w_j
0.11572 21173 58	2.64731 371055 (-1)
0.61175 74845 15	3.77759 275873 (-1)
1.51261 02697 76	2.44082 011320 (-1)
2.83375 13377 44	9.04492 222117 (-2)
4.59922 76394 18	2.01023 811546 (-2)
6.84452 54531 15	2.66397 354197 (-3)
9.62131 68424 57	2.03232 592663 (-4)
13.00605 49933 06	8.36505 585682 (-6)
17.11685 51874 62	1.66849 387654 (-7)
22.15109 03793 97	1.34239 103052 (-9)
28.48796 72509 84	3.06160 163504 (-12)
37.09912 10444 67	8.14807 746743 (-16)

6 Extrapolation Methods

Principle: **Richardson extrapolation** $h \rightarrow 0$

Application: Discretization (quadrature, numerical differentiation, differential equations)

Example: Numerical differentiation

$$\underbrace{f'(0)}_{=:L} = \underbrace{\frac{f(h) - f(0)}{h}}_{=:L(h)} + Rf \quad (h > 0) \quad (\text{difference quotient})$$

The smaller the step size h becomes the better is the approximation (consistency):

$$L - L(h) = O(h) \quad (h \rightarrow 0), \quad L = L(0)$$

In applications h cannot become arbitrarily small

(rounding errors, computational costs)!

Better approximation by combination of the values $L(h_0)$ und $L(h_1)$ with h_0, h_1 .

Error: Asymptotic evaluation (in powers of h)

$$\left. \begin{array}{ll} h_0 := h : & L - L(h) = -\frac{h}{2!}f_0'' - \frac{h^2}{3!}f_0''' - \dots = O(h) \quad (h \rightarrow 0) | \cdot -1 \\ h_1 := \frac{h}{2} : & L - L(\frac{h}{2}) = -\frac{h}{2 \cdot 2!}f_0'' - \frac{h^2}{4 \cdot 3!}f_0''' - \dots = O(h) \quad (h \rightarrow 0) | \cdot 2 \end{array} \right\} \Rightarrow$$

$$L - \underbrace{\{2L(\frac{h}{2}) - L(h)\}}_{=: \tilde{L}(h)} = \frac{h^2}{12}f_0''' + \dots = O(h^2) \quad (h \rightarrow 0)$$

Approximation order for $\tilde{L}(h)$ is higher than for $L(h)$

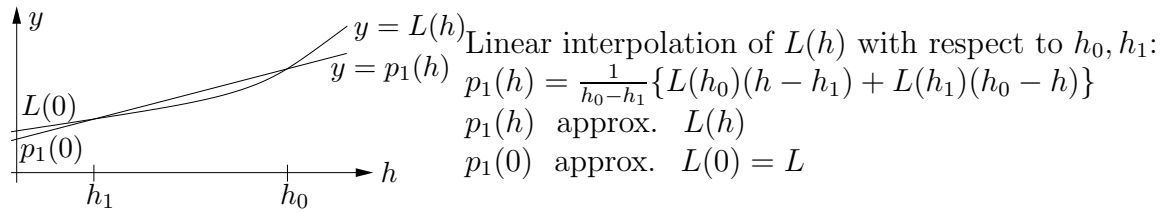
New formula (difference)

$$f'(0) = \frac{1}{h} \underbrace{\{-f(h) + 4f(\frac{h}{2}) - 3f(0)\}}_{=: \tilde{L}(h)} + O(h^2) \quad (h \rightarrow 0)$$

Acceleration of convergence: step size series $h_0 > h_1 > h_2 > \dots$

$$\begin{array}{llll} L(h_0) & & & \\ & \rangle & \tilde{L}(h_0) & \\ L(h_1) & & & \\ & \rangle & \tilde{\tilde{L}}(h_0) & \\ & & \tilde{L}(h_1) & \\ L(h_2) & & \vdots & \ddots \\ \vdots & & \vdots & \\ \tilde{\tilde{L}}(h) & := & \frac{1}{3}\{2^2\tilde{L}(\frac{h}{2}) - \tilde{L}(h)\} = \frac{1}{3h}\{f(h) - 12f(\frac{h}{2}) + 32f(\frac{h}{4}) - 21f(0)\} \end{array}$$

Order of approximation for $\tilde{\tilde{L}}(h)$ is $O(h^3)$ ($h \rightarrow 0$)

Extrapolation for $h \rightarrow 0$ 

Step size $h_0 := h, h_1 := \frac{h}{2} \implies p_1(0) = 2L(\frac{h}{2}) - L(h)$

Interpolation of $L(h)$ with respect to the nodes $h_0 > h_1 > \dots > h_k$: $p_k(h)$

$$L(h) - p_k(h) = \frac{1}{(k+1)!} L^{(k+1)}(\xi) (h - h_0)(h - h_1) \dots (h - h_k)$$

$$L(0) - p_k(0) = O(h_0^{k+1}) \quad (h_0 \rightarrow 0) \quad \text{approximation order } k+1$$

General procedure

Given: problem with the solution L
 Discretization: problem with the solution $L(h)$ $\left. \vphantom{\begin{matrix} \text{Given:} \\ \text{Discretization:} \end{matrix}} \right\} L - L(h) = O(h) \quad (h \rightarrow 0)$

Asymptotic evaluation (the existence is assumed)

$$L - L(h) = c_1 h + c_2 h^2 + c_3 h^3 + \dots \quad (c_1 = \dots = c_{p-1} = 0 \text{ possible})$$

Step size series $h_0 > h_1 > h_2 > \dots$ (suitable)

Interpolation polynomial of $L(h)$ with respect to $h_0, h_1, \dots, h_k$:

$$p_k(0) \text{ approximates } L(0) = L$$

Aitken–Neville (§1): Computation of $p_k(0)$

$$\begin{array}{ccccccc}
 h_0 & L(h_0) & =: & B_{00} & & & \\
 & & & & \rangle & & \\
 h_1 & L(h_1) & =: & B_{01} & & B_{10} & \\
 & & & & \rangle & & \\
 h_2 & L(h_2) & =: & B_{02} & & B_{11} & B_{20} \\
 & & & & \rangle & & \\
 h_3 & L(h_3) & =: & B_{03} & & B_{12} & B_{21} & B_{30} \\
 \vdots & \vdots & & \vdots & & \vdots & \vdots & \ddots \\
 \text{order } p & & = & 1 & 2 & 3 & 4 & \dots
 \end{array}$$

$$L - B_{ij} = O(h^{i+1}) \quad (h \rightarrow 0)$$

Computation of functions only in the first column, otherwise combinations!

Formulas (see Aitken–Neville):

$$B_{ij} := \frac{1}{h_j - h_{j+i}} \{h_j B_{i-1, j+1} - h_{j+i} B_{i-1, j}\}$$

Bisection of the step size $h_{j+1} := \frac{1}{2}h_j$, $j = 0, 1, 2, \dots$:

$$\begin{aligned} B_{1j} &:= 2B_{0,j+1} - B_{0,j} \\ B_{2j} &:= \frac{4B_{1,j+1} - B_{1,j}}{3} \\ &\vdots \\ \text{special case } B_{2\nu,j} &:= \frac{4^\nu B_{2\nu-1,j+1} - B_{2\nu-1,j}}{4^\nu - 1} \rightarrow \text{Romberg method} \end{aligned}$$

Convergence: in columns linear, in diagonals superlinear

Effectivity: Suitable choice of the step size series (to save computations)
Special asymptotic evaluations (e.g., in powers of h^2)

Example: Numerical differentiation for $f(x) = \tan \frac{\pi}{2}x$

Compute $f'(0) = \pi/2 = 1.570\,7963\dots$

by $L(h) := \frac{1}{h}(f(h) - f(0))$

Scheme: Step size series $h_0 = \frac{1}{2}, h_1 = \frac{1}{4}, \dots$

$$\begin{array}{ccccccc} \frac{1}{2} & 2.000 & & & & & \\ \frac{1}{4} & \underline{1.656\dots} & \underline{1.542\,47\dots} & & & & \\ \frac{1}{8} & \underline{1.591\dots} & \underline{1.569\,45\dots} & \underline{1.571\,25\dots} & & & \\ \frac{1}{16} & \underline{1.575\dots} & \underline{1.570\,72\dots} & \underline{1.570\,80\dots} & \underline{1.570\,79\dots} & & \end{array}$$

Refinements

$$\left. \begin{array}{l} \text{Order } p: \quad L - L(h) = \underline{c_p h^p} + c_{p+1} h^{p+1} + \dots \quad | \cdot -1 \\ \quad \quad \quad \underline{L - L(\frac{h}{2}) = c_p (\frac{h}{2})^p + c_{p+1} (\frac{h}{2})^{p+1} + \dots} \quad | \cdot 2^p \end{array} \right\} \text{Addition}$$

$$(2^p - 1)L + L(h) - 2^p L(\frac{h}{2}) = -\frac{1}{2}c_{p+1}h^{p+1} + \dots$$

$$L - \underbrace{\frac{2^p L(\frac{h}{2}) - L(h)}{2^p - 1}} = -\frac{c_{p+1}}{2(2^p - 1)}h^{p+1} + \dots$$

$$=: \tilde{L}(h) \quad \text{approximation order } p+1$$

$$\frac{2^{p+1}\tilde{L}(\frac{h}{2}) - \tilde{L}(h)}{2^{p+1} - 1} =: \tilde{\tilde{L}}(h) \quad \text{approximation order } p+2$$

Error estimation

Leading term of the error $\boxed{c_p h^p}$: subtract the second row from the first row

$$\begin{aligned} L\left(\frac{h}{2}\right) - L(h) &= \left(1 - \frac{1}{2^p}\right) c_p h^p + O(h^{p+1}) \quad (h \rightarrow 0) \\ c_p h^p &= \frac{L\left(\frac{h}{2}\right) - L(h)}{1 - \frac{1}{2^p}} + O(h^{p+1}) \quad (h \rightarrow 0) \\ c_p \left(\frac{h}{2}\right)^p &= \frac{L\left(\frac{h}{2}\right) - L(h)}{2^p - 1} + O(h^{p+1}) \quad (h \rightarrow 0) \end{aligned}$$

Useful for step size control solving differential equations and for the Romberg method.

Special evaluations

$$L - L(h) = c_2 h^2 + c_4 h^4 + c_6 h^6 + \dots \quad (h \rightarrow 0) \quad (\text{see trapezoidal rule})$$

In each extrapolation step the order will be increased by two.

Romberg method “Automatic integration” (Numer. Math. I)

Compute an approximation N for $I := \int_a^b f(x) dx$ with $|N - I| \leq \varepsilon$

Trapezoidal sum (n subintervals of length $h = \frac{b-a}{n}$)

$$\int_a^b f(x) dx = h \underbrace{\sum_{\nu=0}^n {}''f(a + \nu h)}_{=: T(h)} + Rf \quad (\Sigma'' : \text{first and last term are cut in halves})$$

Convergence $T(h) \rightarrow I = T(0)$ for $h \rightarrow 0$

$$\text{Error (Numer. Math. I): } Rf = -\frac{(b-a)^3}{12} \frac{1}{n^2} f''(\xi) = O(h^2) \quad (h \rightarrow 0)$$

Asymptotic evaluation (f sufficiently smooth)

$$I - T(h) = c_2 h^2 + c_4 h^4 + \dots \quad (h \rightarrow 0), \quad (\text{powers in } h^2)$$

Romberg series $h_0, h_1, h_2, \dots, h_\nu := \frac{h_0}{2^\nu}, h_0 = b - a$ (bisection)

Combinations

$$\begin{aligned} \tilde{T}(h) &:= \frac{4T(\frac{h}{2}) - T(h)}{3} && \text{Kepler } \Sigma, && \text{degree of exactness 3} \\ \tilde{\tilde{T}}(h) &:= \frac{4^2 \tilde{T}(\frac{h}{2}) - \tilde{T}(h)}{4^2 - 1} && \text{Boole } \Sigma, && \text{degree of exactness 5} \\ &\vdots && && \end{aligned}$$

Abbreviations

$$T_{0n} := h_n \sum_{j=0}^{2^n} f(a + jh_n), \quad n = 0, 1, \dots$$

$$M_{0n} := h_n \sum_{j=1}^{2^n} f(a + \frac{2j-1}{2}h_n), \quad n = 0, 1, \dots$$

$$T_{0,n+1} := \frac{1}{2}(T_{0,n} + M_{0,n}) \quad \text{trapezoidal } \Sigma + \text{mid-point } \Sigma$$

T-Scheme: Elimination of the power h_n^{2m}

$$T_{m,n} := \frac{4^m T_{m-1,n+1} - T_{m-1,n}}{4^m - 1} = I + O(h_n^{2m+2}) \quad (h_n \rightarrow 0)$$

T_{00}					
T_{01}	T_{10}				
T_{02}	T_{11}	T_{20}			
T_{03}	T_{12}	T_{21}	T_{30}		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
trapezoidal	Kepler	Boole	QF		
order	2	4	6	8	10 ...

*Properties of the T-scheme*Realization row by row: T_{m0} is the best approximationComputational costs: Number of functional computations until $T_{0m} = 2^m + 1$

Convergence: in columns linear, in the main diagonal superlinear

Stability: Positive weights

Error: Peano kernel is definite

$$I - T_{m,n} = 4^{-n(m+1)} c_m f^{(2m+2)}(\xi_{mn}) = O(4^{-n(m+1)}) \quad (m, n \rightarrow \infty)$$

Error estimates

$$|I - T_{m,0}| = |T_{m,1} - T_{m,0}| + \text{remainder term}$$

$$|I - T_{m,1}| = \left| \frac{T_{m,1} - T_{m,0}}{4^{m+1}} \right| + \text{remainder term}$$

Stopping criterion (fast convergence!)

$$|T_{m,1} - T_{m,0}| \cdot d_m \leq \varepsilon \implies |If - T_{m+1,0}| \leq \varepsilon$$

d_m damping factor, e.g. $d_m = 4^{-m-1}$

Example: $\int_1^2 \frac{dx}{x} = \ln 2 = 0.693\,147\,180\dots$

0.75

0.708... 0.694...

0.697... 0.693 25... 0.693 17...

0.694... 0.693 15... 0.693 148... 0.693 147 47...

$$|T_{21} - T_{20}| = 3 \cdot 10^{-5}, \quad d_2 |T_{21} - T_{20}| = 4.7 \cdot 10^{-7},$$

$$|I - T_{20}| = 3 \cdot 10^{-5}, \quad |I - T_{21}| = 1 \cdot 10^{-6}, \quad |I - T_{30}| = 3 \cdot 10^{-7}$$

Asymptotic evaluation for the trapezoidal sum

Theorem 6.1 (*Formula of Euler–MacLaurin*)

Let $f \in C^{2k+1}[0, 1]$, then it holds

$$\begin{aligned} \frac{1}{2}[f(0) + f(1)] &= \int_0^1 f(x)dx + \frac{b_2}{2!}[f'(1) - f'(0)] + \frac{b_4}{4!}[f'''(1) - f'''(0)] + \dots \\ &\quad \dots + \frac{b_{2k}}{(2k)!}[f^{(2k-1)}(1) - f^{(2k-1)}(0)] + R_{2k+1}f, \quad \text{where} \end{aligned}$$

$$R_{2k+1}f = \int_0^1 B_{2k+1}(x)f^{(2k+1)}(x)dx,$$

$B_n(x)$ Bernoulli polynomials,

$b_n := B_n(0)n!$ Bernoulli numbers.

Remark: Euler–MacLaurin $f(0), f(1), f'(0), f'(1), f'''(0), f'''(1)$

Taylor evaluation $f(0), f'(0), f''(0), \dots$,

Bernoulli polynomials

$$B_0(x) := 1$$

$$B_{n+1}(x) := \int B_n(x)dx \quad \text{with a constant such that} \quad \int_0^1 B_{n+1}(x)dx = 0$$

$$B_1(x) = x - \frac{1}{2}$$

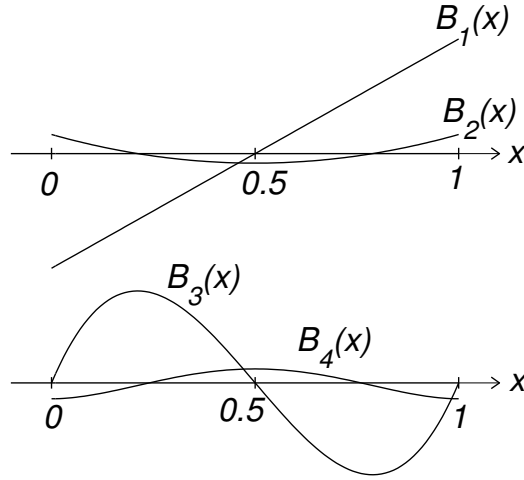
$$B_2(x) = \frac{1}{2}x^2 - \frac{1}{2}x + \frac{1}{12}$$

$$B_3(x) = \frac{1}{6}x^3 - \frac{1}{4}x^2 + \frac{1}{12}x$$

$$B_4(x) = \frac{1}{24}x^4 - \frac{1}{12}x^3 + \frac{1}{24}x^2 - \frac{1}{720}$$

$\vdots$

Leading coefficient of $B_n : \frac{1}{n!}$

Examples

Symmetry with respect to $\frac{1}{2}$, i.e., B_{2k} even: $B_{2k}(\frac{1}{2} + x) = B_{2k}(\frac{1}{2} - x)$

B_{2k+1} odd: $B_{2k+1}(\frac{1}{2} + x) = -B_{2k+1}(\frac{1}{2} - x)$

Bernoulli numbers $b_n := B_n(0)n!$

$$B_{2k}(0) = B_{2k}(1) = \frac{b_{2k}}{(2k)!}$$

$$B_{2k+1}(0) = B_{2k+1}(1) = 0 \quad (k \geq 1) \quad \text{i.e., } b_{2k+1} = 0$$

$$B_n(x) = \frac{1}{n!} \sum_{\nu=0}^n \binom{n}{\nu} b_{\nu} x^{n-\nu} \implies B'_{n+1}(x) = B_n(x)$$

$$\text{Recurrence: } b_0 = 1, b_n = -\frac{1}{n+1} \sum_{\nu=0}^{n-1} \binom{n+1}{\nu} b_{\nu}, \quad n = 1, 2, \dots$$

$$b_0 = 1, \quad b_1 = -\frac{1}{2}, \quad b_2 = \frac{1}{6}, \quad b_4 = -\frac{1}{30}, \quad b_6 = \frac{1}{42}, \quad b_8 = -\frac{1}{30}, \quad b_{10} = \frac{5}{16}, \\ b_{12} = -\frac{691}{2730}, \quad b_{14} = \frac{7}{6}, \quad b_{16} = -\frac{3617}{510}, \quad \dots$$

Proof of Theorem 6.1:

$$\begin{aligned} \text{Integration by parts } \int_0^1 u'v &= uv \Big|_0^1 - \int_0^1 uv' \\ \int_0^1 \underbrace{B_0(x)}_{=1} f(x) dx &= \underbrace{B_1(x)}_{B_1(1)=-B_1(0)=\frac{1}{2}} f(x) \Big|_0^1 - \int_0^1 B_1(x) f'(x) dx \end{aligned}$$

$$\frac{1}{2} (f(0) + f(1)) = \int_0^1 f(x) dx + \int_0^1 B_1(x) f'(x) dx$$

$$\begin{aligned}
\int_0^1 B_1(x) f'(x) dx &= \underbrace{B_2(x)}_{B_2(1)=B_2(0)=\frac{b_2}{2!}} f'(x) \Big|_0^1 - \int_0^1 B_2(x) f''(x) dx \\
&= \frac{b_2}{2!} [f'(1) - f'(0)] - \int_0^1 B_2(x) f''(x) dx \\
- \int_0^1 B_2(x) f''(x) dx &= - \underbrace{B_3(x)}_{B_3(1)=B_3(0)=0} f''(x) \Big|_0^1 + \int_0^1 B_3(x) f'''(x) dx \\
&= \underbrace{B_4(x)}_{B_4(1)=B_4(0)=\frac{b_4}{4!}} f'''(x) - \int_0^1 B_4(x) f^{(4)}(x) dx \\
&= \frac{b_4}{4!} [f'''(1) - f'''(0)] - \int_0^1 B_4(x) f^{(4)}(x) dx
\end{aligned}$$

In general ($\ell \geq 1$)

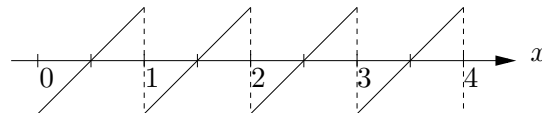
$$\begin{aligned}
- \int_0^1 B_{2\ell}(x) f^{(2\ell)}(x) dx &= \int_0^1 B_{2\ell+1}(x) f^{(2\ell+1)}(x) dx \\
\int_0^1 B_{2\ell+1}(x) f^{(2\ell+1)}(x) dx &= \frac{b_{2\ell+2}}{(2\ell+2)!} [f^{(2\ell+1)}(1) - f^{(2\ell+1)}(0)] \\
&\quad - \int_0^1 B_{2\ell+2}(x) f^{(2\ell+2)}(x) dx \quad \square
\end{aligned}$$

Trapezoidal Σ : Subintervals $[0, 1], [1, 2], \dots, [n-1, n]$
 $\rightarrow$ compensation of derivatives

$$\begin{aligned}
\sum_{j=0}^n {}''f(j) &= \int_0^n f(x) dx + \sum_{j=1}^k \frac{b_{2j}}{(2j)!} [f^{(2j-1)}(n) - f^{(2j-1)}(0)] \\
&\quad + \int_0^n B_{2k+1}^*(x) f^{(2k+1)}(x) dx
\end{aligned}$$

$B_{2k+1}^*(x)$ by *periodic continuation* of $B_{2k+1}(x)$ in $[0, 1)$

Example: $B_1(x) = x - \frac{1}{2}$ in $[0, 1)$
 $B_1^*(x) = x - [x] - \frac{1}{2} = B_1(x - [x])$
in $[0, n]$ *Saw-tooth function*



$$B_{2k+1}^*(x) = B_{2k+1}(x - [x])$$

Theorem 6.2 *Let $f \in C^{2k+1}[a, b]$, then the asymptotic evaluation holds*

$$T_f(h) = \int_a^b f(x)dx + \sum_{j=1}^k \frac{b_{2j}}{(2j)!} [f^{(2j-1)}(b) - f^{(2j-1)}(a)] h^{2j} + h^{2k+1} R_{2k+1}^* f.$$

Remark: Evaluation in powers of h^2

Proof: Transform $[0, n] \rightarrow [a, b]$ with grid length $h = \frac{b-a}{n}$
 $x \mapsto t = a + xh, \quad dt = h \, dx$
 $f(x) = f\left(\frac{t-a}{h}\right) =: \tilde{f}(t), \quad f^{(\nu)}(x) = h^\nu \tilde{f}^{(\nu)}(t)$

$$\begin{aligned} \frac{1}{h} \underbrace{T_{\tilde{f}}(h)}_{=h \sum'' \tilde{f}(a+jh)} &= \frac{1}{h} \int_a^b \tilde{f}(t) dt + \sum_{j=1}^k \frac{b_{2j}}{(2j)!} h^{2j-1} [\tilde{f}^{(2j-1)}(b) - \tilde{f}^{(2j-1)}(a)] \\ &\quad + \underbrace{\frac{h^{2k+1}}{h} \int_a^b \tilde{B}_{2k+1}^*(t) \tilde{f}^{(2k+1)}(t) dt}_{=: R_{2k+1}^* \tilde{f}} \quad \square \end{aligned}$$

Periodic function $f \in C_{[a,b]}^{2k+1}$, i.e.,

$$f'(a) = f'(b), \dots, f^{(2k-1)}(a) = f^{(2k-1)}(b)$$

Corollary 6.3 *Let $f \in C_{[a,b]}^{2k+1}$, then the error of the trapezoidal rule satisfies*

$$\int_a^b f(x)dx - T(h) = O(h^{2k+1}) \quad (h \rightarrow 0).$$

Remarks: The trapezoidal rule is the best possible approximation, but the integrand f does not always satisfy the smoothness. $T(h)$ is used for the computation of Fourier coefficients.

SUPPLEMENTARY EXAMPLES – No. 7

EXTRAPOLATION FOR $h \rightarrow 0$ (RICHARDSON EXTRAPOLATION)**Example 1: Numerical differentiation**

Given $f(x) = \tan \frac{\pi}{2}x$, $L(h) := \frac{f(h) - f(0)}{h}$

Compute $f'(0) = \frac{\pi}{2} = 1.5707963\dots$ by extrapolation

$\frac{1}{2}$	2.000			
$\frac{1}{4}$	<u>1.656</u> ...	<u>1.54247</u> ...		
$\frac{1}{8}$	<u>1.591</u> ...	<u>1.56945</u> ...	<u>1.57125</u> ...	
$\frac{1}{16}$	<u>1.575</u> ...	<u>1.57072</u> ...	<u>1.57080</u> ...	<u>1.57079</u> ...

Example 2: Romberg algorithm

Compute $\int_1^2 \frac{dx}{x} = \ln 2 = 0.693147180\dots$

1	0.75			
$\frac{1}{2}$	0.708...	<u>0.694</u> ...		
$\frac{1}{4}$	<u>0.697</u> ...	<u>0.69325</u> ...	<u>0.69317</u> ...	
$\frac{1}{8}$	<u>0.694</u> ...	<u>0.69315</u> ...	<u>0.693148</u> ...	<u>0.69314747</u> ...

Example 3: Euler method with extrapolation

Consider the initial value problem $y' = y$, $y(0) = 1$

Compute $y(1) = e = 2.718\ 281$

Approximation $L^{(1)}(h) = (1 + h)^{1/h}$

Extrapolation $L^{(k+1)}(h) := \frac{2^k L^{(k)}(\frac{h}{2}) - L^{(k)}(h)}{2^k - 1}$, $k = 1, 2, \dots$

1	2				
$\frac{1}{2}$	2.25	2.5			
$\frac{1}{4}$	2.441...	2.632...	2.677		
$\frac{1}{8}$	2.565...	2.690...	<u>2.709</u> ...	<u>2.7138</u> ...	
$\frac{1}{16}$	2.637...	<u>2.710</u> ...	<u>2.7167</u> ...	<u>2.7177</u> ...	<u>2.71802</u> ...

Chapter II

Eigenvalue Problems for Matrices

7 Bounds for the Eigenvalues

Matrix $A = (a_{ij})_{i,j=1,\dots,n}$ in $K \in \{\mathbb{R}, \mathbb{C}\}$

Definition

$\lambda \in \mathbb{C}$ is called **eigenvalue** of A , if there exists $0 \neq x \in K^n$ satisfying $Ax = \lambda x$; such an x is called **eigenvector** of A for the eigenvalue λ .

Characteristic polynomial

$$\begin{aligned}\varphi_A(t) := \det(A - tI) &= (-1)^n (t - \lambda_1)^{\sigma_1} \dots (t - \lambda_k)^{\sigma_k} \\ &= (-1)^n \{t^n + \alpha_{n-1}t^{n-1} + \dots + \alpha_0\} \\ \lambda \text{ zero of } \varphi_A &\Leftrightarrow \lambda \text{ eigenvalue of } A\end{aligned}$$

Spectral radius $\rho(A) = \max_j |\lambda_j|$, $\rho(A) \leq N(A)$ for each matrix norm N

Rayleigh quotient $r_A(x) := \frac{x^H A x}{x^H x} : x \text{ eigenvector} \implies \lambda = r_A(x)$

Range $G[A] := \{r_A(x), x \in K^n, n \neq 0\}$ convex, contains all eigenvalues

Homogeneous linear system

$$(A - \lambda I)x = 0, \lambda \text{ eigenvalue, singular, solution eigenvector } x$$

$$\text{rank}(A - \lambda I) = n - \ell, \text{ i.e., } \ell \text{ free parameters}$$

Similarity transformation $T^{-1}AT =: B$, it holds $\varphi_A = \varphi_B$

Matrix A **symmetric** resp. **Hermitian**: the eigenvalues are real

Matrix U **orthogonal** resp. **unitary**: $\text{cond}_2(U) = 1$ (lub₂ norm)

$$U^T U = I, U^T = U^{-1} : \text{lub}_2(U) = N_\rho(U)$$

$$\text{Spectral norm } N_\rho(U) = \sqrt{\rho(U^T U)} = 1, N_\rho(U^{-1}) = N_\rho(U^T) = 1$$

Real matrix A positive definite

A symmetric and $x^T A x > 0$ f.a. $0 \neq x \in \mathbb{R}^n \Leftrightarrow$ the eigenvalues are positive

Matrix A normal: $A^H A = A A^H$ ($\Leftrightarrow$ unitarily similar to a diagonal matrix)

$\rightarrow$ Important theorems (see Stoer & Bulirsch)

Theorem 7.1 (*Theorem of Schur*)

For each $A \in K^{n \times n}$ exists an unitary U with

$$U^H A U = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}.$$

Theorem 7.2 Each Hermitian A is unitarily similar to a diagonal matrix, i.e., $U^H A U = \text{diag}(\lambda_1, \dots, \lambda_n)$ with unitary $U = (u_1, \dots, u_n)$.

The j -th column vector u_j is eigenvector of the eigenvalue $\lambda_j : A u_j = \lambda_j u_j$. Hence A has n linearly independent eigenvectors which are orthogonal.

Singular-values of A

Let A be an $(m \times n)$ -Matrix ($m \leq n$), then $A^H A$ is positive semidefinite. The eigenvalues $\lambda_j = \lambda_j(A^H A)$ are real and ≥ 0 .

The singular-values are $\sigma_j := \sqrt{\lambda_j(A^H A)}$ with $\sigma_1 \geq \dots \geq \sigma_n \geq 0$

$$m = n : \quad \sigma_1 = \sqrt{\rho(A^H A)} = \text{lub}_2(A) = \max_{0 \neq x} \frac{\|Ax\|_2}{\|x\|_2}$$

$$\sigma_n = \min_{0 \neq x} \frac{\|Ax\|_2}{\|x\|_2}$$

$$\text{cond}_2(A) = \text{lub}_2(A) \text{lub}_2(A^{-1}) = \frac{\sigma_1}{\sigma_n}.$$

Singular-value decomposition

$A = U \Sigma V^H$, U $m \times m$ and V $n \times n$ unitary

$$\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \quad D = \text{diag}(\sigma_1, \dots, \sigma_r), \quad \text{rank } A = r \quad (\sigma_r > 0)$$

Bounds for the eigenvalues of A ($n \times n$)-matrix

range $G[A]$: contains all the eigenvalues of A

A normal: $G[A] = \text{convex hull of the eigenvalues}$

A Hermitian: $\lambda_1 \leq \dots \leq \lambda_n$, $\lambda_1 = \min_{x \neq 0} r_A(x)$, $\lambda_n = \max_{x \neq 0} r_A(x)$.

Spectrum $\sigma[A] := \{\lambda_j(A), j = 1, \dots, n\}$

Gerschgorin circles

$$K_j := \{z \in \mathbb{C}, |z - a_{jj}| \leq \sum_{\substack{k=1 \\ k \neq j}}^n |a_{jk}|\}, \quad j = 1, \dots, n$$

Theorem 7.3 (*Theorem of Gerschgorin*)

The union of all Gerschgorin circles $\bigcup K_j$ contains all the eigenvalues of A .

Examples

$$A = \begin{pmatrix} 4 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & 4 \end{pmatrix} \quad \text{eigenvalues in } [2, 6], \text{ cond}_2(A) \leq 3$$

$$A = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix} \quad \text{eigenvalues in } [0, 4]$$

Condition of the eigenvalue problem $Ax = \lambda x$

Question: How does a perturbation of A effect λ ? Condition number K ?

$$\begin{array}{ll} Ax = \lambda x, & \text{EW } \lambda(A) \\ Bx = \lambda x, & \text{EW } \lambda(B) \end{array} \left\| \begin{array}{l} \text{perturbation } A + \Delta A =: B, \ N(\Delta A) \leq \varepsilon \\ \text{effect } |\lambda(A) - \lambda(B)| \leq K \cdot \varepsilon \end{array} \right.$$

$$\text{Linear system } Ax = b, \ A\tilde{x} = \tilde{b} \Rightarrow \frac{\|\tilde{x} - x\|}{\|x\|} \leq \text{cond}(A) \frac{\|\tilde{b} - b\|}{\|b\|}$$

Theorem 7.4 Let A be similar to a diagonal matrix $D = P^{-1}AP$ and B arbitrary. For each eigenvalue $\lambda(B)$ there exists an eigenvalue $\lambda(A)$ satisfying

$$|\lambda(B) - \lambda(A)| \leq \text{cond}_\nu(P) \text{lub}_\nu(B - A), \quad \nu = 1, 2, \infty.$$

Remarks:

The condition number depends on the matrix P .

If A is a diagonal matrix, then $P = I$ and $\text{cond}(I) = 1$.

If P is unitary matrix, then $\text{cond}_2(P) = \text{lub}_2(P) \text{lub}_2(P^H) = 1$.

A Hermitian matrix is unitarily similar to a diagonal matrix (Th.7.2).

Result 7.5 The eigenvalue problem of a Hermitian matrix is well conditioned.

Corollary 7.6 *Let λ be an eigenvalue of B , but not of A , then it holds*

$$1 \leq \text{lub} \left((\lambda I - A)^{-1}(B - A) \right) \leq \text{lub} \left((\lambda I - A)^{-1} \right) \text{lub} (B - A).$$

Proof: $(B - A)x = (\lambda I - A)x \implies \underbrace{(\lambda I - A)^{-1}}_{=:H} \underbrace{(B - A)}_{=:G} x = x$

$$x = H \cdot Gx : \|x\| = \|HGx\| \leq N(HG)\|x\| \leq N(H)N(G)\|x\|. \quad \square$$

Corollary 7.6 implies the proof of Theorem 7.3 (Gerschgorin) for the matrix B :

$$\begin{aligned} B &= (b_{ij}), \quad A := \text{diag}(b_{11}, \dots, b_{nn}) \\ 1 &\leq \text{lub}_\infty((\lambda I - A)^{-1}(B - A)) \leq \max_{1 \leq j \leq n} \frac{1}{|\lambda - b_{jj}|} \sum_{\substack{k=1 \\ k \neq j}}^n |b_{jk}| \quad \square \end{aligned}$$

Proof of Theorem 7.4:

Assumptions

$$A = PDP^{-1}, \quad D = \text{diag}(\lambda_1(A), \dots, \lambda_n(A))$$

$\lambda = \lambda(B)$ eigenvalue of B , but not of A (else the statement is trivial)

$$(\lambda I - A)^{-1} = P(\lambda I - D)^{-1}P^{-1}$$

$\lambda \neq \lambda_j(A)$ for all j : estimation of $\min_j |\lambda - \lambda_j(A)|$

$\lambda - \lambda_j(A)$ is eigenvalue of $\lambda I - D$

$\frac{1}{\lambda - \lambda_j(A)}$ is eigenvalue of $(\lambda I - D)^{-1}$

$$\max_j \frac{1}{|\lambda - \lambda_j(A)|} = \text{lub}_\nu \left(\underbrace{(\lambda I - D)^{-1}}_{\text{diagonal matrix}} \right), \quad \nu = 1, 2, \infty$$

$\uparrow$
 does not hold
 for all norms

$$\begin{aligned} \text{lub}((\lambda I - A)^{-1}) &\leq \text{lub}(P) \text{lub}((\lambda I - D)^{-1}) \text{lub}(P^{-1}) \\ &= \text{cond}(P) \text{lub}((\lambda I - D)^{-1}) \end{aligned}$$

$$\text{Corollary 7.6} \Rightarrow 1 \leq \text{cond}_\nu(P) \max_j \frac{1}{|\lambda - \lambda_j(A)|} \text{lub}_\nu(B - A)$$

hence $\min_j |\lambda - \lambda_j(A)| \leq \text{cond}_\nu(P) \text{lub}_\nu(B - A). \quad \square$

Similar matrices by suitable transformations

Given: eigenvalue problem $Ax = \lambda x$ with condition number $\text{cond}(P)$

Wanted: equivalent problem $\tilde{A}x = \lambda x$ with an easier matrix

by transforming matrix T with $\tilde{A} = T^{-1}AT$.

Question: How does T effect the condition of the eigenvalue problem?

Theorem 7.7 *Let A be similar to a diagonal matrix $D = P^{-1}AP$. Under the transformation $T^{-1}AT$ the condition of the eigenvalue problem gets worse at most by the factor $\text{cond}(T)$.*

Proof : $A = PDP^{-1}$

$$\begin{aligned}\tilde{A} &= T^{-1}AT = T^{-1}PDP^{-1}T \\ &= \underbrace{(T^{-1}P)}_{=: \tilde{P}} D (T^{-1}P)^{-1}\end{aligned}$$

condition of $\tilde{A}x = \lambda x$ (Theorem 7.4): $\text{cond}(\tilde{P}) \leq \text{cond}(P)\text{cond}(T)$. $\square$

Reduction of A to an easier form

$$\begin{aligned}A &=: A^{(1)} \longrightarrow A^{(2)} \longrightarrow \dots \longrightarrow A^{(m)} =: \tilde{A}, \quad A^{(i+1)} := T_i^{-1}A^{(i)}T_i \\ \tilde{A} &= T_{m-1}^{-1} \dots T_1^{-1}A \underbrace{T_1 \dots T_{m-1}}_{=: T} = T^{-1}AT\end{aligned}$$

Under a series of similarity transformations $T := T_1 \dots T_{m-1}$ the condition gets worse at most by the factor

$$\text{cond}(T) \leq \text{cond}(T_1) \dots \text{cond}(T_{m-1}).$$

Unitary transformation matrices $T_1, \dots, T_{m-1}$:

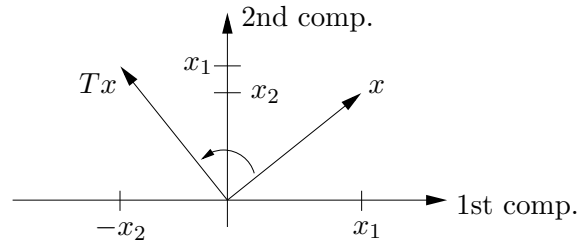
$$\text{cond}_2(T) \leq \text{cond}_2(T_1) \dots \text{cond}_2(T_{m-1}) = 1$$

Examples: Rotation matrices orthogonal, $\text{cond}_2(T_{ij}) = 1$, $\text{cond}_\infty(T_{ij}) \leq 2$

$$T_{ij}(\alpha) := \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \boxed{\cos \alpha} & \dots & \boxed{-\sin \alpha} & \dots \\ & & & 1 & & & \\ & & & & \ddots & & \\ & & & & & 1 & \\ & & \dots & \boxed{\sin \alpha} & \dots & \boxed{\cos \alpha} & \dots \\ & & & & & & 1 \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \begin{matrix} \\ \\ \\ \text{i-th row} \\ \\ \\ \text{j-th row} \\ \\ \text{i-th} \quad \text{j-th} \quad \text{column} \end{matrix}$$

Transformation: $x \mapsto T(\alpha)x$

$$n = 2 : \quad \text{e.g.,} \quad \alpha = \frac{\pi}{2} : T\left(\frac{\pi}{2}\right) = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto Tx = \begin{pmatrix} -x_2 \\ x_1 \end{pmatrix}$$

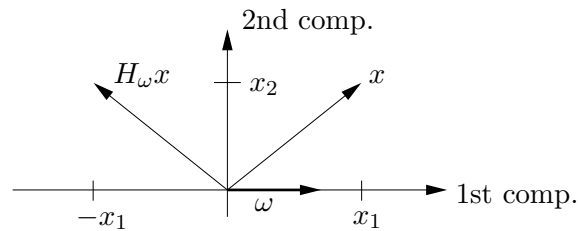
Figure: Rotation of the vector x by the angle $\frac{\pi}{2}$

Householder matrices: Hermitian and unitary

$$\begin{aligned}
 H_w &= I - 2ww^H \text{ for } w \in K^n \text{ with } w^H w = 1 \\
 (H_w)^H &= H_w \text{ Hermitian} \\
 H_w^H H_w &= (I - 2ww^H)(I - 2ww^H) = I - 4ww^H + 4w \underbrace{w^H w}_{=1} w^H = I
 \end{aligned}$$

Transformation: $x \mapsto H_w x$

$$n = 2 : \text{e.g. } w = \begin{pmatrix} 1 \\ 0 \end{pmatrix} : H_w = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}, x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto H_w x = \begin{pmatrix} -x_1 \\ x_2 \end{pmatrix}$$

Figure: Reflection at the orthogonal hyperplane of w

Reduction of matrices

In general the *Jordan normal form* resp. the diagonal resp. triangular form can not be reached in a finite number of steps (because the eigenvalues are not rational also in the case of entire elements of the matrix)!

Iterative approach to the diagonal resp. triangular form is possible.

In a finite number of steps it is possible to reach

$$\text{Hessenberg form} \quad \begin{pmatrix} * & * & \dots & * \\ * & \ddots & \ddots & \vdots \\ & \ddots & \ddots & * \\ 0 & & * & * \end{pmatrix}$$

$$\text{or symmetric tridiagonal form} \quad \begin{pmatrix} * & * & \dots & 0 \\ * & \ddots & \ddots & \vdots \\ & \ddots & \ddots & * \\ 0 & & * & * \end{pmatrix} \quad (\text{if } A \text{ symmetric})$$

Computation of the eigenvalues of a matrix

- In principle:* evaluation of the characteristic polynomial,
then methods for computing zeros of polynomials
- Direct methods:* reduction of the matrix to an easier form,
characteristic polynomial by recurrence formulas,
then special methods for computing the zeros
- Iterative methods:* construction of a series of matrices $\{A^{(i)}\}_{i \geq 0}$
by transformations $T_i^{-1}A^{(i)}T_i$
with $A^{(i)} \rightarrow \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$ if $i \rightarrow \infty$
- Vector iteration:* Raising to a higher power $A^k x = \lambda^k x$
("distilling" the largest eigenvalues in absolute value!)

8 Eigenvalues of Symmetric Matrices

$$\text{Real matrix } A_n = \begin{pmatrix} a_{11} & a_{12} & & 0 \\ a_{12} & a_{22} & \ddots & \\ & \ddots & \ddots & a_{n-1,n} \\ 0 & & a_{n-1,n} & a_{nn} \end{pmatrix} \quad \text{eigenvalues real}$$

A_n is called *irreducible*, if $a_{j-1,j} \neq 0$, $j = 2, \dots, n$ and *reducible* otherwise.

If A_n is irreducible, then the eigenvalues are *real* and *simple*.

If A_n is reducible, then reduction of the problem to

$$A = \left(\begin{array}{c|c} A_k & 0 \\ \hline 0 & B_{n-k} \end{array} \right),$$

i.e., the characteristic polynomial satisfies $\varphi_{A_n} = \varphi_{A_k} \cdot \varphi_{B_{n-k}}$.

Main matrices

$$A_k := \begin{pmatrix} a_{11} & a_{12} & & 0 \\ a_{12} & \ddots & \ddots & \\ & \ddots & \ddots & a_{k-1,k} \\ 0 & & a_{k-1,k} & a_{kk} \end{pmatrix}, \quad k = 1, \dots, n$$

characteristic polynomial $\varphi_k(t) = \det(A_k - tI)$, degree $\varphi_k = k$

Theorem 8.1 (*3-term recurrence relation*)

The characteristic polynomial φ_n of a symmetric tridiagonal matrix A_n is deduced by the recurrence formula

$$\begin{aligned} \varphi_0(t) &:= 1, \quad \varphi_1(t) := a_{11} - t \\ \varphi_k(t) &:= (a_{kk} - t) \varphi_{k-1}(t) - a_{k-1,k}^2 \varphi_{k-2}(t), \quad k = 2, 3, \dots, n. \end{aligned}$$

Proof: Evaluation of the determinant by the last column. □

Theorem 8.2 (*Zeros*)

Let A_n be an irreducible symmetric tridiagonal matrix.

The characteristic polynomial φ_k of A_k ($1 \leq k \leq n$) has k real and simple zeros $\lambda_1^{(k)} < \dots < \lambda_k^{(k)}$.

The zeros of φ_{k+1} are separated by the zeros of φ_k , i.e., $\lambda_\nu^{(k+1)} < \lambda_\nu^{(k)} < \lambda_{\nu+1}^{(k+1)}$ for $1 \leq \nu \leq k$.

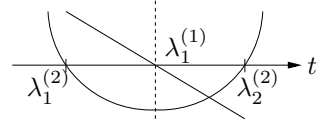
Proof: The zeros of φ_k are real because of the symmetry of A_k .

Show the property of separation by induction:

Begin $k = 1$: $\varphi_1(t) = a_{11} - t$, zero $\lambda_1^{(1)} = a_{11}$

$$\begin{aligned}\varphi_2(t) &= (a_{22} - t) \varphi_1(t) - a_{12}^2 \varphi_0(t) \\ &= t^2 - (a_{11} + a_{22})t + a_{11}a_{22} - a_{12}^2\end{aligned}$$

$$\left. \begin{aligned}\varphi_2(\lambda_1^{(1)}) &= -a_{12}^2 < 0 \\ \varphi_2(t) &\rightarrow +\infty \text{ for } t \rightarrow \pm\infty\end{aligned} \right\} \text{ i.e., } \lambda_1^{(2)} < \lambda_1^{(1)} < \lambda_2^{(2)}$$

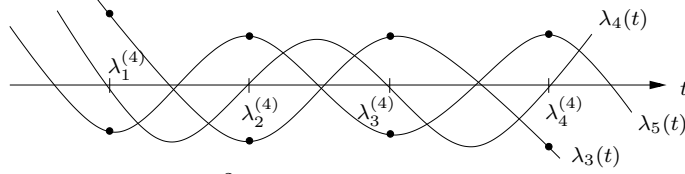


Assumption: Statement holds for $k - 1$

$$\lambda_1^{(k)} < \lambda_1^{(k-1)} < \lambda_2^{(k)} < \dots < \lambda_{k-1}^{(k)} < \lambda_{k-1}^{(k-1)} < \lambda_k^{(k)}$$

Conclusion to k :

$k = 4$:



$$\begin{aligned}\varphi_{k+1}(t) &= (a_{k+1,k+1} - t)\varphi_k(t) - a_{k,k+1}^2 \varphi_{k-1}(t) \\ &= (-1)^{k+1} t^{k+1} + \dots\end{aligned}$$

$$\varphi_{k+1}(\lambda_\nu^{(k)}) = -a_{k,k+1}^2 \varphi_{k-1}(\lambda_\nu^{(k)}), \quad \nu = 1, \dots, k$$

$$\text{i.e., } \boxed{\varphi_{k+1}(\lambda_\nu^{(k)}) \varphi_{k-1}(\lambda_\nu^{(k)}) < 0} \quad (\text{the 4-th axiom of the chain of Sturm})$$

$\varphi_{k-1}(t)$ has exactly one zero in $(\lambda_\nu^{(k)}, \lambda_{\nu+1}^{(k)})$ ($\nu = 1, \dots, k-1$) (assumption)

i.e., $\varphi_{k+1}(t)$ has one change of sign in $(\lambda_\nu^{(k)}, \lambda_{\nu+1}^{(k)})$ ($\nu = 1, \dots, k-1$)

i.e., $\varphi_{k+1}(t)$ has one zero in $(\lambda_\nu^{(k)}, \lambda_{\nu+1}^{(k)})$ ($\nu = 1, \dots, k-1$)

i.e., $\varphi_{k+1}(t)$ has at least $k-1$ zeros in $(\lambda_1^{(k)}, \lambda_k^{(k)})$

Subinterval $(-\infty, \lambda_1^{(k)})$:

$$\left. \begin{aligned}\varphi_{k+1}(t) \text{ and } \varphi_{k-1}(t) &\rightarrow +\infty, \quad t \rightarrow -\infty \\ \varphi_{k+1}(\lambda_1^{(k)}) &< 0\end{aligned} \right\} \rightarrow \varphi_{k+1}(t) \text{ has one zero}$$

Subinterval $(\lambda_k^{(k)}, \infty)$:

$$\left. \begin{aligned}\varphi_{k+1}(t) \text{ and } \varphi_{k-1}(t) &\rightarrow (-1)^{k+1} \cdot \infty, \quad t \rightarrow +\infty \\ \text{sign } \varphi_{k+1}(\lambda_k^{(k)}) &= -\text{sign } \varphi_{k-1}(\lambda_k^{(k)}) = (-1)^k\end{aligned} \right\} \rightarrow \varphi_{k+1}(t) \text{ one zero}$$

Hence: $\varphi_{k+1}(t)$ has $k+1$ simple zeros, separated by the zeros of $\varphi_k(t)$. $\square$

Remark: Compare this result with the zeros of orthogonal polynomials.

Example

$$\begin{pmatrix} 0 & \frac{1}{\sqrt{2}} & 0 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \ddots \\ & & & \ddots & \ddots \end{pmatrix} \quad \begin{aligned} &\varphi_0(t) = 1, \varphi_1(t) = -t, \varphi_2(t) = t^2 - \frac{1}{2}, \\ &\varphi_3(t) = -(t^3 - \frac{3}{4}t), \varphi_4(t) = t^4 - t^2 + \frac{1}{8}, \dots \\ &\rightarrow \text{chain of Sturm} \\ &\rightarrow \text{Tschebyscheff polynomials (notice the sign)} \\ &T_0 = 1, T_1 = t, T_2 = 2t^2 - 1, T_3 = 4t^3 - 3t, \\ &T_4 = 8t^4 - 8t^2 + 1, \dots \end{aligned}$$

Theorem 8.3 (*Chain of Sturm*)

Let A_n be an irreducible symmetric tridiagonal matrix.

The characteristic polynomials φ_k of A_k , $k = 1, \dots, n$ (with alternating signs)

$$(-1)^n \varphi_n, (-1)^{n-1} \varphi_{n-1}, \dots, (-1) \varphi_1, \varphi_0$$

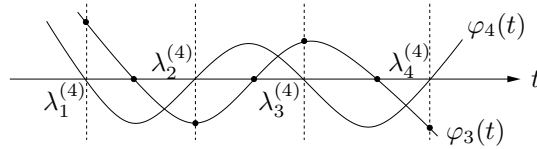
form a chain of Sturm in $[\alpha, \beta]$ with $\varphi_n(\alpha)\varphi_n(\beta) \neq 0$.

Proof: Check the axioms of the chain of Sturm (see Numer. Math. I)

$$(1) \quad \varphi_n(\alpha)\varphi_n(\beta) \neq 0$$

$$(2) \quad \varphi_0(t) \neq 0$$

$$(3) \quad \varphi_n(t) \text{ only simple zeros } \xi \\ \text{and } \varphi'_n(\xi)\varphi_{n-1}(\xi) < 0$$



$$(4) \quad \varphi_k(\eta) = 0 \text{ then} \\ \varphi_{k+1}(\eta)\varphi_{k-1}(\eta) < 0$$

Number of sign-changes $W(t) := W((-1)^n \varphi_n(t), \dots, \varphi_0(t))$

Theorem of Sturm (see Numer. Math. I):

The number of zeros in (α, β) satisfies $Z_\alpha^\beta(\varphi_n) = W(\alpha) - W(\beta)$.

Bisection method: Determine an interval $[\alpha, \beta]$ of length ε with at least (or exactly) one zero

Newton method with a suitable initial value

Derivative by recurrence $\varphi'_0(t) = 0, \varphi'_1(t) = -1$

$$\varphi'_k(t) = -\varphi_{k-1}(t) + (a_{kk} - t)\varphi'_{k-1}(t) - a_{k-1,k}^2 \varphi'_{k-2}(t)$$

SUPPLEMENTARY EXAMPLES – NO. 8

EIGENVALUES OF SYMMETRIC TRIDIAGONAL MATRICES

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \rightarrow \text{Gerschgorin: eigenvalues in } [1, 6]$$

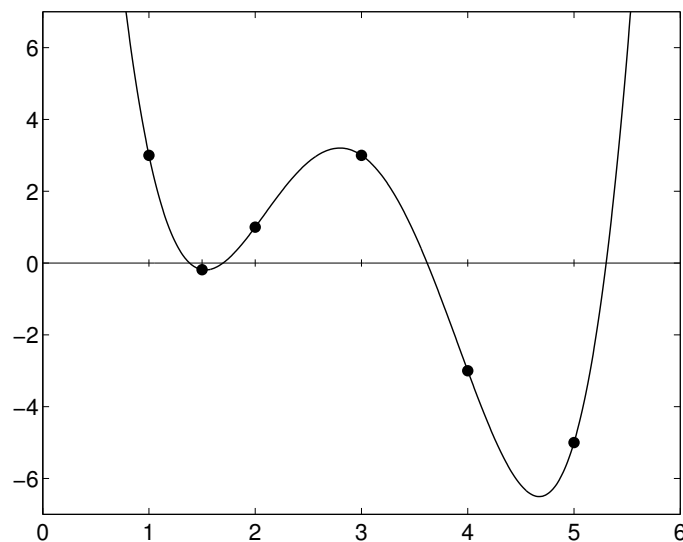
Chain of Sturm: $\varphi_0, -\varphi_1, \varphi_2, -\varphi_3, \varphi_4$

$$\begin{aligned} \varphi_0(t) &= 1 \\ \varphi_1(t) &= 2 - t \\ \varphi_2(t) &= (4 - t)\varphi_1(t) - \varphi_0(t) \\ \varphi_3(t) &= (4 - t)\varphi_2(t) - \varphi_1(t) \\ \varphi_4(t) &= (2 - t)\varphi_3(t) - \varphi_2(t) \end{aligned}$$

t	$\varphi_0(t)$	$-\varphi_1(t)$	$\varphi_2(t)$	$-\varphi_3(t)$	$\varphi_4(t)$	$W(t)$
1	1	-1	2	-5	3	4
1.5	1	-0.5	0.25	-0.125	-0.1875	3
2	1	0	-1	2	1	2
3	1	1	-2	1	3	2
4	1	2	-1	-2	-3	1
5	1	3	2	-1	-5	1
6	1	4	7	10	33	0

Characteristic polynomial $y = \varphi_4(t)$:

Exactly one zero in each of the intervals $[1, 1.5]$, $[1.5, 2]$, $[3, 4]$ and $[5, 6]$.



Hessenberg matrix: *Method of Hyman*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ & \ddots & \ddots & \vdots \\ 0 & & a_{n,n-1} & a_{nn} \end{pmatrix} \quad \text{irreducible } (a_{j,j-1} \neq 0)$$

characteristic polynomial $\varphi_A(t) = \det(A - tI)$

Newton method (for real eigenvalues):

Computation of $\varphi_A(t)$ and $\varphi'_A(t)$ without having the explicit form of the polynomial.

initial values by Gerschgorin

The alternative problem

the linear system $(A - \mu I)x = q(\mu)e_1$ $x = (x_1, \dots, x_n)^T$
 μ fixed and $x_n = 1$ $e_1 = (1, 0, \dots, 0)^T$

Determine numbers $x_1, \dots, x_{n-1}$ and $q(\mu)$ such that equality holds.

Recurrence (see below)

$$x_{n-1} = -\frac{a_{nn} - \mu}{a_{n,n-1}}, \quad x_{n-2} = -\frac{1}{a_{n-1,n-2}}(\dots), \quad \dots, \quad x_1 = -\frac{1}{a_{21}}(\dots)$$

the first row reads $(a_{11} - \mu)x_1 + \dots + a_{1n}x_n = q(\mu)$

and implies the unique solution for $q(\mu)$

Cramer's rule for $x_n = 1$:

$$1 = x_n = \frac{\det(\dots, q(\mu)e_1)}{\varphi_A(\mu)} \quad \text{i.e., } \varphi_A(\mu) = \det(\dots, q(\mu)e_1)$$

$$\det \begin{pmatrix} a_{11} - \mu & a_{12} & \cdots & a_{1,n-1} & q(\mu) \\ a_{21} & a_{22} - \mu & \cdots & a_{2,n-1} & 0 \\ \vdots & \ddots & & \vdots & \vdots \\ 0 & & & a_{n,n-1} & 0 \end{pmatrix}$$

$$= (-1)^{n-1} q(\mu) \cdot \det \begin{pmatrix} a_{21} & \cdots & a_{2,n-1} \\ & \ddots & \vdots \\ 0 & & a_{n,n-1} \end{pmatrix}$$

$$= (-1)^{n-1} q(\mu) \cdot \prod_{j=2}^n a_{j,j-1} \neq 0$$

Result: $\varphi_A(\mu) = \text{constant} \cdot q(\mu)$

i.e., λ zero of $q(\mu)$ if and only if λ eigenvalue of A .

$q(\mu)$ is implicitly defined by the linear system

$$\left. \begin{array}{rcl} -q(\mu) + (a_{11} - \mu)x_1 + a_{12}x_2 + \dots & + a_{1,n-1} x_{n-1} = & -a_{1n} \\ a_{21} x_1 + (a_{22} - \mu)x_2 + \dots & + a_{2,n-1} x_{n-1} = & -a_{2n} \\ & \vdots & \vdots \\ & a_{n,n-1} x_{n-1} = & -(a_{nn} - \mu) \end{array} \right\}$$

$q(\mu)$ will be computed by recurrence (from below)

Derivative: $x'_j = x'_j(\mu)$

$$\left. \begin{array}{rcl} -q'(\mu) + (a_{11} - \mu)x'_1 + a_{12}x'_2 + \dots & + a_{1,n-1} x'_{n-1} = & x_1 \\ a_{21} x'_1 + (a_{22} - \mu)x'_2 + \dots & + a_{2,n-1} x'_{n-1} = & x_2 \\ & \vdots & \vdots \\ & a_{n,n-1} x'_{n-1} = & x_n = 1 \end{array} \right\}$$

$q'(\mu)$ will be computed by recurrence (from below)

Newton method $\mu_{k+1} := \mu_k - \frac{q(\mu_k)}{q'(\mu_k)}, \quad k = 0, 1, \dots$

9 Reduction Method of Householder

Real symmetric (or Hermitian) matrix A : eigenvalue λ , eigenvector x

Similar matrix $\tilde{A} = T^{-1}AT$: eigenvalue λ , eigenvector $T^{-1}x$

$$A = \begin{pmatrix} a_{11} & \dots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{pmatrix} \begin{array}{c} \text{similar matrix} \\ \longrightarrow \\ \text{orthogonal } T \end{array} \tilde{A} = \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{22} & & 0 \\ \tilde{a}_{21} & \tilde{a}_{22} & \ddots & \\ 0 & & \ddots & \\ & & & \ddots \end{pmatrix}$$

$$A =: A^{(1)} \rightarrow \left(\begin{array}{c|ccc} a_{11} & a_{21}^{(2)} & 0 & \dots & 0 \\ \hline a_{21}^{(2)} & a_{22}^{(2)} & \dots & & a_{n2}^{(2)} \\ 0 & a_{32}^{(2)} & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & a_{n2}^{(2)} & \dots & & a_{nn}^{(2)} \end{array} \right) =: A^{(2)} \rightarrow \dots \rightarrow A^{(n-1)} =: \tilde{A}$$

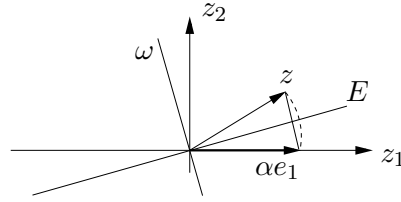
Transformation by the orthogonal matrix T_i : $T_i^T A^{(i)} T_i =: A^{(i+1)}$

$$T_1^T A^{(1)} T_1 = \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & U \end{array} \right) \left(\begin{array}{c|c} a_{11} & a_1'^T \\ \hline a_1' & A' \end{array} \right) \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & U^T \end{array} \right) = \left(\begin{array}{c|c} a_{11} & (U a_1')^T \\ \hline U a_1' & U A' U^T \end{array} \right)$$

$$U \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} = \alpha \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \boxed{U : z \mapsto \alpha e_1}$$

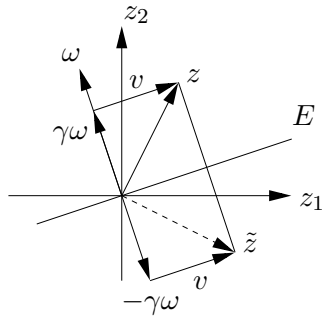
Householder matrices: Reflection

$$\boxed{H_w := I - 2ww^H, \quad w^H w = 1}$$



Householder transformation $H_w : z \mapsto H_w z$

corresponds to a reflection at the hyperplane E orthogonal to w .



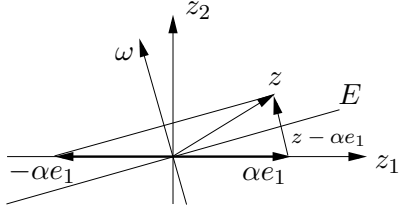
$$z = \gamma w + v, \quad \gamma = w^H z \in K \quad (\text{orthogonal projection})$$

$$\tilde{z} = -\gamma w + v \quad (\text{reflected})$$

$$H_w z = (I - 2ww^H)z = z - 2w \underbrace{w^H z}_{=\gamma}$$

$$= \gamma w + v - 2\gamma w = \tilde{z}$$

The choice of w for given $z \neq \beta e_1 \in \mathbb{R}^n$:



2 possibilities (reflection to the right or to the left side)

$|\alpha| = \|z\|_2$ (the Euclidean norm is invariant)

w is a multiple of $z - \alpha e_1$,

because $H_w z = z - 2ww^T z \stackrel{!}{=} \alpha e_1$,

hence $z - \alpha e_1 = 2 \underbrace{(w^T z)}_w w$

Hence $w = \frac{z - \alpha e_1}{\|z - \alpha e_1\|_2}$ (normed to $\|w\|_2 = 1$)

Sign of α : $\alpha = \pm \|z\|_2$

Choose the sign such that $\|z - \alpha e_1\|_2$ is as large as possible (“division”)

$$\|z - \alpha e_1\|_2^2 = (z_1 - \alpha)^2 + z_2^2 + \cdots + z_n^2 \rightarrow \alpha = -\text{sign}(z_1) \|z\|_2.$$

Theorem 9.1 (Householder transformation)

The vector $0 \neq z \in \mathbb{R}^n$ with first component $z_1 = \text{sign}(z_1)|z_1|$ will be reflected to a multiple of the first unit vector by the Householder transformation $H_w : z \mapsto H_w z$ with

$$w = \frac{z - \alpha e_1}{\|z - \alpha e_1\|_2}, \quad \alpha = -\text{sign}(z_1) \|z\|_2.$$

If $z_1 = 0$, then choose the sign arbitrarily $\alpha = \pm \|z\|_2$.

Problem: For $z = (4, 3, 0)^T$ discuss the result of Theorem 9.1.

$\|z\|_2 = 5$, $z_1 > 0$ implies $\alpha = -5$ and $H_w : z \rightarrow (-5, 0, 0)^T$.

Then $z - \alpha e_1 = z + 5e_1 = (9, 3, 0)^T$ and $\|z - \alpha e_1\|_2 = \sqrt{90}$.

It follows $w = \frac{z - \alpha e_1}{\|z - \alpha e_1\|_2} = \frac{\sqrt{90}}{90}(9, 3, 0)^T$ and $\|w\|_2 = 1$,

$$\text{and further, } H_w = I - 2ww^T = \frac{1}{5} \begin{pmatrix} -4 & -3 & 0 \\ -3 & 4 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

Please, check the orthogonality and the symmetry.

Householder reduction for the eigenvalue problem $Ax = \lambda x$

A real and symmetric (respectively Hermitian)

$$\begin{aligned} A &= A^{(1)} \rightarrow T_1 A^{(1)} T_1 = A^{(2)} \rightarrow T_2 A^{(2)} T_2 = A^{(3)} \rightarrow \dots \rightarrow A^{(n-1)} \\ A^{(n-1)} &= T_{n-2} \dots T_1 A T_1 \dots T_{n-2} \end{aligned}$$

Transformation matrices: $T_1, \dots, T_{n-2}$ symmetric and orthogonal

$$T_1 := \left(\begin{array}{c|c} 1 & 0 \\ \hline 0 & H_1 \end{array} \right), \quad H_1 := H_{w_1} \text{ corresponding to } z_1 = \begin{pmatrix} a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}$$

$$A^{(2)} = \begin{pmatrix} a_{11} & a_{21}^{(2)} & 0 & \dots & 0 \\ \hline a_{21}^{(2)} & a_{22}^{(2)} & \dots & a_{n2}^{(2)} \\ 0 & a_{32}^{(2)} & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix} \quad \text{symmetric}$$

$$T_2 := \left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & \\ \hline 0 & & H_2 \end{array} \right), \quad H_2 := H_{w_2} \longrightarrow z_2 = \begin{pmatrix} a_{32}^{(2)} \\ \vdots \\ \vdots \\ a_{n2}^{(2)} \end{pmatrix}$$

$$A^{(3)} = \begin{pmatrix} a_{11}^{(1)} & a_{21}^{(2)} & 0 & 0 & \dots & 0 \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{32}^{(3)} & 0 & \dots & 0 \\ 0 & a_{32}^{(3)} & a_{33}^{(3)} & a_{43}^{(3)} & \dots & a_{n3}^{(3)} \\ 0 & 0 & a_{43}^{(3)} & * & * & * \\ \vdots & \vdots & \vdots & * & * & * \\ 0 & 0 & a_{n3}^{(3)} & * & * & * \end{pmatrix} \quad \text{symmetric}$$

$$\vdots \\ T_{k-1} := \left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & H_{k-1} \end{array} \right), \quad H_{k-1} := H_{w_{k-1}} \longrightarrow z_{k-1} = \begin{pmatrix} a_{k,k-1}^{(k-1)} \\ \vdots \\ \vdots \\ a_{n,k-1}^{(k-1)} \end{pmatrix}$$

$$A^{(k)} := \left(\begin{array}{ccc|cc} \ddots & \ddots & \ddots & 0 & 0 & 0 \\ 0 & & & & a_{k,k-1}^{(k)} & 0 \\ \hline 0 & & a_{k,k-1}^{(k)} & & a_{k,k}^{(k)} & z_k^T \\ 0 & & 0 & & z_k & * \end{array} \right)$$

$$\vdots \\ A^{(n-1)} = \left(\begin{array}{ccc|cc} * & * & & 0 & & \\ * & \ddots & \ddots & & & \\ & \ddots & \ddots & & * & \\ 0 & & * & * & & \end{array} \right)$$

Chain of n-2 similarity transformations

Algorithmic implementation (see Stoer & Bulirsch): Use the symmetry!

Natural stability (condition:) A symmetric, unitary transformation

Numerical stability: Effect of the *rounding errors*

$$A =: A^{(1)} \xrightarrow{\tilde{T}_1} \tilde{A}^{(2)} \xrightarrow{\tilde{T}_2} \tilde{A}^{(3)} \longrightarrow \dots \xrightarrow{\tilde{T}_{n-2}} \tilde{A}^{(n-1)} =: \hat{A} \text{ instead of } A^{(n-1)}$$

Backward analysis: Reducing to perturbations of A

$$A + F \xrightarrow[\text{transformation}]{\text{without rounding errors}} \hat{A}$$

Floating-point arithmetic

$$\text{lub}_2(\tilde{T}_i - T_i) \leq g(n) \cdot \text{eps} \quad (g(n) = O(n^\alpha), n \rightarrow \infty, \alpha \approx 1)$$

$$\text{lub}_2(F) \leq K(n) \cdot \text{eps} \cdot \text{lub}_2(A) \quad (\text{in the first approximation})$$

$$K(n) = O(n^2), \quad \text{lub}_2(A) = \rho(A) \quad (A \text{ symmetric})$$

$$\left. \begin{array}{l} A \xrightarrow{T} T^{-1}AT =: \tilde{A} \\ A + F \longrightarrow T^{-1}AT + T^{-1}FT =: \hat{A}, \text{ perturbation } F \end{array} \right\} \text{ effect}$$

$$\text{lub}_2(\hat{A} - \tilde{A}) = \text{lub}_2(T^{-1}FT) \leq \underbrace{\text{cond}_2(T)}_{=1 \text{ (} T \text{ unitary)}} \text{lub}_2(F)$$

Result 9.2 *The Householder reduction method is numerically stable. More precise it holds:*

$$\boxed{\text{lub}_2(\hat{A} - A^{(n-1)}) \leq K(n) \cdot \text{lub}_2(A) \cdot \text{eps}, \quad K(n) = O(n^2)}$$

Remark: If A is not symmetric, then the Householder reduction leads to the Hessenberg form.

Matrix decompositions (see Numer. Math. I)

$$LR \text{ decomposition } A = LR = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix} \begin{pmatrix} * & \dots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix},$$

diagonal choice of the pivot elements in the Gaussian elimination.

$$Cholesky \text{ decomposition } A = LL^H, L = \begin{pmatrix} * & & 0 \\ \vdots & \ddots & \\ * & \dots & * \end{pmatrix}, \quad A \text{ positive definite.}$$

$$QR \text{ decomposition } A = QR = \begin{pmatrix} * & \dots & * \\ \vdots & & \vdots \\ * & \dots & * \end{pmatrix} \begin{pmatrix} * & \dots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix}, \quad Q \text{ unitary.}$$

Theorem 9.3 *The quadratic matrix A can be decomposed into $A = QR$.*

Proof: Householder transformation (Theorem 9.1)

$$A =: A^{(1)} \rightarrow T_1 A^{(1)} =: A^{(2)} \rightarrow T_2 A^{(2)} =: A^{(3)} \rightarrow \dots \rightarrow A^{(n)} =: R$$

Transformation matrices $T_1, \dots, T_{n-1}$ Hermitian and unitary

$$T_j := \left(\frac{I_{j-1} \mid 0}{0 \mid H_{w_j}} \right), \quad H_{w_j} \text{ corresponding to } z_j := \begin{pmatrix} a_{jj}^{(j)} \\ \vdots \\ a_{nj}^{(j)} \end{pmatrix}.$$

If $z_j = 0$, then $T_j = I$.

Altogether: $T_{n-1} T_{n-2} \dots T_1 A = R$, i.e., $A = \underbrace{T_1 \dots T_{n-1}}_{=: Q \text{ unitary}} \cdot R$. $\square$

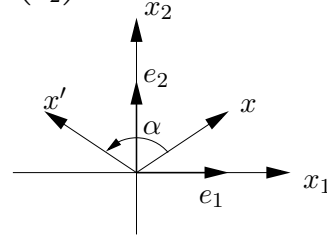
Remarks: QR decomposition by Householder is numerically stable.
 QR decomposition for symmetric tridiagonal matrix (see below).
Reduction of a linear system into the triangular form by Householder.

10 Methods of Givens and Jacobi

Rotation matrix $T := \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}$, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ real

$$T : x \mapsto Tx =: x' = \begin{pmatrix} x_1 \cos \alpha - x_2 \sin \alpha \\ x_1 \sin \alpha + x_2 \cos \alpha \end{pmatrix}$$

Rotation by the angle α with $\|x\|_2 = \|x'\|_2$



Rotation in e_1 - resp. e_2 -direction ($x_1 \neq 0$, $x_2 \neq 0$)

$$\begin{aligned} x_1 \sin \alpha + x_2 \cos \alpha &= 0, \quad \text{i.e.,} \quad \cos \alpha = \pm \frac{|x_1|}{\|x\|_2}, \quad \sin \alpha = -\frac{x_2}{x_1} \cos \alpha \\ x_1 \cos \alpha - x_2 \sin \alpha &= 0, \quad \text{i.e.,} \quad \cos \alpha = \pm \frac{|x_2|}{\|x\|_2}, \quad \sin \alpha = +\frac{x_1}{x_2} \cos \alpha \end{aligned}$$

Since $\|x\|_2^2 = x_1^2 + x_2^2$;

$x'_2 = 0$ then $x_1 \cos \alpha - x_2 \sin \alpha = \pm \|x\|_2$;

from these two equations $\cos \alpha$ and $\sin \alpha$ follow.

Rotation matrix T_{ij} ($i < j$) orthogonal

$$T_{ij}(\alpha) := \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \cos \alpha & & & -\sin \alpha \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & \sin \alpha & & & \cos \alpha \\ & & & & & & 1 \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} \begin{matrix} \leftarrow i \\ \\ \\ \leftarrow j \end{matrix}$$

$$\begin{matrix} \uparrow & \uparrow \\ i & j \end{matrix}$$

$x \in \mathbb{R}^n : T_{ij} : x \mapsto x'$ Rotation of x by the angle α in the plane spanned by the vectors e_i and e_j

Transformation: $A' = T_{ij}^T A T_{ij}$ and A are similar

$$A \longrightarrow T_{ij}^T A \longrightarrow T_{ij}^T A T_{ij} = A'$$

Changes, first in the i -th and j -th row, second in the i -th and j -th column.

A symmetric then A' is symmetric!

Special choices of α deliver that $a'_{kl} = 0 \longrightarrow$ series of transformations!

Transformation formulas (A real and symmetric)

$$\begin{aligned}
a'_{\nu\mu} &= a_{\nu\mu}, & \nu &\neq i, j, \mu \neq i, j \\
a'_{i\mu} &= a'_{\mu i} = a_{i\mu} \cos \alpha + a_{j\mu} \sin \alpha, & \mu &\neq i, j \\
a'_{j\mu} &= a'_{\mu j} = -a_{i\mu} \sin \alpha + a_{j\mu} \cos \alpha, & \mu &\neq i, j \\
a'_{ii} &= a_{ii} \cos^2 \alpha + 2a_{ij} \sin \alpha \cos \alpha + a_{jj} \sin^2 \alpha \\
a'_{jj} &= a_{ii} \sin^2 \alpha - 2a_{ij} \sin \alpha \cos \alpha + a_{jj} \cos^2 \alpha \\
a'_{ij} &= a'_{ji} = a_{ij}(\cos^2 \alpha - \sin^2 \alpha) - (a_{ii} - a_{jj}) \sin \alpha \cos \alpha, & i &\neq j
\end{aligned}$$

$\mu = i - 1$

The reduction method of Givens

reduces a symmetric matrix $A = (a_{ij})_{i,j=1,\dots,n}$ to a symmetric tridiagonal matrix

Series of transformations (not destroying already produced zeros);
each transformation produces two zeros (symmetry).

Choose $T_{ij}(\alpha)$ with α such that $a'_{i-1,j} = a'_{j,i-1} = 0$

Angle of rotation: $-a_{i,i-1} \sin \alpha + a_{j,i-1} \cos \alpha = 0$

$$\cos \alpha = \frac{\pm a_{i,i-1}}{\sqrt{a_{i,i-1}^2 + a_{j,i-1}^2}}, \quad \sin \alpha = \frac{\pm a_{j,i-1}}{\sqrt{a_{i,i-1}^2 + a_{j,i-1}^2}}$$

$\neq 0$, if irreducible

Series of the indexes (i, j) :

$$\begin{array}{llll}
(2, 3), & (2, 4), & \dots, & (2, n) & \Rightarrow & a'_{13} = \dots = a'_{1n} = 0 \\
& & & & & a'_{31} = \dots = a'_{n1} = 0 \\
(3, 4), & \dots, & (3, n) & \Rightarrow & a'_{24} = \dots = a'_{2n} = 0 \\
& & & & & a'_{42} = \dots = a'_{n2} = 0 \\
& \ddots & & \vdots & & \vdots \\
& & (n-1, n) & \Rightarrow & a'_{n-2,n} = a'_{n,n-2} = 0
\end{array}$$

After $\frac{1}{2}(n-1)(n-2)$ transformations we have a symmetric tridiagonal matrix (numerically stable algorithm).

Costs $\approx n^3$ multiplications, twice as many as with Householder!

Fast Givens transformation by factorization of A (Schwarz p. 253).

In general: A not symmetric, then Givens produces the Hessenberg matrix.

QR decomposition of tridiagonal resp. Hessenberg matrices

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & \ddots & \ddots & \\ & \ddots & \ddots & a_{n-1,n} \\ 0 & & a_{n,n-1} & a_{nn} \end{pmatrix} \text{ real}$$

Idea $A =: A^{(1)} \xrightarrow{a_{21}^{(1)}=0} T_1^T A^{(1)} =: A^{(2)} \xrightarrow{a_{32}^{(2)}=0} T_2^T A^{(2)} \rightarrow \dots \xrightarrow{a_{n,n-1}^{(n-1)}=0} A^n =: R$

Hence $T_{n-1}^T T_{n-2}^T \dots T_1^T A = R$, i.e., $A = \underbrace{T_1 \dots T_{n-1}}_{=:Q \text{ orthogonal}} \cdot R = QR$

Choose $T_{i,i+1}(\alpha)$ with α such that $a'_{i+1,i} = 0$

Index series: $(1, 2), (2, 3), \dots, (n-1, n) \Rightarrow a'_{21} = a'_{32} = \dots = a'_{n,n-1} = 0$

Angle of rotation: Formula for $a'_{j\mu}$ with $j = i+1$, $\mu = i$ (see above)

$$a'_{i+1,i} = -a_{ii} \sin \alpha + a_{i+1,i} \cos \alpha \stackrel{!}{=} 0$$

$$\cos \alpha = \frac{\pm a_{ii}}{\sqrt{a_{ii}^2 + a_{i+1,i}^2}}, \quad \sin \alpha = \frac{\pm a_{i+1,i}}{\sqrt{a_{ii}^2 + \underbrace{a_{i+1,i}^2}_{\neq 0, \text{ if irreducible}}}}$$

Transformation

$$T_{i,i+1}^T A = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & \cos \alpha & \sin \alpha & \\ & & & -\sin \alpha & \cos \alpha & \\ & & & & & 1 & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix} \begin{pmatrix} a_{11}^{(i)} & & \dots \\ & \ddots & \\ & & a_{ii}^{(i)} \\ & & & a_{i+1,i}^{(i)} & \dots \\ & & & & \ddots \end{pmatrix}$$

Changes given only in the i -th and $(i+1)$ -th row.

QR decomposition for symmetric tridiagonal matrix with Givens is easier then with Householder!

Remarks:

A symmetric tridiagonal matrix: $A = QR$ implies $A' = RQ$

A' similar to A : $A' = Q^T A Q$, A' symmetric.

First, A' Hessenberg matrix, then tridiagonal matrix because of symmetry.

Jacobian rotation: $T_{ij}^T A T_{ij} =: A' \quad (a_{ij} \neq 0)$

Choose $T_{ij}(\alpha)$ with α such that $a'_{ij} = 0 \quad (i \neq j)$

$$\begin{aligned} \text{Angle of rotation} \quad a'_{ij} &= a_{ij} \underbrace{(\cos^2 \alpha - \sin^2 \alpha)}_{=\cos 2\alpha} - (a_{ii} - a_{jj}) \underbrace{\sin \alpha \cos \alpha}_{=\frac{1}{2} \sin 2\alpha} \stackrel{!}{=} 0 \\ a_{ij} \cos 2\alpha &\stackrel{!}{=} (a_{ii} - a_{jj}) \frac{1}{2} \sin 2\alpha \end{aligned}$$

Different cases $a_{ii} = a_{jj}$: set $\alpha = \frac{\pi}{4}$, i.e., $\cos \alpha = \sin \alpha = \frac{1}{2}\sqrt{2}$
 $a_{ii} \neq a_{jj}$: set $\tan 2\alpha = \frac{2a_{ij}}{a_{ii}-a_{jj}}$, $|\alpha| < \frac{\pi}{4}$ (main value)
 (implies $\cos \alpha, \sin \alpha$)

More stable formulas: $\vartheta := \cot 2\alpha = \frac{a_{ii} - a_{jj}}{2a_{ij}}$
 $t := \tan \alpha$ the smallest solution in absolut value of
 $\boxed{t^2 + 2t\vartheta - 1 = 0}$ identity
 $t = \frac{s(\vartheta)}{|\vartheta| + \sqrt{1 + \vartheta^2}}, \quad s(\vartheta) := \begin{cases} 1, & \text{if } \vartheta \geq 0 \\ -1, & \text{if } \vartheta < 0 \end{cases}$

Angle of rotation: $\cos \alpha = \frac{1}{\sqrt{1+t^2}}, \quad \sin \alpha = t \cdot \cos \alpha$
 $(\tan \alpha = \frac{\sin \alpha}{\cos \alpha}, \quad \cos^2 \alpha = \frac{1}{1+\tan^2 \alpha})$

Jacobi 1804 – 1851, Givens 1954, Householder 1964

Jacobian method: Iteration to diagonal form

A real and symmetric matrix

Infinite series of similarity transformations

$A =: A^{(1)} \rightarrow A^{(2)} \rightarrow A^{(3)} \rightarrow \dots \rightarrow D$

Jacobian rotation $T_{ij}(\alpha)$ with $a'_{ij} = a'_{ji} = 0 \quad (i \neq j)$ (see above)

Iteration: $A^{(1)} := A, \quad A^{(k+1)} := T_k^T A^{(k)} T_k, \quad k = 1, 2, 3, \dots$

$T_k := T_{ij}(\alpha)$, where (i, j) such that $a_{ij}^{(k)}$ is the
largest non-diagonal element in absolut value of $A^{(k)}$
and α such that $a_{ij}^{(k+1)} = a_{ji}^{(k+1)} = 0$.

Convergence

$S(A^{(k)}) := 2 \sum_{\nu > \mu} (a_{\nu\mu})^2$ square- Σ of the non-diagonal elements

$S(A^{(k)}) \rightarrow 0$ if $k \rightarrow \infty$, i.e., $A^{(k)} \rightarrow D = \text{diag}(\lambda_1, \dots, \lambda_n)$.

Theorem 10.1 *The Jacobian method converges at least linear, i.e.,*

$$S(A^{(k+1)}) \leq c S(A^{(k)}), \quad c < 1, \quad k = 1, 2, \dots$$

Proof

$$S(A^{(k+1)}) = 2 \left\{ \sum_{\substack{\nu > \mu \\ \nu, \mu \neq i, j}} (a_{\nu\mu}^{(k+1)})^2 + \sum_{\nu \neq i, j} \left((a_{\nu i}^{(k+1)})^2 + (a_{\nu j}^{(k+1)})^2 \right) + (a_{ij}^{(k+1)})^2 \right\}$$

Transformation $A^{(k)} \rightarrow A^{(k+1)}$ (see above)

Consider the non-diagonal elements!

The first sum: There are no changes in the coefficients!

The second sum: $a_{\nu i}^{(k+1)} = a_{\nu i}^{(k)} \cos \alpha \pm a_{\nu j}^{(k)} \sin \alpha$ (symmetry $a_{\nu i}^{(k)} = a_{i \nu}^{(k)}$)

$$(a_{\nu i}^{(k+1)})^2 = (a_{\nu i}^{(k)})^2 \cos^2 \alpha + (a_{\nu j}^{(k)})^2 \sin^2 \alpha \pm 2a_{\nu i}^{(k)} a_{\nu j}^{(k)} \sin \alpha \cos \alpha$$

Addition of the two equations yields

$$(a_{\nu i}^{(k+1)})^2 + (a_{\nu j}^{(k+1)})^2 = (a_{\nu i}^{(k)})^2 + (a_{\nu j}^{(k)})^2$$

i.e., there are no changes in the sums.

The third sum: $a_{ij}^{(k)}$ the largest non-diagonal element in absolut value,
i.e., $a_{ij}^{(k+1)} = 0$

Hence, it follows $S(A^{(k+1)}) < S(A^{(k)})$.

Consider a rough estimation for the constant c

$$S(A^{(k)}) \leq n(n-1)(a_{ij}^{(k)})^2$$

(number of the non-diagonal elements $\times$ largest element in absolut value to the square) or

$$(a_{ij}^{(k)})^2 \geq \frac{1}{n(n-1)} S(A^{(k)}).$$

Hence, it follows

$$\begin{aligned} S(A^{(k+1)}) &\leq S(A^{(k)}) - 2(a_{ij}^{(k)})^2 \\ &\leq S(A^{(k)}) - \frac{2}{n(n-1)} S(A^{(k)}) = \underbrace{\left(1 - \frac{2}{n(n-1)}\right)}_{=:c} S(A^{(k)}) \\ &\text{with } 0 \leq c < 1. \quad \square \end{aligned}$$

Theorem 10.2 (*Error bound*)

Let $\lambda_1 \leq \dots \leq \lambda_n$ be the eigenvalues of the symmetric matrix A and

$$d_1^{(k)} \leq d_2^{(k)} \leq \dots \leq d_n^{(k)}$$

the diagonal elements $a_{jj}^{(k)}$ of $A^{(k)}$. Then it follows

$$\max_j |d_j^{(k)} - \lambda_j| \leq \sqrt{S(A^{(k)})}, \quad k = 1, 2, \dots$$

Remarks

i) Stopping criterion with a given accuracy ε :

$$S(A^{(k)}) \leq \varepsilon^2 \quad \text{or simple} \quad n(n-1)|a_{ij}^{(k)}|^2 \leq \varepsilon^2$$

- ii) The Jacobian method converges quadratically: $S(A^{(k+1)}) \leq \tilde{c} (S(A^{(k)}))^2$, as soon as the non-diagonal elements are sufficiently small in absolute value ($\rightarrow$ local convergency, c.f. Newton's method)!
- iii) The structure of a sparse matrix will be destroyed by the Jacobian method!

The cyclic Jacobian method

Instead of determining the largest non-diagonal element in absolute value, choose a fixed index series, which will be repeated several times:

$$\left. \begin{array}{cccc} (1, 2) & (1, 3) & \dots & (1, n) \\ & (2, 3) & \dots & (2, n) \\ & & \ddots & \vdots \\ & & & (n-1, n) \end{array} \right\} 1 \text{ cycle}$$

Convergence behaviour as in the classical case, but more costs (a difficult proof)!

Computation of the eigenvectors by the Jacobian method

$$\begin{aligned} A^{(k+1)} &= T_k^T A^{(k)} T_k = T_k^T \dots T_1^T A T_1 \dots T_k =: V_k^T A V_k \\ V_k &:= T_1 \dots T_k \text{ orthogonal} \end{aligned}$$

Theorem 7.1 implies that the symmetric matrix A is orthogonally diagonalizable, i.e., $U^T A U = \text{diag}(\lambda_1, \dots, \lambda_n)$, $U = (x_1, \dots, x_n)$, x_j orthogonal eigenvector of the eigenvalue λ_j ($j = 1, \dots, n$).

Hence $A^{(k+1)} = V_k^T A V_k$ approximates $\text{diag}(\lambda_1, \dots, \lambda_n)$, and the column vectors of V_k approximates the eigenvectors of the eigenvalues λ_k .

Summary

The Jacobian method for real and symmetric matrices:

Iterative computation of all eigenvalues and the corresponding eigenvectors

The matrix T_k will be chosen classically or the cyclic version is used

Stopping criterion is used or a given number of cycles is chosen

Approximation for the eigenvalues λ_j : diagonal elements $a_{jj}^{(k+1)}$

Approximation for the eigenvectors x_j : column vectors $v_j^{(k)}$ of V_k

Local quadratic convergence for $k \geq m_0$

The structure of a sparse matrix (tridiagonal matrix) will be destroyed

11 Vector Iteration of Mises and Wielandt

Raising the dominant eigenvalue to a high power: $A^k x = \lambda^k x$

Assumption: A a real matrix with dominant eigenvalue λ_1 and eigenvector x_1 ,
 A diagonalizable (n linearly independent eigenvectors)
 (the assumptions can be essentially weaker)

Vector iteration $z^{(k+1)} := Az^{(k)} = A^{k+1}z^{(0)}$, $k = 0, 1, 2, \dots$

Theorem 11.1 *The vector iteration converges linearly with a suitable initial vector $z^{(0)}$, i.e., if $k \rightarrow \infty$ the vectors $z^{(k)}$ become parallel to the eigenvector x_1 and the quotients $z_j^{(k+1)}/z_j^{(k)}$ converge to λ_1 (for suitable components $z_j^{(k)} \neq 0$, $x_{1j} \neq 0$).*

Proof

Eigenvalues $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$

Eigenvectors $x_1, \dots, x_n$ linearly independent (a simplification)

Initial vector $z^{(0)} = \alpha_1 x_1 + \dots + \alpha_n x_n$ ($z^{(0)}$ suitable, i.e., $\alpha_1 \neq 0$)

$$\begin{aligned} \text{Iteration } z^{(k+1)} = A^{k+1}z^{(0)} &= \alpha_1 \lambda_1^{k+1} x_1 + \dots + \alpha_n \lambda_n^{k+1} x_n \\ &= \alpha_1 \lambda_1^{k+1} x_1^{k+1} \left\{ x_1 + \underbrace{\frac{\alpha_2}{\alpha_1} \left(\frac{\lambda_2}{\lambda_1} \right)^{k+1}}_{\rightarrow 0} x_2 + \dots + \frac{\alpha_n}{\alpha_1} \underbrace{\left(\frac{\lambda_n}{\lambda_1} \right)^{k+1}}_{\rightarrow 0} x_n \right\} \end{aligned}$$

$$z_j^{(k+1)}/z_j^{(k)} \rightarrow \lambda_1 \quad (z_j^{(k)} \neq 0, x_{1j} \neq 0) \quad \text{if } k \rightarrow \infty \quad \square$$

Example

$$A = \begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}, \quad z^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad z^{(1)} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad z^{(2)} = \begin{pmatrix} 9 \\ 8 \end{pmatrix}, \quad z^{(3)} = \begin{pmatrix} 41 \\ 42 \end{pmatrix},$$

$$z^{(4)} = \begin{pmatrix} 209 \\ 208 \end{pmatrix}, \quad z^{(5)} = \begin{pmatrix} 1041 \\ 1042 \end{pmatrix}, \quad \dots$$

quotients ($j = 1$): 9, 4.55..., 5.097..., 4.9808..., ...
 quotients ($j = 2$): 4, 5.25..., 4.952..., 5.0096..., ...

The eigenvalues are $\lambda_1 = 5$ $\lambda_2 = 1$;

the eigenvectors are $x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $x_2 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$.

Remarks The speed of convergence depends on the factor $\left| \frac{\lambda_2}{\lambda_1} \right|$.
 The initial vector can easily be found (because of rounding errors we usually have $\alpha_1 \neq 0$).
 The normalization of the vectors $z^{(k)}$ is convenient.

The method of Mises

Initial vector $y^{(0)}$ with $\|y^{(0)}\| = 1$ ($\alpha_1 \neq 0$, z.B. $\|\cdot\|_\infty$)

$$\tilde{y}^{(k+1)} := Ay^{(k)}, \quad y^{(k+1)} := \frac{\tilde{y}^{(k+1)}}{\|\tilde{y}^{(k+1)}\|}, \quad k = 0, 1, 2, \dots$$

$$\tilde{y}_j^{(k+1)} / y_j^{(k+1)} \rightarrow \lambda_1 \quad \text{if } k \rightarrow \infty$$

$$(\text{sign} \lambda_1)^k y^{(k)} \rightarrow \frac{\alpha_1 x_1}{\|\alpha_1 x_1\|} \quad \text{if } k \rightarrow \infty$$

Linear convergence with the factor $q := \left| \frac{\lambda_2}{\lambda_1} \right| < 1$:

$$\tilde{y}_j^{(k+1)} / y_j^{(k)} = \lambda_1 + O(q^k) \quad (k \rightarrow \infty).$$

Modifications

1. A not diagonalizable
2. complex eigenvalue $\lambda_1 = |\lambda_1|e^{i\varphi_1}$
3. dominant eigenvalue of multiplicity p
4. distinct dominant eigenvalues
5. raising the zeros of a polynomial to high power

Acceleration of convergence

Rayleigh quotient $r_A(z) = \frac{z^H A z}{z^H z}$

Series $\{z^{(k)}\}_{k \geq 0}$ resp. $\{y^{(k)}\}_{k \geq 0}$

New series

$$\{r_A(z^{(k)})\}_{k \geq 0} \quad \text{with} \quad r_A(z^{(k)}) = \frac{(z^{(k)})^H z^{(k+1)}}{(z^{(k)})^H z^{(k)}} = \frac{(y^{(k)})^H \tilde{y}^{(k+1)}}{(y^{(k)})^H y^{(k)}}$$

Theorem 11.2 *Let A be Hermitian with eigenvalues $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$, then*

$$r_A(z^{(k)}) = \lambda_1 + O(q^{2k}) \quad (k \rightarrow \infty).$$

Proof

Real eigenvalues, linearly independent orthogonal eigenvectors (Th. 7.1)

$z^{(0)} = \alpha_1 x_1 + \dots + \alpha_n x_n$ ($\alpha_1 \neq 0$), $\{x_1, \dots, x_n\}$ orthogonal normalized system

$$z^{(k)} = \alpha_1 \lambda_1^k x_1 + \dots + \alpha_n \lambda_n^k x_n$$

$$\begin{aligned}
(z^{(k)})^H z^{(k)} &= \left(\sum_{\nu} \bar{\alpha}_{\nu} \lambda_{\nu}^k x_{\nu}^H \right) \left(\sum_{\mu} \alpha_{\mu} \lambda_{\mu}^k x_{\mu} \right) = \sum_{\nu} |\alpha_{\nu}|^2 \lambda_{\nu}^{2k} \\
&= |\alpha_1|^2 \lambda_1^{2k} \left(1 + \sum_{\nu} \underbrace{\left| \frac{\alpha_{\nu}}{\alpha_1} \right|^2 \left(\frac{\lambda_{\nu}}{\lambda_1} \right)^{2k}}_{\rightarrow 0} \right) \\
(z^{(k)})^H z^{(k+1)} &= \dots = |\alpha_1|^2 \lambda_1^{2k+1} \left(1 + \sum_{\nu} \underbrace{\left| \frac{\alpha_{\nu}}{\alpha_1} \right|^2 \left(\frac{\lambda_{\nu}}{\lambda_1} \right)^{2k+1}}_{\rightarrow 0} \right) \\
r_A(z^{(k)}) &= \lambda_1 \frac{1 + \sum_{\nu=2}^n \left| \frac{\alpha_{\nu}}{\alpha_1} \right|^2 \left(\frac{\lambda_{\nu}}{\lambda_1} \right)^{2k+1}}{1 + \sum_{\nu=2}^n \left| \frac{\alpha_{\nu}}{\alpha_1} \right|^2 \left(\frac{\lambda_{\nu}}{\lambda_1} \right)^{2k}} = \lambda_1 (1 + O(q^{2k})) = \lambda_1 + O(q^{2k}) \quad (k \rightarrow \infty) \quad \square
\end{aligned}$$

The inverse iteration of Wielandt

Application of the vector iteration to A^{-1} (A nonsingular)

$$z^{(k+1)} := A^{-1} z^{(k)}, \quad k = 0, 1, 2, \dots \quad (z^{(0)} \text{ suitable})$$

yields the inverse of the largest eigenvalue of A in absolute value.

Instead of computing A^{-1} , the linear system $Az^{(k+1)} = z^{(k)}$ will be solved:

$$A = LR, \quad Lv^{(k)} = z^{(k)}, \quad Rz^{(k+1)} = v^{(k)} \quad (\text{with normalization}).$$

Eigenvalues $|\lambda_1| > |\lambda_2| > \dots > |\lambda_{n-1}| > |\lambda_n|$

Factor of convergence $\left| \frac{\lambda_2}{\lambda_1} \right|$ respectively $\left| \frac{\lambda_n}{\lambda_{n-1}} \right|$

Shift: shift of the eigenvalues

Matrix $A_s := A - sI$ with *shift parameter* s

λ eigenvalue of A if and only if $\lambda - s$ eigenvalue of A_s

i) Acceleration of convergence:

$$s \text{ such that } \left| \frac{\lambda_n - s}{\lambda_{n-1} - s} \right| < \left| \frac{\lambda_n}{\lambda_{n-1}} \right|$$

ii) Computation of the eigenvalues λ_j ($j \in \{1, 2, \dots, n\}$):

$$s \text{ such that } |\lambda_j - s| < |\lambda_k - s| \text{ for all } k \neq j$$

Then the inverse iteration is applied to A_s

Choice of the shift parameter:

An approximation t_j of λ_j is known, then choose $s = t_j$;

fitting of the parameter s after some steps!

SUPPLEMENTARY EXAMPLES – No. 9

INVERSE VECTOR ITERATION WITH SHIFT

For a given matrix A the series $\{z^{(k)}\}_{k \geq 0}$ with $A_s z^{(k+1)} := z^{(k)}$ has to be computed ($z^{(0)}$ suitable initial value), where the shift parameter s is chosen such that

$$|\lambda_j - s| < |\lambda_k - s| \quad \text{for all } k \neq j.$$

Then approximations of the eigenvalue λ_j follow.

For the matrix

$$A = \begin{pmatrix} 23 & 26 & -51 \\ -25 & 74 & -51 \\ -25 & -26 & 49 \end{pmatrix}$$

let one eigenvalue λ be approximately known as 50.

Compute this eigenvalue λ more precisely with the inverse iteration.

Choose $s = 50$:

Inverse iteration for A_{50} with the initial vector $z^{(0)} = (1, 0, 0)^T$ produces the series $z^{(1)}, z^{(2)}, z^{(3)}, z^{(4)}, \dots$ as follows:

$$\begin{pmatrix} -0.259\dots \\ 0.240\dots \\ 0.240\dots \end{pmatrix}, \begin{pmatrix} -0.125\dots \\ 0.124\dots \\ 0.124\dots \end{pmatrix}, \begin{pmatrix} -0.0625\dots \\ 0.0624\dots \\ 0.0624\dots \end{pmatrix}, \begin{pmatrix} -0.03125\dots \\ 0.03124\dots \\ 0.03124\dots \end{pmatrix}, \dots$$

and hence the series of approximations $\frac{z_1^{(k+1)}}{z_1^{(k)}} \quad (k = 0, 1, 2, \dots)$

$$-0.25\dots, -0.48\dots, -0.4992\dots, -0.499972\dots, \dots$$

respectively using the Rayleigh quotient $\frac{z^{(k)T} z^{(k+1)}}{z^{(k)T} z^{(k)}} \quad (k = 0, 1, 2, \dots)$

$$-0.25\dots, -0.505\dots, -0.5002\dots, -0.5000091\dots, \dots$$

After 4 iteration steps the Rayleigh quotient yields the approximation $N = -0.5000091\dots$ of the inverse of the smallest eigenvalue of A_{50} , and hence the approximation $\hat{\lambda} = 50 + \frac{1}{N} = 48.00003\dots$ of the wanted eigenvalue.

The matrix A has the eigenvalues 100, 48 and -2 .

12 LR and QR method

Iteration method with the series $\{A_k\}_{k \geq 1}$ and $A_k \rightarrow \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$

Similarity transformations using *LR* and *QR* decompositions!
 Compute simultaneously the eigenvalues and the eigenvectors!

Idea: Raising A to a high power A^k similar to the vector iteration

Factorization $A^k = \tilde{L}_k \tilde{R}_k$ (*LR* decomposition)

Refinement: compute all eigenvalues $\lambda_1, \dots, \lambda_n$

Application: Symmetric and non-symmetric matrices

First, reduction to symmetric tridiagonal respectively Hessenberg matrix,
LR resp. *QR* decomposition is simpler and the structure is preserved.

LR method

Assumption: Real nonsingular matrix A with $|\lambda_1| > \dots > |\lambda_n| > 0$

Existence of the necessary *LR* decompositions

Algorithm: $A_1 := A$
 $A_k =: L_k R_k, \quad A_{k+1} := R_k L_k, \quad k = 1, 2, \dots$

i) *Similarity transformation*

$$A_{k+1} = L_k^{-1} A_k L_k = (L_1 \dots L_k)^{-1} A (L_1 \dots L_k) \quad (\text{cond}(L_k) = ?)$$

ii) *Factorization of A^k :*

$$A^k = \underbrace{L_1 \dots L_k}_{\tilde{L}_k} \underbrace{R_k \dots R_1}_{\tilde{R}_k} =: \tilde{L}_k \tilde{R}_k$$

$$\text{Since } A = A_1 = L_1 R_1, \quad A_2 = L_2 R_2, \dots$$

$$\begin{aligned} A^k &= L_1 \underbrace{(R_1 L_1)}_{A_2} (R_2 \dots L_2) R_2 = L_1 A_2^{k-1} R_1 \\ &= L_1 \underbrace{(L_2 (R_2 L_2))}_{A_3} (R_3 \dots L_3) R_3 = L_1 L_2 A_3^{k-2} R_2 R_1 \\ &= L_1 \dots L_{k-1} A_k R_{k-1} \dots R_1 \\ &= (L_1 \dots L_{k-1} L_k) (R_k R_{k-1} \dots R_1) = \tilde{L}_k \tilde{R}_k \end{aligned}$$

Hence $A^k = \tilde{L}_k \tilde{R}_k$, where

$$\tilde{L}_k = \tilde{L}_{k-1} L_k = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix}, \quad \tilde{R}_k = R_k \tilde{R}_{k-1} = \begin{pmatrix} \tilde{r}_{11}^{(k)} & & * \\ & \ddots & \\ 0 & & \tilde{r}_{nn}^{(k)} \end{pmatrix},$$

$$R_k = \begin{pmatrix} r_{11}^{(k)} & & * \\ & \ddots & \\ 0 & & r_{nn}^{(k)} \end{pmatrix}, \quad \tilde{r}_{11}^{(k)} = r_{11}^{(k)} \tilde{r}_{11}^{k-1},$$

$$\text{vector iteration} \quad z^{(k)} = A^k e_1 = \begin{pmatrix} \tilde{r}_{11}^{(k)} \\ \vdots \end{pmatrix}, \quad z_1^{(k)} / z_1^{(k-1)} = r_{11}^{(k)},$$

$$r_{11}^{(k)} \rightarrow \lambda_1 \quad \text{if } k \rightarrow \infty.$$

Under additional assumptions of A it follows

$$R_k \rightarrow \begin{pmatrix} \boxed{\lambda_1} & & * \\ & \ddots & \\ & & \lambda_n \end{pmatrix}, \quad A_k \rightarrow \begin{pmatrix} \boxed{\lambda_1} & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \quad L_k \rightarrow I \quad \text{if } k \rightarrow \infty.$$

Theorem 12.1 (*LR method*) Let $A = T \operatorname{diag} (\lambda_1, \dots, \lambda_n) T^{-1}$ with LR decomposable matrices T and T^{-1} and with

$$|\lambda_1| > \dots > |\lambda_n| > 0.$$

Let the LR method be feasible for A , then

$$\lim_{k \rightarrow \infty} A_k = \lim_{k \rightarrow \infty} R_k = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \quad \lim_{k \rightarrow \infty} L_k = I.$$

Proof: Investigation of the factorization of A^k

$$A^k = \tilde{L}_k \tilde{R}_k, \quad \tilde{L}_k = \tilde{L}_{k-1} L_k, \quad \tilde{R}_k = R_k \tilde{R}_{k-1}$$

Assumption on A : $A = T D T^{-1}$, $D := \operatorname{diag} (\lambda_1, \dots, \lambda_n)$, D^{-1} exists

$$T = L_T R_T, \quad T^{-1} = L_{T^{-1}} R_{T^{-1}}$$

$$i > j : \left(\frac{\lambda_i}{\lambda_j} \right)^k \rightarrow 0, \quad k \rightarrow \infty.$$

Lemmas

- (1) $A^k = TD^kT^{-1} (= (TDT^{-1})^k = TD \underbrace{T^{-1}T}_{} D \underbrace{T^{-1} \dots T}_{} DT^{-1})$
- (2) $D^k L_{T^{-1}} D^{-k} = I + F_k$ with $F_k \rightarrow 0$ if $k \rightarrow \infty$
(coefficients in the diagonal are one, below $\ell_{ij}(\frac{\lambda_i}{\lambda_j})^k \rightarrow 0$ ($i > j$))
- (3) $R_T F_k R_T^{-1} \rightarrow 0$ if $k \rightarrow \infty$
(R_T^{-1} exists, because of $R_T = L_T^{-1}T$; $F_k \rightarrow 0$ if $k \rightarrow \infty$)
- (4) *Decomposition* $I + \underbrace{R_T F_k R_T^{-1}}_{\rightarrow 0} =: \underbrace{\hat{L}_k \hat{R}_k}_{\rightarrow I}$ exists for sufficiently large k
- (5) $\hat{L}_k \rightarrow I$ and $\hat{R}_k \rightarrow I$ if $k \rightarrow \infty$.

Corollaries

- a) $A^k = TD^kT^{-1} = L_T R_T D^k L_{T^{-1}} R_{T^{-1}} = L_T R_T (\underbrace{D^k L_{T^{-1}} D^{-k}}_{\text{assumption}}) D^k R_{T^{-1}}$
(1)
 $= L_T (R_T (I + F_k) R_T^{-1}) R_T D^k R_{T^{-1}} = L_T (\underbrace{I + R_T F_k R_T^{-1}}_{\text{assumption}}) R_T D^k R_{T^{-1}}$
(2)
 $= (L_T \hat{L}_k) (\hat{R}_k R_T D^k R_{T^{-1}}) \underset{\text{uniqueness}}{=} \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ * & & 1 \end{pmatrix} \begin{pmatrix} * & \dots & * \\ & \ddots & \vdots \\ 0 & & * \end{pmatrix} = \tilde{L}_k \tilde{R}_k$
(4)
- b) $A_k = L_k R_k$ (LR decomposition)
 $L_k = \tilde{L}_{k-1}^{-1} \tilde{L}_k = (L_T \hat{L}_{k-1})^{-1} (L_T \hat{L}_k) = \hat{L}_{k-1}^{-1} \hat{L}_k \rightarrow I$ if $k \rightarrow \infty$
(a) (5)
 $R_k = \tilde{R}_k \tilde{R}_{k-1}^{-1} = (\hat{R}_k R_T D^k R_{T^{-1}}) (\hat{R}_{k-1} R_T D^{k-1} R_{T^{-1}})^{-1}$
(a)
 $= \underbrace{\hat{R}_k}_{\rightarrow I} R_T D R_T^{-1} \underbrace{\hat{R}_{k-1}^{-1}}_{\rightarrow I} \xrightarrow{(5)} R_T D R_T^{-1} = \begin{pmatrix} \lambda_1 & \dots & * \\ & \ddots & \vdots \\ 0 & & \lambda_n \end{pmatrix} =: R_\lambda$
- c) *Altogether*

$$\lim_{k \rightarrow \infty} A_k = \lim_{k \rightarrow \infty} L_k R_k = R_\lambda. \quad \square$$

Remarks

1. The method breaks down, if A_k is not LR decomposable, and in general it does not converge, if T or T^{-1} is not LR decomposable.
2. The computation of the LR decomposition can be ill-conditioned.
3. If the quotients $|\frac{\lambda_{i+1}}{\lambda_i}|$ are close to 1, then the convergence is slow. Shift techniques can accelerate the convergence.
4. Stopping criterion: $\max_{\nu > \mu} |a_{\nu\mu}^{(k)}| \leq \delta$ (δ given)
5. Numerical stability of the method (compare with $\text{cond}(L_k)$): Gaussian elimination method without pivot searching is not always stable!
6. If A symmetric, then $A_k \rightarrow \text{diag}(\lambda_1, \dots, \lambda_n)$ if $k \rightarrow \infty$.
7. Expensive: one step $A_k \rightarrow A_{k+1}$ needs $\frac{2}{3}n^3$ multiplications; therefore application mostly on reduced matrices: Tridiagonal or Hessenberg form (this structure is preserved)
8. In case of a positive definite matrix A the Cholesky decomposition is used:
 $A_{k+1} := L_k^T L_k$, $A_k := L_k L_k^T$.
9. A weakening of the assumption $|\lambda_1| > \dots > |\lambda_n|$ is possible.

QR decomposition

$$A = QR = \begin{pmatrix} * & \dots & * \\ \vdots & & \vdots \\ * & \dots & * \end{pmatrix} \begin{pmatrix} r_{11} & \dots & * \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{pmatrix}, \quad Q \text{ unitary}$$

It exists for nonsingular A (see Theorem 9.3); the computation is numerically stable using Householder respectively Givens transformations, but not unique! If A is real, then Q and R are real, and Q is an orthogonal matrix!

QR method

Assumption Nonsingular matrix A with eigenvalues $|\lambda_1| > \dots > |\lambda_n| > 0$

Algorithm $A_1 := A$
 $A_k := Q_k R_k$, $A_{k+1} := R_k Q_k$, $k = 1, 2, \dots$

i) Similarity transformation

$$A_{k+1} = Q_k^T A_k Q_k = (Q_1 \dots Q_k)^T A_1 (Q_1 \dots Q_k), \quad Q_k \text{ unitary}$$

ii) Factorization of A^k

$$A^k = Q_1 \dots Q_k R_k \dots R_1 =: \tilde{Q}_k \tilde{R}_k \quad (\text{compare with the } LR \text{ method})$$

$$\tilde{R}_k = R_k \tilde{R}_{k-1}, \quad \tilde{r}_{11}^{(k)} = r_{11}^{(k)} \tilde{r}_{11}^{(k-1)}$$

$$\tilde{Q}_k = \tilde{Q}_{k-1} Q_k = (\tilde{q}_1^{(k)}, \dots, \tilde{q}_n^{(k)}) \quad \text{unitary, i.e., } \|\tilde{q}_1^{(k)}\|_2 = 1$$

iii) **Preservation of the structure** A_k Hessenberg $\implies A_{k+1}$ Hessenberg A_k symmetric tridiagonal matrix $\implies A_{k+1}$ symmetric tridiagonal matrixiv) **Relation to the vector iteration** A diagonalizable, dominating eigenvalue λ_1 with eigenvector x_1
decomposition such that $r_{11}^{(k)} > 0, \tilde{r}_{11}^{(k)} > 0$ ($k \geq 1$) $z^{(0)} = e_1 : z^{(k)} = A^k e_1 = \tilde{Q}_k \tilde{R}_k e_1 = \tilde{r}_{11}^{(k)} \tilde{q}_1^{(k)}$ with $\|z^{(k)}\|_2 = \tilde{r}_{11}^{(k)}$ $normalized: (\text{sign } \lambda_1)^k \underbrace{\frac{z^{(k)}}{\|z^{(k)}\|_2}}_{=\tilde{q}_1^{(k)}} \rightarrow \frac{\alpha_1 x_1}{\|\alpha_1 x_1\|_2}, k \rightarrow \infty$ $j\text{-th component: } (\text{sign } \lambda_1)^k \tilde{q}_{1j}^{(k)} \rightarrow \frac{\alpha_1 x_{1j}}{\|\alpha_1 x_1\|_2}, k \rightarrow \infty$ $(\text{sign } \lambda_1)^{k-1} \underbrace{\frac{z^{(k)}}{\|z^{(k-1)}\|_2}}_{=r_{11}^{(k)} \tilde{q}_1^{(k)}} \rightarrow \lambda_1 \frac{\alpha_1 x_1}{\|\alpha_1 x_1\|_2}, k \rightarrow \infty$ $j\text{-th component:}$ $\underbrace{(\text{sign } \lambda_1)^k \tilde{q}_{1j}^{(k)}}_{\rightarrow \frac{\alpha_1 x_{1j}}{\|\alpha_1 x_1\|_2}} \underbrace{(\text{sign } \lambda_1) r_{11}^{(k)}}_{\rightarrow \lambda_1} \rightarrow \lambda_1 \frac{\alpha_1 x_{1j}}{\|\alpha_1 x_1\|_2}, k \rightarrow \infty$ $hence (\text{sign } \lambda_1) r_{11}^{(k)} \rightarrow \lambda_1, \text{ i.e., } r_{11}^{(k)} \rightarrow |\lambda_1|, k \rightarrow \infty.$ Convergence of the series $\{r_{11}^{(k)}\}_{k \geq 1}$ (element of R_k) to $|\lambda_1|$ with the factor of convergence $\left| \frac{\lambda_2}{\lambda_1} \right|$.

The close relation to the method of inverse iteration also shows

$$r_{nn}^{(k)} \rightarrow |\lambda_n|, k \rightarrow \infty.$$

Theorem 12.2 (*QR method*) Let A be real and diagonalizable, i.e.,
 $A = T \text{diag}(\lambda_1, \dots, \lambda_n) T^{-1}$ with LR decomposable matrix T^{-1} .The eigenvalues of A satisfy $|\lambda_1| > \dots > |\lambda_n| > 0$.The matrices R_k ($k \geq 1$) have real positive diagonal elements r_{jj}
(suitable QR decomposition). Then it holds

$$\lim_{k \rightarrow \infty} A_k = \begin{pmatrix} \lambda_1 & & * \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}, \quad \lim_{k \rightarrow \infty} R_k = \begin{pmatrix} |\lambda_1| & & * \\ & \ddots & \\ 0 & & |\lambda_n| \end{pmatrix},$$
$$\lim_{k \rightarrow \infty} Q_k = \text{diag}(\text{sign } \lambda_j).$$

Proof: Similar to Theorem 12.1 (see Stoer & Bulirsch). $\square$

Remarks

1. If A symmetric, then $\lim_{k \rightarrow \infty} A_k = \text{diag}(\lambda_1, \dots, \lambda_n)$.
2. Application of the QR method to reduced matrices: structure is preserved, QR decomposition simpler, a simple stopping criterion.

3. Example: $A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 4 \end{pmatrix}$

eigenvalues $\lambda_1 = 4.7321$, $\lambda_2 = 3.0$, $\lambda_3 = 1.2679$. $\frac{\lambda_2}{\lambda_1} = 0.64$, $\frac{\lambda_3}{\lambda_2} = 0.42$

$$A_2 = \begin{pmatrix} 3.00 & 1.09 & 0 \\ 1.09 & 3.00 & 1.34 \\ 0 & 1.34 & 3.00 \end{pmatrix}, \quad A_3 = \begin{pmatrix} 3.705 & 0.955 & 0 \\ 0.955 & 3.521 & 0.973 \\ 0 & 0.973 & 1.772 \end{pmatrix}$$

$$A_7 = \begin{pmatrix} 4.679 & 0.297 & 0 \\ 0.297 & 3.052 & 0.027 \\ 0 & 0.027 & 1.268 \end{pmatrix}, \quad A_{10} = \begin{pmatrix} 4.7285 & 0.0781 & 0 \\ 0.0781 & 3.0035 & 0.0020 \\ 0 & 0.0020 & 1.2680 \end{pmatrix}$$

$$|a_{32}^{(10)}| \leq 0.002, \quad |a_{33}^{(10)} - \lambda_3| \leq 0.0001$$

A_{10} decomposable in two pieces: continuation with a smaller matrix!
Study the relation $|a_{n,n-1}^{(k)}| \leq \delta \Rightarrow |a_{nn}^{(k)} - \lambda_n| \leq \varepsilon(\delta)$ in more detail!

4. Weakening of the assumption:

A real with a pair of conjugated complex eigenvalues

$$|\lambda_1| > \dots > |\lambda_r| = |\lambda_{r+1}| > \dots > |\lambda_n|$$

$$A_k = \begin{pmatrix} a_{11}^{(k)} & & & & \\ & \ddots & & & \\ & & \boxed{\begin{matrix} a_{rr}^{(k)} & a_{r,r+1}^{(k)} \\ a_{r+1,r}^{(k)} & a_{r+1,r+1}^{(k)} \end{matrix}} & & \\ & & & \ddots & \\ & & & & a_{nn}^{(k)} \end{pmatrix} =: C_k$$

$\lim_{k \rightarrow \infty} a_{jj}^{(k)} = \lambda_j (j \neq r, r+1)$, eigenvalues of C_k converge to λ_r and λ_{r+1}

5. Acceleration of convergence: shift technique

Choose the shift parameter $s \in \mathbb{R}$ satisfying

$$|\lambda_1 - s| \geq \dots \geq |\lambda_{n-1} - s| \gg |\lambda_n - s| > 0, \quad \text{i.e.,} \quad \left| \frac{\lambda_n - s}{\lambda_{n-1} - s} \right| < \left| \frac{\lambda_n}{\lambda_{n-1}} \right|$$

QR method with shift (for symmetric tridiagonal matrices)

$$A = \begin{pmatrix} \alpha_1 & \beta_1 & & 0 \\ \beta_1 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \beta_{n-1} \\ 0 & & \beta_{n-1} & \alpha_n \end{pmatrix} \quad (\beta_j \neq 0)$$

QR decomposition by the Givens transformation (§10)

$$A_k = \begin{pmatrix} \alpha_1^{(k)} & \beta_1^{(k)} & & 0 \\ \beta_1^{(k)} & \ddots & \ddots & \\ & \ddots & \ddots & \beta_{n-1}^{(k)} \\ 0 & & \beta_{n-1}^{(k)} & \alpha_n^{(k)} \end{pmatrix}$$

$$A_k =: Q_k R_k, \quad Q_k = T_k \dots T_1 \quad (\text{rotations})$$

$$A_{k+1} := Q_k^T A_k Q_k \quad (\text{preserving the structure})$$

Shift parameter s_k for A_k :

choose $s_k := \alpha_n^{(k)}$ or $s_k := \mu_{1/2}^{(k)}$ eigenvalue of C_k

$$C_k := \begin{pmatrix} \alpha_{n-1}^{(k)} & \beta_{n-1}^{(k)} \\ \beta_{n-1}^{(k)} & \alpha_n^{(k)} \end{pmatrix},$$

$$\text{eigenvalue } \mu_{1/2}^{(k)} = \alpha_n^{(k)} + d \pm \sqrt{d^2 + \beta_{n-1}^{(k)2}}, \quad d = \frac{\alpha_{n-1}^{(k)} - \alpha_n^{(k)}}{2}$$

choose s_k as that eigenvalue $\mu_{1/2}^{(k)}$, that is closer to $\alpha_n^{(k)}$, i.e.,

$$s_k := \alpha_n^{(k)} + d - \text{sign } d \sqrt{d^2 + \beta_{n-1}^{(k)2}}$$

Implementation

$$A_k - s_k I =: Q_k R_k, \quad A_{k+1} := R_k Q_k + s_k I = Q_k^T A_k Q_k$$

Costs for one QR step: about $15(n-1)$ multiplications and $n-1$ square roots

Order of convergence at least quadratic, for symmetric tridiagonal matrices the order is cubic for the most cases

$$\text{Stopping criterion } |\beta_{n-1}^{(k)}| \leq \delta, \text{ because } |s_k - \alpha_n^{(k)}| \leq \left| |d| - \sqrt{d^2 + \beta_{n-1}^{(k)2}} \right|$$

Reduction of the matrix A_k : cancellation of the n -th row and column, then repetition of the method.

Comparison of the QR method with the Jacobian method

QR method: Application to a symmetric tridiagonal matrix
shift technique and reduction of the matrix
cubic order of convergence

Jacobian method: the structure of a sparse matrix will be destroyed
quadratic convergence

QR method is about **4 times faster** than the Jacobian method, in the case that the eigenvalues and the eigenvectors have to be computed, and at most **10 times faster**, if only the eigenvalues have to be computed.

Computation of the eigenvectors (see Jacobian method)

Computation of $\tilde{Q}_k$: $\tilde{Q}_k = \tilde{Q}_{k-1}Q_k$, $\tilde{Q}_k = (\tilde{q}_1^{(k)}, \dots, \tilde{q}_n^{(k)})$

The column vectors $q_j^{(k)}$ approximate the eigenvectors x_j , $j = 1, \dots, n$.

Since: $A_{k+1} = \tilde{Q}_k^T A \tilde{Q}_k \longrightarrow U^T A U = \text{diag}(\lambda_1, \dots, \lambda_n)$.

Very expensive!

Inverse vektor iteration method

For the eigenvalue λ_j let the approximation $\tilde{\lambda}$ be known:

Shift parameter $s = \tilde{\lambda}$: $\tilde{A} := A - \tilde{\lambda}I$

(Eigenvector is invariant with respect to shift and inverse iteration)

Linear system $\tilde{A}z^{(k+1)} = z^{(k)}$, $k = 0, 1, 2, \dots$ (nearly singular)

$$z^{(k)} \longrightarrow \text{EV } x_j, \quad k \rightarrow \infty$$

In case of an appropriate initial vector only a few iteration steps are necessary.

Solution of the linear system: LR decomposition, special methods

SUPPLEMENTARY EXAMPLES – NO. 10

QR METHOD WITH SHIFT APPLIED TO SYMMETRIC TRIDIAGONAL MATRICES

QR method with shift: Determine the series $\{A_k = (a_{jj}^{(k)})\}_{k \geq 1}$ with

$$A_1 := A, A_k - s_k I =: Q_k R_k \quad (QR \text{ decomposition}) \quad , A_{k+1} := R_k Q_k + s_k I.$$

Shift strategy: Choose parameter s_k as the eigenvalue λ of the matrix

$$\begin{pmatrix} \alpha_{n-1}^{(k)} & \beta_{n-1}^{(k)} \\ \beta_{n-1}^{(k)} & \alpha_n^{(k)} \end{pmatrix} \quad (\text{part of } A_k)$$

for which $|\alpha_n^{(k)} - \lambda|$ is the smallest.

As $\alpha_n^{(k)}$ and s_k if $k \rightarrow \infty$ tend to the eigenvalue λ_n of A , A_k tends more and more to a decomposable matrix, i.e., $\beta_{n-1}^{(k)} \rightarrow 0$, $k \rightarrow \infty$. This effect can be used for cancelling out the last row and the last column. The reduced matrix can be treated appropriately.

Numerical example

$$A = \begin{pmatrix} 12 & 1 & & & \\ & 1 & 9 & 1 & \\ & & 1 & 6 & 1 \\ & & & 1 & 3 & 1 \\ & & & & 1 & 0 \end{pmatrix}, \quad \text{Gerschgorin: Eigenvalues in } [-1, 13]$$

(the eigenvalues are lying symmetrically to 6, especially 6 is eigenvalue of A).

For the QR method with shift strategy the elements $\beta_{n-1}^{(k)}$, $\alpha_n^{(k)}$ as well as the shift parameters s_k are given in the following table:

k	$\beta_4^{(k)}$	$\alpha_5^{(k)}$	s_k
1	1	0	-0.302775637732
2	$-0.454544295102 \cdot 10^{-2}$	-0.316869782391	-0.316875874226
3	$+0.106774452090 \cdot 10^{-9}$	-0.316875952616	-0.316875952619
4	$+0.918983519419 \cdot 10^{-22}$	-0.316875952617	
$= \lambda_5$			

Continuation with the (4×4) -matrix

k	$\beta_3^{(k)}$	$\alpha_4^{(k)}$	s_k
4	$+0.143723850633 \cdot 10^0$	$+0.299069135875 \cdot 10^1$	$+0.298389967722 \cdot 10^1$
5	$-0.171156231712 \cdot 10^{-5}$	$+0.298386369683 \cdot 10^1$	$+0.298386369682 \cdot 10^1$
6	$-0.111277687663 \cdot 10^{-17}$	$+0.298386369682 \cdot 10^1$	
$= \lambda_4$			

Continuation with the (3×3) -matrix

k	$\beta_2^{(k)}$	$\alpha_3^{(k)}$	s_k
6	$+0.780088052879 \cdot 10^{-1}$	$+0.600201597254 \cdot 10^1$	$+0.600000324468 \cdot 10^1$
7	$-0.838854980961 \cdot 10^{-7}$	$+0.599999999996 \cdot 10^1$	$+0.599999999995 \cdot 10^1$
8	$+0.12781135623 \cdot 10^{-19}$	$+0.599999999995 \cdot 10^1$	
$= \lambda_3$			

The remaining (2×2) -matrix has the eigenvalues

$+0.9016136303414 \cdot 10^1 = \lambda_2$
$+0.123168759526 \cdot 10^2 = \lambda_1$

The result of the *QR method without shift* after 11 iteration steps is given for comparison. The elements of the matrix A_{12} , where the wanted approximations of the eigenvalues are in the diagonal, read:

i	$\beta_{i-1}^{(12)}$	$\alpha_i^{(12)}$
1		$+0.123165309125 \cdot 10^2$
2	$+0.337457586637 \cdot 10^{-1}$	$+0.901643819611 \cdot 10^1$
3	$+0.114079951421 \cdot 10^{-1}$	$+0.600004307566 \cdot 10^1$
4	$+0.463086759853 \cdot 10^{-3}$	$+0.298386376789 \cdot 10^1$
5	$+0.202188244733 \cdot 10^{-10}$	$+0.316875952617 \cdot 10^0$

By continuation with the (4×4) -matrix 23 iteration steps are needed to reach $\beta_3^{(35)} \approx 0.5 \cdot 10^{-10}$.

(See Stoer & Bulirsch).

Chapter III

Numerical Treatment of Ordinary Differential Equations

13 Basic Ideas

Initial Value Problem (IVP)

$y' = f(x, y)$, $y(a) = y_0$, $f : I \times \mathbb{R} \rightarrow \mathbb{R}$, $I := [a, b]$
sector (rectangle), Lipschitz continuous

Wanted function $u : I \rightarrow \mathbb{R}$ satisfying

- 1) u differentiable in I
- 2) $(x, u(x)) \in I \times \mathbb{R}$ for all $x \in I$
- 3) $u'(x) = f(x, u(x))$ for all $x \in I$ and $u(a) = y_0$

Then u is called **solution** of the IVP in the interval I .

Examples

$y' = qy$, $y(0) = 1 : u(x) = e^{qx}$
 $y' = y^2$, $y(0) = 1 : u(x) = \frac{1}{1-x}$ ($x \neq 1$)
 $y' = Ay$, $y(a) = y_0$ linear system
 $y' = f(x, y)$, $y(a) = y_0$ nonlinear system
($f : I \times \mathbb{R}^n \rightarrow \mathbb{R}^n$, solution $u : I \rightarrow \mathbb{R}^n$)
 $y'' = f(x, y, y')$ second order $\rightarrow$ reduction to a first order system

Theory existence, uniqueness, continuous dependence on the
initial value/right-hand side $\rightarrow$ **well posed**
continuation of the solution
solution theory (linear differential equation, Bernoulli differential
equation, exact differential equation, ...)

The necessity of approximation methods

$y' = x^2 + y^2$, $y'' = 6y^2 + x \leadsto$ not solvable by elementary functions!

$y' = 1 - 2xy$, $u(x) = e^{-x^2}(\int_0^x e^{t^2} dt + c)$, IV $y(0) = 1 \leadsto c = 1$
 $\rightarrow$ computation of the integral?

Volterra integral equation

$$u(x) = y_0 + \int_a^x f(t, u(t)) dt = y_0 + \int_0^x u'(t) dt$$

equivalent to the IVP $y' = f(x, y(x))$, $y(a) = y_0$

Continuous methods

Picard iteration $\varphi_{n+1}(x) := y_0 + \int_a^x f(t, \varphi_n(t)) dt$, $n = 0, 1, \dots$

$$\lim_{n \rightarrow \infty} \varphi_n(x) = u(x)$$

yields (infinite) series ($\leadsto$ function) as the solution

[Usually not suitable for applications!]

Discretization methods

Wanted: a function u_h defined on a **grid** I_h , which approximates the exact solution u of the IVP in I best possible.

Grid $I_h = \{x \in I \mid x := x_m, m = 0, 1, \dots, N; x_0 := a, x_{m+1} := x_m + h_m, x_N = b\}$

Uniformly spaced grid $h_m = h$

exact solution $u : I \rightarrow \mathbb{R}$

approximate solution $u_h : I_h \rightarrow \mathbb{R}$, $u_h(x_m) \equiv y_m$

Discretization: derivatives are approximated by difference quotients on the grid

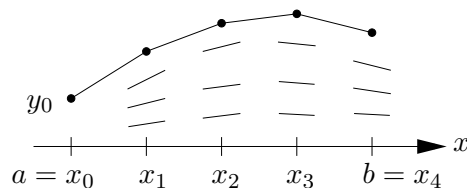
Example: Euler polygon method

$u'(x) = f(x, u(x))$ approximation $\frac{u(x+h) - u(x)}{h} = f(x, u(x)) + O(h)$
 $x \in I$ of order $O(h)$ $x \in I_h$
 difference quotient

$$u_h(x+h) = u_h(x) + hf(x, u_h(x))$$

$$y_{m+1} = y_m + hf(x_m, y_m), \quad m = 0, 1, \dots, N-1$$

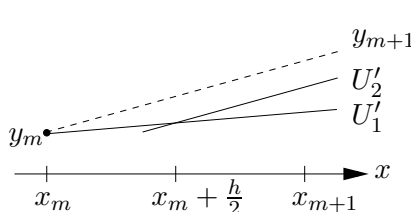
marching along the vector field



Taylor evaluation

$$\begin{aligned}
u(x+h) &= u(x) + hu'(x) + \frac{h^2}{2!}u''(x) + \frac{h^3}{3!}u'''(x) + \dots \\
&= u(x) + hf(x, u(x)) + \frac{h^2}{2}\{f_x + ff_y\}_{(x, u(x))} + \frac{h^3}{3!}\{\dots\} + \dots
\end{aligned}$$

Example: *Half-step method*



2 gradients

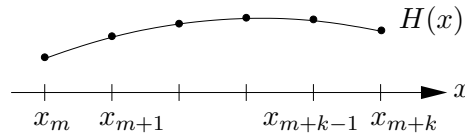
$$\begin{aligned}
U'_1 &:= f(x_m, y_m) \\
U'_2 &:= f\left(x_m + \frac{h}{2}, \underbrace{y_m + \frac{h}{2}f(x_m, y_m)}_{\text{Euler step with } h/2}\right) \\
&= f(x_m, y_m) + \frac{h}{2}\{f_x + ff_y\}_{(x_m, y_m)} + \dots
\end{aligned}$$

$$y_{m+1} := y_m + hU'_2$$

In general

$$\begin{aligned}
U'_1 &:= f(x_m, y_m) \\
U'_2 &:= f(x_m + ch, y_m + chU'_1) \\
y_{m+1} &= y_m + h(b_1U'_1 + b_2U'_2)
\end{aligned}$$

$$\begin{aligned}
c = \frac{1}{2} : b_1 = 0, b_2 = 1 & \quad \text{half-step method} \\
c = 1 : b_1 = b_2 = \frac{1}{2} & \quad \text{improved polygon method}
\end{aligned}$$

2 gradients imply order of approximation $O(h^2)$ **One-step method**Initial value $y_0 : y_m \mapsto y_{m+1}, y_{m+1} := y_m + h\phi(x_m, y_m, h)$ **Multistep methods**Starting values $y_0, y_1, \dots, y_{k-1}$ (approximations) $y_m, y_{m+1}, \dots, y_{m+k-1} \mapsto y_{m+k}$ k -step methodHermite interpolation polynomial $H : y_{m+k} := H(x_{m+k})$

$$H(x) = \sum_{j=0}^{k-1} U_j(x)y_{m+j} + \sum_{j=0}^{k-1} V_j(x)f(x_{m+j}, y_{m+j})$$

$$\sum_{j=0}^k \alpha_j y_{m+j} = h \sum_{j=0}^{k-1} \beta_j f(x_{m+j}, y_{m+j})$$

Examples: *Euler polygon method* $k = 1$

Midpoint rule $k = 2$

$$u'(x) = f(x, u(x)) \rightarrow \frac{u(x+2h) - u(x)}{2h} = f(x+h, u(x+h)) + O(h^2)$$

$$y_{m+2} - y_m = 2hf(x_{m+1}, y_{m+1})$$

Another two-step method

$$y_{m+2} - 4y_{m+1} + 3y_m = -2hf(x_m, y_m) \text{ order } O(h^2)$$

Linear multistep method

$$\sum_{j=0}^k \alpha_j y_{m+j} = h \sum_{j=0}^k \beta_j f_{m+j}, \quad f_{m+j} := f(x_{m+j}, y_{m+j})$$

$$\alpha_j, \beta_j \in \mathbb{R} \quad (\alpha_k = 1)$$

Explicit: $\beta_k = 0$; implicit: $\beta_k \neq 0$

Construction

interpolation and quadrature	}	many possibilities
Taylor evaluation		
divided differences		
algebraic conditions		

Efficient algorithm: formula
error control
stability
implementation/costs

Discretization Error (DE)

Local DE (consider one step): $\tau(x, y, h), y \equiv y_m \equiv u(x_m)$

Error transport: depends mostly on the differential equations

Global DE = $h \times$ local DE + error transport

In the most cases the local error implies information on the global error.

Example: Euler polygon method $y_{m+1} = y_m + hf(x_m, y_m)$

Consider one step $y \rightarrow y + hf(x, y)$

$$\begin{aligned} \text{Local DE} \quad \tau(x, y, h) &:= \frac{u(x+h) - y}{h} - f(x, y) = O(h) \quad (h \rightarrow 0) \\ \text{Global DE} \quad e_h(x+h) &:= u(x+h) - u_h(x+h) \\ &= u(x+h) - u_h(x) - hf(x, u_h(x)) \\ &= \underbrace{u(x+h) - u(x) - hf(x, u(x))}_{= h \times \text{local DE}} + \underbrace{u(x) - u_h(x) + h\{f(x, u(x)) - f(x, u_h(x))\}}_{= \{1+hf_y(x, \dots)\}(u(x)-u_h(x)) = \text{error transport}} \end{aligned}$$

$$\text{Global DE } e_h(x+h) = h \cdot \tau(x, u(x), h) + \underbrace{\{1 + hf_y(x, \dots)\}}_{\text{condition number}} e_h(x)$$

Choice of the step size: rule of thumb $hL < 1$ (Lipschitz constant)

$$\begin{aligned} \text{Error bound } |e_h(x)| &\leq \underbrace{\frac{e^{(b-a)L} - 1}{L}}_{\text{condition number}} \cdot \underbrace{M h^p}_{= \text{local DE}} \quad (\text{roughly}) \\ &\longrightarrow \text{Error estimates} \end{aligned}$$

Important properties

Consistency approximation becomes better for smaller h ,
i.e., local DE $\rightarrow 0$ if $h \rightarrow 0$

Order of consistency local DE $= O(h^p)$ (p the largest possible number)

Convergence global DE $\rightarrow 0$ if $h \rightarrow 0$
step size series $h_0 > h_1 > h_2 > \dots$
series of approximations $u_{h_0}(x), u_{h_1}(x), \dots \rightarrow u(x)$

Stability insensibility against perturbations

Example: 2-step method $y_{m+2} - 4y_{m+1} + 3y_m = -2hf(x_{m+1}, y_{m+1})$
order of consistency $p = 2$

Special IVP: $y' = 0, y(0) = 0$: solution $u = 0$

Method: $y_{m+2} - 4y_{m+1} + 3y_m = 0, m = 0, 1, 2, \dots$
homogeneous difference equation

Starting values: $y_0 = 0, y_1 = 0 \curvearrowright y_2 = 0, y_3 = 0, \dots$

Perturbation $y_0 = 0, y_1 = \varepsilon (\neq 0)$:
 $y_2 = 4\varepsilon, y_3 = 13\varepsilon, y_4 = 40\varepsilon, y_5 = 121\varepsilon, \dots$

generally $y_m = \frac{1}{2}(3^m - 1)\varepsilon \rightarrow \infty$ für $m \rightarrow \infty$

Proof: $\frac{1}{2}(3^{m+2} - 1) - \frac{4}{2}(3^{m+1} - 1) + \frac{3}{2}(3^m - 1) = \frac{3^m}{2}(3^2 - 4 \cdot 3 + 3) = 0$

Large sensibility against small perturbations implies **instability!**

Characteristic polynomial $\rho(t) = t^2 - 4t + 3$, zeros 1 and 3 (> 1)!

14 Discretization Methods

History Euler 1768, Cauchy 1840,

Adams 1883, Runge 1895, Kutta 1901

[New stimulation beginning at 1960 by computers!]

IVP: $y' = f(x, y), y(x_0) = y_0, \quad f : D \rightarrow \mathbb{R}$ Lipschitz continuous (w.r.t. y)
 $D = I \times \mathbb{R}$ sector (rectangle), $I = [x_0, b]$

Solution $u : I \rightarrow \mathbb{R}$

Notation: $F_k(I)$ the set of f having all partial derivatives up to the order k existing on D , being continuous and bounded there.
 $f \in F_1(I)$ implies that f is Lipschitz continuous.

Wanted: Approximation u_h defined on the grid I_h

Uniformly spaced grid $h = \frac{b-x_0}{N}$

Algorithmic notation $y_m \equiv u_h(x_m)$

One-step method $u_h(x+h) = u_h(x) + h\phi(x, u_h(x), h), \quad x \in I_h$

Increment function $\phi = \phi(x, y, h)$

$$\phi : D \times [0, h_0] \rightarrow \mathbb{R} \quad (h_0 > 0)$$

Algorithm $y_{m+1} := y_m + h\phi(x_m, y_m, h), \quad m = 0, 1, \dots, N-1$

Examples

Euler polygon method $\phi(x, y, h) := f(x, y)$

Half-step method $\phi(x, y, h) := f(x + \frac{h}{2}, y + \frac{h}{2}f(x, y))$

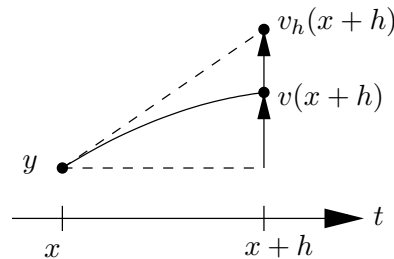
Improved polygon method $\phi(x, y, h) := \frac{1}{2}f(x, y) + \frac{1}{2}f(x+h, y+hf(x, y))$

Combination of gradients (nonlinear) $\rightarrow$ **Runge-Kutta methods**

Consideration of one step

IVP $v'(t) = f(t, v(t)), \quad \boxed{v(x) = y}$

Initial value possible for each $(x, y) \in D$



Increment function $\phi(x, y, h) := \frac{v_h(x+h)-y}{h}$

Exact relative increment $\Delta(x, y, h) := \begin{cases} \frac{v(x+h)-y}{h}, & h \neq 0 \\ f(x, y), & h = 0 \end{cases}$
 ϕ approximates Δ

Local DE $\tau(x, y, h) := \Delta(x, y, h) - \phi(x, y, h)$

Consistency $\tau(x, y, h) = O(h)$ ($h \rightarrow 0$) f.a. $(x, y) \in D$ and f.a. $f \in F_1(I)$

Definition: A one-step method has *order of consistency* p , if it satisfies
 $\tau(x, y, h) = O(h^p)$ ($h \rightarrow 0$) f.a. $(x, y) \in D$ and f.a. $f \in F_p(I)$.

$$\text{Euler polygon method } \tau(x, y, h) = \begin{cases} \frac{v(x+h)-y}{h} - f(x, y), & h \neq 0 \\ 0, & h = 0 \end{cases}$$

$$\begin{aligned} \text{Taylor } \tau(x, y, h) &= \frac{1}{h}\{v(x) + h v'(x) + \frac{h^2}{2}v''(x) + \cdots - y\} - f(x, y) \\ &= \frac{h}{2}v''(x) + \cdots = O(h) \quad (h \rightarrow 0) \quad \text{f.a. } (x, y) \in D \end{aligned}$$

Half-step method $\tau(x, y, h) = O(h^2)$

$$\Delta(x, y, h) = \frac{v(x+h)-y}{h} = v'(x) + \frac{h}{2}v''(x) + \cdots = f(x, y) + \frac{h}{2}\{f_x + f f_y\}_{(x,y)} + O(h^2)$$

$$\phi(x, y, h) = f(x + \frac{1}{2}h, y + \frac{1}{2}hf(x, y)) = f(x, y) + \frac{h}{2}\{f_x + f f_y\}_{(x,y)} + O(h^2)$$

implies order of consistency $p = 2$

Linear multistep methods

$$\sum_{j=0}^k \alpha_j y_{m+j} = h \sum_{j=0}^h \beta_j f_{m+j}, \quad f_{m+j} := f(x_{m+j}, y_{m+j}), \quad y'_{m+j} \equiv f_{m+k}$$

Real parameters $\alpha_0, \dots, \alpha_k, \beta_0, \dots, \beta_k$; $\alpha_k = 1, |\alpha_0| + |\beta_0| > 0$
 (the reduction to $k-1$ steps is not possible)

Explicit: $\beta_k = 0$

Implicit: $\beta_k \neq 0$ iteration $\curvearrowright$ **Predictor–Corrector method**

Abbreviation

$$\text{Characteristic polynomials} \quad \boxed{\rho(\xi) := \sum_{j=0}^k \alpha_j \xi^j, \quad \sigma(\xi) := \sum_{j=0}^k \beta_j \xi^j}$$

$$\text{Shift operator } E \quad (j \in \mathbb{N}) : \quad \begin{array}{l|l} Ey_m := y_{m+1} & Eu(x) := u(x+h) \\ E^j y_m := y_{m+j} & E^j u(x) := u(x+jh) \end{array}$$

$$\text{Linear } k\text{-step method } (\rho, \sigma) : \quad \boxed{\rho(E)y_m = h\sigma(E)y'_m}$$

Consideration of one step (without error transport):

Exact values $u(x), u(x+h), \dots, u(x+(k-1)h) \rightarrow \hat{u}_h(x+kh)$

Linear difference operator $L_h : C^1 \rightarrow C$

$$\begin{aligned} L_h(u(\cdot)) &:= \sum_{j=0}^k \alpha_j u(\cdot + jh) - h \sum_{j=0}^k \beta_j u'(\cdot + jh) \\ &= u(\cdot + kh) - \underbrace{\left\{ - \sum_{j=0}^{k-1} \alpha_j u(\cdot + jh) + h \sum_{j=0}^k \beta_j u'(\cdot + jh) \right\}}_{=:\hat{u}_h(\cdot + kh)} \end{aligned}$$

Taylor evaluation $f \in F_p(I)$, i.e. $u \in C^{p+1}(I)$

$$\begin{aligned} u(x + jh) &= u(x) + jhu'(x) + \cdots + \frac{j^p}{p!} h^p u^{(p)}(x) + O(h^{p+1}) \quad (h \rightarrow 0) \\ u'(x + jh) &= u'(x) + jhu''(x) + \cdots + \frac{j^{p-1}}{(p-1)!} h^{p-1} u^{(p)}(x) + O(h^p) \quad (h \rightarrow 0) \\ L_h(u(x)) &= C_0 u(x) + C_1 h u'(x) + \cdots + C_p h^p u^{(p)}(x) + O(h^{p+1}) \quad (h \rightarrow 0) \\ \text{with } C_0 &= \sum_{j=0}^k \alpha_j, \quad C_\nu = \sum_{j=0}^k \left(\alpha_j \frac{j^\nu}{\nu!} - \beta_j \frac{j^{\nu-1}}{(\nu-1)!} \right), \quad \nu = 1, \dots, p \end{aligned}$$

Local discretization error

IVP $(x, y) \in D : v'(t) = f(t, v(t)), v(x) = y$, solution v

Local DE $\tau(x, y, h) := \frac{1}{h} \rho(E)v(x) - \sigma(E)v'(x) = \frac{1}{h} L_h(v(x))$

To emphasize the step number k use $\tau_k := \tau_k(x, y, h)$

Definition A linear multistep method is called *consistent*, if

$\tau(x, y, h) = O(h)$ ($h \rightarrow 0$) f.a. $(x, y) \in D$ and f.a. $f \in F_1(I)$.

Condition of consistency $C_0 = C_1 = 0$, i.e., $\boxed{\rho(1) = 0, \rho'(1) = \sigma(1)}$

Definition A linear multistep method has *order of consistency* p , if

$\tau(x, y, h) = O(h^p)$ ($h \rightarrow 0$) f.a. $(x, y) \in D$ and f.a. $f \in F_p(I)$.

Order of consistency $p : C_0 = \cdots = C_p = 0 \curvearrowright \boxed{\text{linear system}}$

$$\begin{aligned} 0 &: \sum_{j=0}^k \alpha_j = 0 \\ 1 &: \sum_{j=0}^k (\alpha_j \cdot j - \beta_j) = 0 \\ \nu &: \sum_{j=0}^k \left(\alpha_j \frac{j^\nu}{\nu!} - \beta_j \frac{j^{\nu-1}}{(\nu-1)!} \right) = 0, \quad \nu = 2, \dots, p \quad (\alpha_k = 1) \end{aligned}$$

$$\left. \begin{array}{l} \text{Number of coefficients: } 2k+1 \\ \text{Number of equations: } p+1 \end{array} \right\} \Rightarrow$$

$$\text{maximal order of consistency } p^*(k) = 2k$$

Examples

$$y_{m+2} - y_m = 2hf_{m+1}, \quad p = 2$$

$$y_{m+2} - 4y_{m+1} + 3y_m = -2hf_m, \quad p = 2$$

$$y_{m+2} - y_{m+1} = \frac{1}{2}h(3f_{m+1} - f_m), \quad p = 2$$

$$y_{m+1} - y_m = \frac{1}{2}h(f_m + f_{m+1}), \quad k = 1, \quad p = 2 \quad (\text{trapezoidal rule})$$

Consistent starting values $y_0, y_1, \dots, y_{k-1}$

$$\text{IVP } v'(t) = f(t, v(t)), v(x) = y$$

Definition The starting values $y_0^h, \dots, y_{k-1}^h$ w.r.t. $x, x+h, \dots, x+(k-1)h$ are called *consistent*, if $\max_{0 \leq j \leq k-1} |y_j^h - v(x+jh)| = O(h)$ ($h \rightarrow 0$) f.a. $(x, y) \in D$.

Remarks

Large approximation errors of the starting values cannot be repaired again!
The starting values must have the same approximation order as the chosen discretization method!

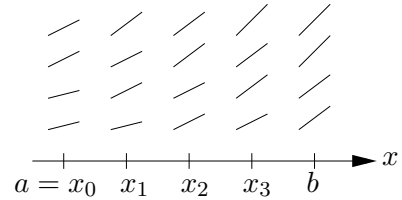
15 Runge–Kutta Methods

IVP $y' = f(x, y), y(a) = y_0$ ($f \in F_k(I)$, k sufficiently large)

solution u in I

Wanted approximation u_h in I_h

Method $u_h(x+h) = u_h(x) + h\phi(x, u_h(x), h)$



Marching along the vector field:

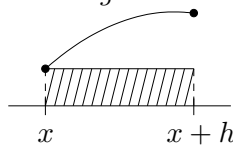
Mean value of gradients, intermediate approximations, Taylor evaluation

Integral equation
$$\underbrace{\frac{u(x+h) - u(x)}{h}} = \frac{1}{h} \int_x^{x+h} f(t, u(t)) dt = \frac{1}{h} \underbrace{\int_x^{x+h} u'(t) dt}_{\underbrace{AP \curvearrowright \boxed{QF}}_{\phi(x,y,h)+\text{approx.error}}}$$

$$\Delta(x, y, h) = u'(x) + \frac{h}{2}u''(x) + \dots$$

Rectangular rule
$$\frac{1}{h} \int_x^{x+h} u'(t) dt = f(x, u(x)) + O(h) \quad \text{Euler polygon}$$

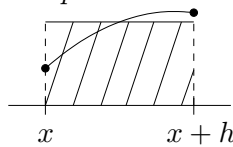
$$= f(x+h, u(x+h)) + O(h) \quad \text{implicit Euler}$$



Midpoint rule
$$\frac{1}{h} \int_x^{x+h} u'(t) dt = f(x + \frac{h}{2}, u(x + \frac{h}{2})) + O(h^2) \quad \text{explicit/implicit}$$

Intermediate approx.: $u(x + \frac{h}{2}) = u(x) + \frac{h}{2}f(x, u(x)) + O(h^2)$

$\phi(x, u(x), h) = f(x + \frac{h}{2}, u(x) + \frac{h}{2}f(x, u(x)))$ half-step method



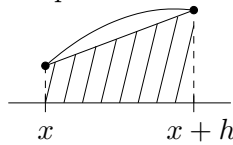
Trapezoidal rule
$$\frac{1}{h} \int_x^{x+h} u'(t) dt = \frac{1}{2}\{f(x, u(x)) + f(x+h, u(x+h))\} + O(h^2)$$

implicit trapezoidal rule $y_{m+1} - y_m = \frac{h}{2}(f_m + f_{m+1})$

intermediate approx.: $u(x+h) = u(x) + hf(x, u(x))$

$\phi(x, u(x), h) = \frac{1}{2}\{f(x, u(x)) + f(x+h, u(x) + hf(x, u(x)))\}$

improved Euler polygon method (explicit)



Algorithm: *Gradients by recurrence* (2-stage method)

$$\begin{aligned}
 U'_1 &:= f(x_m, y_m) \\
 U_2 &:= y_m + ha_{21}U'_1 \\
 U'_2 &:= f(x_m + c_2h, U_2) \\
 y_{m+1} &:= y_m + h(b_1U'_1 + b_2U'_2)
 \end{aligned}
 \quad \text{scheme} \quad \left(\begin{array}{c|cc|c} 0 & 0 & 0 & b_1 \\ c_2 & a_{21} & 0 & b_2 \end{array} \right)$$

Coefficients: $a_{21} = c_2 = \frac{1}{2}$, $b_1 = 0$, $b_2 = 1$ and $a_{21} = c_2 = 1$, $b_1 = b_2 = \frac{1}{2}$

Quadrature formula: Nodes $(0, c_2)$, weights (b_1, b_2)

Pure quadrature formula if $y' = f(x)$

Taylor evaluation

$$\begin{aligned}
 \Delta(x, u(x), h) &= u'(x) + \frac{h}{2}u''(x) + \frac{h^2}{3!}u'''(x) + \dots \quad (u(x) = y) \\
 &= f(x, y) + \frac{h}{2}\{f_x + ff_y\}_{(x,y)} + \frac{h^2}{3!}\{\dots\} + \dots \\
 U'_1 &= f(x, y) \\
 U'_2 &= f(x + c_2h, y + ha_{21}f(x, y)) = f(x, y) + \{c_2f_x + a_{21}ff_y\}_{(x,y)} + \dots \\
 \phi(x, u(x), h) &= b_1U'_1 + b_2U'_2 \\
 &= (b_1 + b_2)f(x, y) + hb_2\{c_2f_x + a_{21}ff_y\}_{(x,y)} + \dots \\
 \tau(x, u(x), h) &= \Delta(x, u(x), h) - \phi(x, u(x), h) \\
 &= \underbrace{(1 - b_1 - b_2)}_{=0}f(x, y) + h[\underbrace{(\frac{1}{2} - b_2c_2)}_{=0}f_x + \underbrace{(\frac{1}{2} - b_2a_{21})}_{=0}ff_y]_{(x,y)} + \dots \\
 &\quad \text{general assumption } \swarrow \boxed{c_2 = a_{21}} \quad (b_2 \neq 0)
 \end{aligned}$$

$$\begin{array}{lll}
 \text{Order of consistency} & p = 1 & \boxed{b_1 + b_2 = 1} \quad (\Sigma \text{ weights} = 1) \\
 & p = 2 & b_2c_2 = \frac{1}{2}, \quad c_2 = a_{21}
 \end{array}$$

Solution: $c_2 \neq 0$, $b_2 \neq 0$ (otherwise only one stage)

$$b_1 = 1 - \frac{1}{2c_2}, \quad b_2 = \frac{1}{2c_2} \quad 1 \text{ independent parameter } s = p = 2$$

half-step method: $c_2 = \frac{1}{2}$, improved Euler method: $c_2 = 1$

Implicit trapezoidal rule $s = p = 2$

$$U'_1 := f(x_m, y_m)$$

$$U'_2 := f(x_m + h, y_m + \frac{h}{2}(U'_1 + U'_2))$$

$$y_{m+1} := y_m + \frac{h}{2}(U'_1 + U'_2)$$

$$y_{m+1} - y_m = \frac{h}{2}(f_m + f_{m+1})$$

The implicit trapezoidal rule is a Runge-Kutta method
and a linear 1-step method!

s-stage Runge-Kutta method

$$U'_j := f(x_m + c_j h, U_j)$$

$$U_j := y_m + h \sum_{k=1}^s a_{jk} U'_k \quad (j = 1, \dots, s)$$

$$y_{m+1} := y_m + h \sum_{j=1}^s b_j U'_j$$

Generating matrix

$$(c, A, b) = \left(\begin{array}{c|ccc|c} c_1 & a_{11} & \dots & a_{1s} & b_1 \\ \vdots & \vdots & & \vdots & \vdots \\ c_s & a_{s1} & \dots & a_{ss} & b_s \end{array} \right) \quad \text{assumption } c_j = \sum a_{jk}$$

Explicit: $a_{jk} = 0$ für $j \leq k \leadsto U'_j$ by recurrence

Implicit: otherwise $\leadsto$ nonlinear system of equations for U_j, U'_j .

Quadrature formulas (c, b)

Examples: Euler $(0, 0, 1)$, implicit Euler $(1, 1, 1)$

Heun method $s = p = 3$:

$$\left(\begin{array}{c|ccc|c} 0 & 0 & 0 & 0 & 1/4 \\ 1/3 & 1/3 & 0 & 0 & 0 \\ 2/3 & 0 & 2/3 & 0 & 3/4 \end{array} \right) \quad \begin{array}{c} U'_3 \\ U'_2 \\ U'_1 \end{array}$$

Radau quadrature formula

$$\int_0^1 g(x) dx = \frac{1}{4} \{g(0) + 3g(\frac{2}{3})\} + Rg, \quad \text{degree of exactness is two}$$

Classical Runge-Kutta method $s = p = 4$

$$\left(\begin{array}{c|ccc|c} 0 & 0 & & & 1/6 \\ 1/2 & 1/2 & 0 & & 1/3 \\ 1/2 & 0 & 1/2 & 0 & 1/3 \\ 1 & 0 & 0 & 1 & 1/6 \end{array} \right) \quad \begin{array}{c} U'_4 \\ U'_3 \\ U'_2 \\ U'_1 \end{array}$$

$$\left. \begin{array}{l} U'_1 = f(x_m, y_m), \quad U_2 = y_m + \frac{h}{2}U'_1, \quad p = 1 \\ U'_2 = f(x_m + \frac{h}{2}, U_2), \quad U_3 = y_m + \frac{h}{2}U'_2, \quad p = 1 \\ U'_3 = f(x_m + \frac{h}{2}, U_3), \quad U_4 = y_m + hU'_3, \quad p = 2 \\ U'_4 = f(x_m + h, U_4), \quad y_{m+1} := y_m + \frac{h}{6}(U'_1 + 2U'_2 + 2U'_3 + U'_4) \end{array} \right\} \text{error compensation}$$

Theorem 15.1 A Runge-Kutta method is consistent if and only if $\sum b_j = 1$.

Proof $\tau(x, y, h) = \{1 - \sum b_j\}f(x, y) + O(h) \quad (h \rightarrow 0). \quad \square$

Theorem 15.2 If the Runge-Kutta method (c, A, b) has order of consistency p , then the basic quadrature formula (c, b) has at least the degree of exactness $p - 1$.

Proof IVP $y' = g(x), \quad y(0) = y_0$

$$\Delta(x, y, h) = \frac{1}{h} \int_x^{x+h} g(t) dt = g(x) + \frac{h}{2}g'(x) + \cdots + \frac{h^{p-1}}{p!}g^{(p-1)}(x) + O(h^p)$$

$$\phi(x, y, h) = \sum b_j g(x + c_j h) = \sum b_j \{g(x) + c_j h g'(x) + \cdots + c_j^{p-1} \frac{h^{p-1}}{(p-1)!} g^{(p-1)}(x)\} + O(h^p)$$

$$\tau(x, y, h) = (1 - \sum b_j)g(x) + h(\frac{1}{2} - \sum b_j c_j)g'(x) + \cdots + \frac{h^{p-1}}{(p-1)!}(\frac{1}{p} - \sum b_j c_j^{p-1})g^{(p-1)}(x) + O(h^p)$$

$$\int_0^1 g(t) dt = \sum b_j g(c_j) \text{ f.a. } g \in \mathcal{P}_{p-1} \iff \sum b_j c_j^\nu = \frac{1}{\nu+1}, \quad \nu = 0, 1, \dots, p-1. \quad \square$$

Construction of (explicit) Runge-Kutta methods

Taylor evaluation

$$\left. \begin{array}{l} \Delta(x, y, h) = \dots \\ \phi(x, y, h) = \dots \end{array} \right\} \curvearrowright \tau(x, y, h) = \dots$$

$$\text{Operator} \quad D := \frac{\partial}{\partial x} + f \frac{\partial}{\partial y}, \quad D^j := (\frac{\partial}{\partial x} + f \frac{\partial}{\partial y})^j \quad \text{symbolic power}$$

$$D^2 f = f_{xx} + 2f f_{xy} + f^2 f_{yy}$$

$$D(Df) \neq D^2 f$$

$$D(Df) = D^2 f + f y Df$$

Assume $u(x) = y$:

$$\Delta(x, y, h) = \frac{u(x+h)-y}{h} = u'(x) + \frac{h}{2}u''(x) + \frac{h^2}{3!}u'''(x) + \frac{h^3}{4!}u^{(4)}(x) + \dots$$

Derivatives of u and partial derivatives of f :

$$u' = f$$

$$u'' = Df = f_x + f f_y$$

$$u''' = D(Df) = f_{xx} + f f_{xy} + f_x f_y + f f_{xy} + f(f_y)^2 + f^2 f_{yy} \quad (6 \text{ partial derivatives})$$

$$= D^2 f + f_y Df$$

$$u^{(4)} = D(D(Df)) = f_{xxx} + f f_{xxy} + \dots \quad (13 \text{ partial derivatives})$$

$$= D^3 f + f_y D^2 f + f_y^2 Df + 3Df Df_y \quad (4 \text{ elementary differentials})$$

$\vdots$

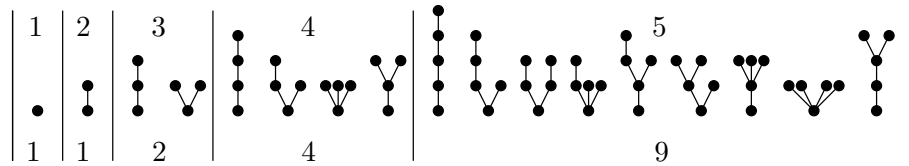
Elementary differentials: f ; Df ; $D^2 f$, $f_y Df$; $D^3 f$, $f_y D^2 f$, $f_y^2 Df$, $Df Df_y$; $\dots$

Elementary differentials: putting certain partial derivatives together!

Each elementary differential implies one equation for the coefficients!

The number of elementary differentials increases like **trees** branch out

$\curvearrowright$ *recurrence formulas*



Unique mapping $\rightarrow$ **rooted trees** (see Butcher, Hairer–Norsett–Wanner)

order of the elementary differentials	1	2	3	4	5	6	7	8
number of the elementary differentials	1	1	2	4	9	20	48	115

order of consistency p	1	2	3	4	5	6	7	8	9	10
number of condition equations	1	2	4	8	17	37	85	200	486	1205

Evaluation of Δ

$$\Delta(x, y, h) = f + \frac{h}{2}Df + \frac{h^2}{3!}\{D^2 f + f_y Df\} + \frac{h^3}{4!}\{D^3 f + f_y D^2 f + f_y^2 Df + 3Df Df_y\} + \dots |_{(x,y)}$$

Evaluation of Φ (evaluation of the U'_j)

$$f(x + \alpha, y + \beta) = f + \{\alpha f_x + \beta f_y\} + \frac{1}{2}\{\alpha^2 f_{xx} + 2\alpha\beta f_{xy} + \beta^2 f_{yy}\} + \dots |_{(x,y)}$$

Note the assumption $c_j = \Sigma a_{jk}$

Evaluation of the gradients

$$U'_1 = f(x, y)$$

$$\begin{aligned} U'_2 &= f(x + c_2 h, y + h a_{21} U'_1) \\ &= f + h c_2 Df + \frac{h^2}{2} c_2^2 D^2 f + \frac{h^3}{3!} c_2^3 D^3 f + \dots |_{(x,y)} \end{aligned}$$

$$\begin{aligned} U'_3 &= f(x + c_3 h, y + h(a_{31} U'_1 + a_{32} U'_2)) \\ &= f + h c_3 Df + h^2 \left\{ \frac{1}{2} c_3^2 D^2 f + a_{32} c_2 f_y Df \right\} \\ &\quad + h^3 \left\{ \frac{1}{3!} c_3^3 D^3 f + \frac{1}{2} a_{32} c_2^2 f_y D^2 f + a_{32} c_2 c_3 Df Df_y \right\} + \dots |_{(x,y)} \end{aligned}$$

$$\begin{aligned} U'_4 &= f(x + c_4 h, y + h(a_{41} U'_1 + a_{42} U'_2 + a_{43} U'_3)) \\ &= f + h c_4 Df + h^2 \left\{ \frac{1}{2} c_4^2 D^2 f + (a_{42} c_2 + a_{43} c_3) f_y Df \right\} \\ &\quad + h^3 \left\{ \frac{1}{3!} c_4^3 D^3 f + \frac{1}{2} (a_{42} c_2^2 + a_{43} c_3^2) f_y D^2 f + a_{43} a_{32} c_2 f_y^2 Df \right. \\ &\quad \left. + (a_{42} c_2 + a_{43} c_3) c_4 Df Df_y \right\} + \dots |_{(x,y)} \end{aligned}$$

$$\vdots$$

$$\begin{aligned} \phi(x, y, h) &= \Sigma b_j U'_j \\ &= \Sigma b_j f + h \Sigma b_j c_j \cdot Df + h^2 \left\{ \frac{1}{2} \Sigma b_j c_j^2 \cdot D^2 f + \Sigma b_j \Sigma a_{j\nu} c_\nu \cdot f_y Df \right\} + \\ &\quad + h^3 \left\{ \frac{1}{3!} \Sigma b_j c_j^3 \cdot D^3 f + \dots f_y D^2 f + \dots f_y^2 Df + \dots Df Df_y \right\} + \dots |_{(x,y)} \end{aligned}$$

Taylor evaluation of $\tau = \Delta - \phi$: $\tau(x, y, h) = O(h^p)$ ($h \rightarrow 0$)

$$p = 1 : \quad \Sigma b_j = 1 \quad \text{q.f. with degree of exactness 0}$$

$$p = 2 : \quad \Sigma b_j c_j = \frac{1}{2} \quad \text{q.f. with degree of exactness 1}$$

$$p = 3 : \quad \Sigma b_j c_j^2 = \frac{1}{3} \quad \text{q.f. with degree of exactness 2}$$

$$\Sigma \Sigma b_j a_{j\nu} c_\nu = \frac{1}{6}$$

$$p = 4 : \quad \Sigma b_j c_j^3 = \frac{1}{4} \quad \text{q.f. with degree of exactness 3}$$

$$\dots$$

$$\dots$$

$$\dots$$

nonlinear system of equations

Explicit s -stage Runge-Kutta method

$$\left(\begin{array}{c|ccc|c} 0 & 0 & \dots & 0 & b_1 \\ c_2 & a_{21} & 0 & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_s & a_{s1} & \dots & a_{s,s-1} & 0 & b_s \end{array} \right) \quad c_j = \sum_{k=1}^{j-1} a_{jk}$$

Number of coefficients a_{jk}, b_j : $A(s) = \frac{1}{2}s(s+1)$

Maximal order of consistency for s stages $p^*(s)$

s	1	2	3	4	5	6	7	8	9	10	11	17	18
$p^*(s)$	1	2	3	4	4	5	6	6	7	7	8	10	10
$A(s)$	1	3	6	10	15	21	28	36	45	55	66	153	171
$B(p^*)$	1	2	4	8	8	17	37	37	85	85	200	1205	1205

Remarks

One method with $s = 11$ and $p = 8$ (Curtis 1970, Cooper–Verner 1972).

An explicit method with $s = 10$ and $p = 8$ does not exist (Butcher 1985).

Methods of highest order of consistency $p = 10$: $s = 18$ (Curtis 1975)

and $s = 17$ (Hairer 1978).

Explicit Runge–Kutta methods of maximal order of consistency

$s = p = 1$: Euler polygon method

$s = p = 2$: $\left. \begin{array}{l} b_1 + b_2 = 1 \\ b_2 c_2 = \frac{1}{2} \end{array} \right\} b_1 = 1 - \frac{1}{2c_2}, b_2 = \frac{1}{2c_2} \ (c_2 \neq 0)$ half-step improved Euler

$p = 3$ is not possible!

$s = p = 3$: $\left. \begin{array}{l} b_1 + b_2 + b_3 = 1 \\ b_2 c_2 + b_3 c_3 = \frac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 = \frac{1}{3} \\ b_3 a_{32} c_2 = \frac{1}{6} \end{array} \right\}$ multitude of solutions

$$\text{Heun} \quad \left(\begin{array}{c|cc|c} 0 & 0 & & 1/4 \\ 1/3 & 1/3 & 0 & 0 \\ 2/3 & 0 & 2/3 & 3/4 \end{array} \right)$$

$$\text{Kutta} \quad \left(\begin{array}{c|cc|c} 0 & 0 & & 1/6 \\ 1/2 & 1/2 & 0 & 2/3 \\ 1 & -1 & 2 & 1/6 \end{array} \right)$$

$$s = p = 4 : \quad \left. \begin{array}{l} b_1 + b_2 + b_3 + b_4 = 1 \\ b_2 c_2 + b_3 c_3 + b_4 c_4 = \frac{1}{2} \\ b_2 c_2^2 + b_3 c_3^2 + b_4 c_4^2 = \frac{1}{3} \\ b_2 c_2^3 + b_3 c_3^3 + b_4 c_4^3 = \frac{1}{4} \\ b_3 a_{32} c_2 + b_4 (a_{42} c_2 + a_{43} c_3) = \frac{1}{6} \\ b_3 a_{32} c_2^2 + b_4 (a_{42} c_2^2 + a_{43} c_3^2) = \frac{1}{12} \\ b_3 a_{32} c_2 c_3 + b_4 (a_{42} c_2 + a_{43} c_3) c_4 = \frac{1}{8} \\ b_4 a_{43} a_{32} c_2 = \frac{1}{24} \end{array} \right\} \begin{array}{l} 10 \text{ coeff., } 8 \text{ equations} \\ \curvearrowright 2 \text{ independent parameters} \\ \text{Kopal '54, Butcher} \\ \text{Hairer–Norsett–Wanner} \end{array}$$

$$\text{Classical Runge-Kutta} \left(\begin{array}{c|ccc|c} 0 & 0 & & & 1/6 \\ 1/2 & 1/2 & 0 & & 1/3 \\ 1/2 & 0 & 1/2 & 0 & 1/3 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & & & & 1/6 \end{array} \right)$$

$$3/8\text{-Runge-Kutta} \left(\begin{array}{c|ccc|c} 0 & 0 & & & 1/8 \\ 1/3 & 1/3 & 0 & & 3/8 \\ 2/3 & -1/3 & 1 & 0 & 3/8 \\ 1 & 1 & -1 & 1 & 0 \\ \hline & & & & 1/8 \end{array} \right)$$

Outview

Implicit Runge-Kutta methods (c, A, b) (Butcher 1963)

$$\begin{aligned} U'_j &= f(x + c_j h, y + h \sum_{k=1}^s a_{jk} U'_k), \quad j = 1, \dots, s \\ u_h(x + h) &= u_h(x) + h \sum b_j U'_j \end{aligned}$$

Nonlinear system for the $U'_j \leadsto$ Newton iteration

Number of the coefficients $A(s) = s(s+1)$

Maximal order of consistency $p^*(s) = 2s$

The basic q.f. (c, b) is the Gauss-Legendre q.f. with degree of exactness $2s - 1$

Approximation order of the intermediate approximates U_j is $O(h^s)$

s	1	2	3	4
$p^*(s)$	2	4	6	8
$A(s)$	2	6	12	20
$B(p^*)$	2	8	37	200

Examples

$$s = 1 : (\frac{1}{2}, \frac{1}{2}, 1), \quad y_{m+1} := y_m + hU'_1, \quad U'_1 = f(x_m + \frac{h}{2}, y_m + \frac{h}{2}U'_1)$$

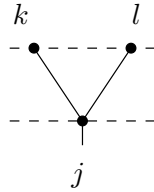
$$s = 2 : \left(\begin{array}{c|cc|c} \frac{1}{2} - \frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{4} - \frac{1}{6}\sqrt{3} & \frac{1}{2} \\ \frac{1}{2} + \frac{1}{6}\sqrt{3} & \frac{1}{4} + \frac{1}{6}\sqrt{3} & \frac{1}{4} & \frac{1}{2} \\ \hline \end{array} \right)$$

There exist also Radau-Runge-Kutta methods with $p(s) = 2s - 1$ and Lobatto-Runge-Kutta methods with $p(s) = 2s - 2$.

Tree theory

Graphic representation of the elementary differential (or order conditions)

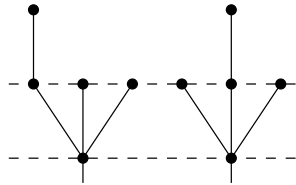
A tree consists of **vertices** and **branches**; the lowest vertex is called **root**; the vertices are noted by indices and the set of vertices is given by $\{j, k, \ell, \dots\}$; the branches represent the mappings $k \rightarrow j, \ell \rightarrow j, \dots$ (downward); the **order** of the tree is the number of elements of the set of vertices.



set of vertices $\{j, k, \ell\}$, root j
 set of branches $\{k \rightarrow j, \ell \rightarrow j\}$
 2 branches from the first floor to the root
 order of the tree is 3 (=number of vertices)

Equivalent trees

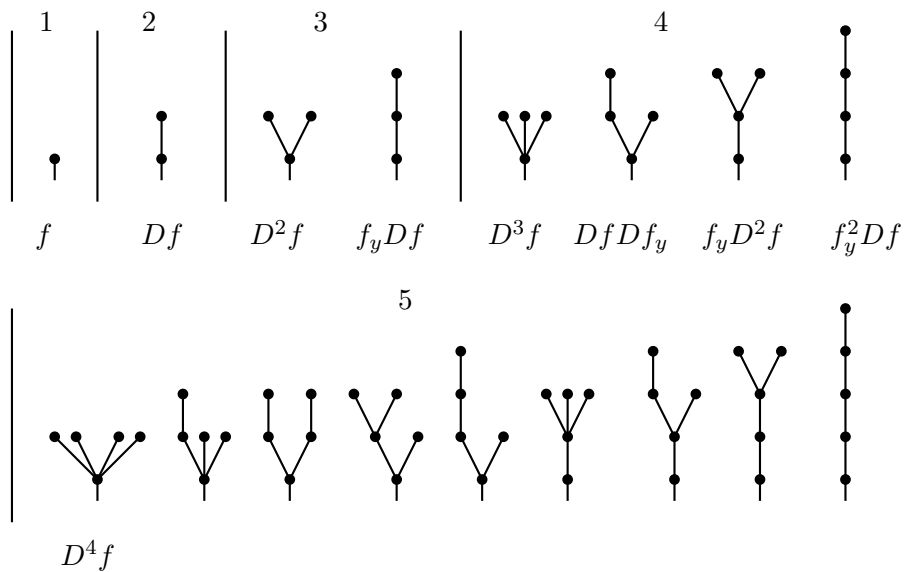
- 1) the same order
- 2) equivalent set of branches

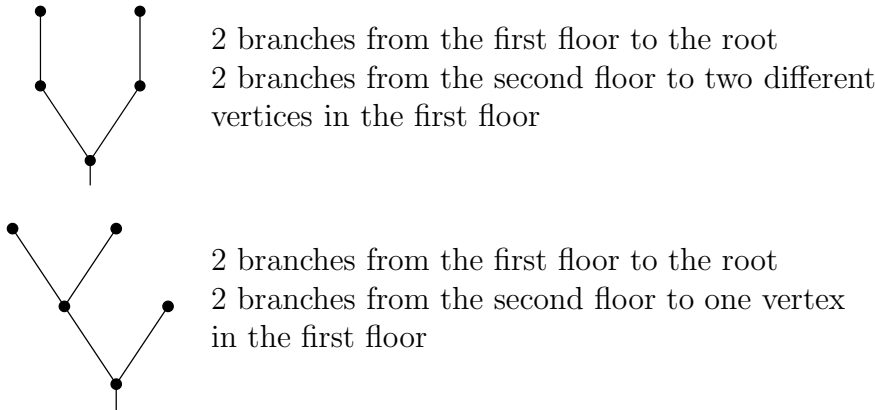


the same order
 3 branches from the first floor to the root
 1 branch from the second to the first floor
 (no matter to which vertex)

Definition by recurrence: order $q \rightarrow q + 1$

- 1) root remains root
- 2) attach to an arbitrary vertex exactly one additional branch upward, i.e., the new tree has one more vertex
- 3) do it for each vertex
- 4) equivalent trees are classified as one tree





Stepsize control for one-step methods (estimates not bounds!)

Extrapolation (see section 6): Use one discretization method and compute two approximations u_h and $u_{\frac{h}{2}}$ with stepsize h and $\frac{h}{2}$.

The main part of the discretization error (for u_h) $c_p h^{p+1}$ is estimated by $\frac{2^p}{2^p-1} \cdot |u_h(x+h) - u_{\frac{h}{2}}(x+h)|$

Embedded methods: Use two methods of order p and $p+1$:

$$u(x+h) = u_h(x+h) + ch^{p+1} + O(h^{p+2})$$

$$u(x+h) = \hat{u}_h(x+h) + O(h^{p+2})$$

The main part of the discretization error ch^{p+1} (method of order p) is estimated by

$$|u_h(x+h) - \hat{u}_h(x+h)|$$

Using two completely distinct methods is very expensive. Find pairs of Runge-Kutta $p(p+1)$ methods having order p and $p+1$ such that the generating matrix of the lower order method is a subset of the generating matrix of the higher order method; then they are called *embedded methods*.

Examples

$$RKM\ 1(2) : (c, A, b, \hat{b}) = \left(\begin{array}{c|cc|c|c} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{array} \middle| \begin{array}{c} \frac{1}{2} \\ \frac{1}{2} \end{array} \right)$$

$$RKM\ 2(3) : (c, A, b, \hat{b}) = \left(\begin{array}{c|ccc|c|c} 0 & 0 & 0 & 0 & 0 & \frac{214}{891} & \frac{533}{2106} \\ \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{33} & 0 \\ \frac{27}{4} & -\frac{189}{800} & \frac{729}{800} & 0 & 0 & \frac{650}{891} & \frac{800}{1053} \\ 1 & \frac{214}{891} & \frac{1}{33} & \frac{650}{891} & 0 & 0 & -\frac{1}{78} \end{array} \right)$$

Famous method DOPRI5 (Dormand & Prince 5(4)): the error of the 5-order method is estimated (see Hairer et al.).

16 Linear Multistep Methods of Adams

k-step methods $\rho(E)y_m = h\sigma(E)f_m$

Consistency $\rho(1) = 0, \rho'(1) = \sigma(1)$

Order of consistency $p : \sum \left(\alpha_j \frac{j^\nu}{\nu!} - \beta_j \frac{j^{\nu-1}}{(\nu-1)!} \right) = 0, \quad \nu = 2, \dots, p$

Maximal order of consistency $p^*(k) = 2k$

Linear multistep methods are easier to handle than Runge–Kutta methods!

Methods of Adams

Construction (interpolation of Gregory–Newton)

$$\begin{aligned} \text{Integral equation } u(x_{m+k}) - u(x_{m+k-1}) &= \int_{x_{m+k-1}}^{x_{m+k}} \underbrace{f(x, u(x))}_{=P(x)+\text{error}} dx \\ y_{m+k} - y_{m+k-1} &= \int_{x_{m+k-1}}^{x_{m+k}} P(x) dx = h \sum \beta_j f_{m+j} \end{aligned}$$

Approximation of $f(x, u(x))$ by the interpolation polynomial $P(x)$

w.r.t. the nodes $x_m, x_{m+1}, \dots, x_{m+k-1}$ and the data $f_m, f_{m+1}, \dots, f_{m+k-1}$

→ explicit methods

or w.r.t. the nodes $x_m, \dots, x_{m+k}$ and the data $f_m, \dots, f_{m+k}$

→ implicit methods

Lagrange representation

$$P(x) = \sum \ell_j(x) f_{m+j} \curvearrowright y_{m+k} - y_{m+k-1} = h \sum \beta_j f_{m+j}, \beta_j = \frac{1}{h} \int_{x_{m+k-1}}^{x_{m+k}} \ell_j(x) dx$$

Gregory–Newton with backward differences (Section 1):

$$P(x) = \sum_{\nu=0}^{k-1} (-1)^\nu \binom{-t}{\nu} \nabla^\nu f_{m+k-1}, \quad t = \frac{x - x_{m+k-1}}{h} \rightarrow \text{explicit methods}$$

$$\nabla^\nu f_{m+i} := \nabla^{\nu-1} f_{m+i} - \nabla^{\nu-1} f_{m+i-1}$$

degree of $P = k - 1$

$$\int_{x_{m+k-1}}^{x_{m+k}} P(x) dx = h \sum_{\nu=0}^{k-1} \gamma_\nu \nabla^\nu f_{m+k-1}, \quad \gamma_\nu = (-1)^\nu \int_0^1 \binom{-t}{\nu} dt \quad (\text{independent of } k)$$

Methods of Adams–Bashforth (explicit)

$$y_{m+k} - y_{m+k-1} = h \sum_{\nu=0}^{k-1} \gamma_\nu \nabla^\nu f_{m+k-1}, \quad p = k$$

Coefficients γ_ν by recurrence relation

$$\gamma_\nu + \frac{1}{2}\gamma_{\nu-1} + \frac{1}{3}\gamma_{\nu-2} + \cdots + \frac{1}{\nu+1}\gamma_0 = 1, \quad \nu = 0, 1, 2, \dots$$

ν	0	1	2	3	4
γ_ν	1	$\frac{1}{2}$	$\frac{5}{12}$	$\frac{3}{8}$	$\frac{251}{720}$

$\{\gamma_\nu\} = \{0.5, 0.42, 0.38, 0.35, \dots\}$ monotone decreasing

Examples

$$\begin{aligned} k = p = 2: \quad y_{m+2} - y_{m+1} &= h(f_{m+1} + \frac{1}{2}\nabla f_{m+1}) \\ &= \frac{h}{2}(3f_{m+1} - f_m) \end{aligned}$$

$$\begin{aligned} k = p = 3: \quad y_{m+3} - y_{m+2} &= h(f_{m+2} + \frac{1}{2}\nabla f_{m+2}) + \frac{5}{12}h\nabla^2 f_{m+2} \\ &= \frac{h}{12}(23f_{m+2} - 16f_{m+1} + 5f_m) \end{aligned}$$

Local discretization error τ

k -step method: $u' = f(t, u(t)), u(x) = y \rightarrow \tau(x, y, h), \tau_k := \tau(x_m, y_m, h)$

$$d_{m+k} := h \times \tau_k = \int_{x_{m+k-1}}^{x_{m+k}} \underbrace{\{f(x, u(x)) - P(x)\}}_{\text{interpolation error}} dx$$

$$\begin{aligned} \underbrace{f(x, u(x)) - P(x)}_{= u'(x)} &= \frac{1}{k!} u^{(k+1)}(\xi(x)) \underbrace{(x - x_m) \cdots (x - x_{m+k-1})}_{\text{no change of sign in } (x_{m+k-1}, x_{m+k})} \\ &= u'(x) \end{aligned}$$

$$\begin{aligned} \text{Substitution } t = \frac{x - x_{m+k-1}}{h} &\curvearrowright \frac{1}{h^k} \frac{1}{k!} (x - x_m) \cdots (x - x_{m+k-1}) = (-1)^k \binom{-t}{k} \\ &\curvearrowright \gamma_k = (-1)^k \int_0^1 \binom{-t}{k} dt \end{aligned}$$

Theorem 16.1 *The Adams–Bashforth methods have the following representation and estimate of the local discretization error (main part):*

$d_{m+k} = h^{k+1} \gamma_k u^{(k+1)}(\xi), \quad x_m < \xi < x_{m+k-1}$
$d_{m+k} = h \gamma_k \nabla^k f_{m+k-1} + O(h^{k+2})$

Error estimate and increasing the order of consistency

$$y_{m+k} - y_{m+k-1} = h \underbrace{\sum_{\nu=0}^{k-1} \gamma_\nu \nabla^\nu f_{m+k-1}}_{p=k} + \underbrace{h \gamma_k \nabla^k f_{m+k-1}}_{\approx d_{m+k}} + h \gamma_{k+1} \nabla^{k+1} f_{m+k-1} + \dots$$

$\underbrace{\hspace{10em}}_{p=k+1}$
 $\underbrace{\hspace{15em}}_{p=k+2 \quad (k+2)\text{-step method}}$

Difference scheme (Section 1)

x_{m-2}	$\nabla^\circ f_{m-2}$	$\dots$	$\dots$	$\dots$	$\nabla^{k-1} f_{m+k-3}$	$\nabla^k f_{m+k-2}$	$\nabla^{k+1} f_{m+k-1}$
x_{m-1}	$\nabla^\circ f_{m-1}$	$\dots$	$\dots$	$\dots$	$\nabla^{k-1} f_{m+k-2}$	$\nabla^k f_{m+k-1}$	
x_m	$\nabla^\circ f_m$	$\nabla^1 f_{m+1}$	$\dots$	$\dots$	$\nabla^{k-1} f_{m+k-1}$		
x_{m+1}	$\nabla^\circ f_{m+1}$		$\dots$	$\dots$			
$\vdots$	$\vdots$	$\dots$	$\dots$				
x_{m+k-2}	$\nabla^\circ f_{m+k-2}$						
x_{m+k-1}	$\nabla^\circ f_{m+k-1}$	$\nabla^1 f_{m+k-1}$					

Efficient algorithm: Adams method with controlling of **order** and **step size**

Resolution of the backward differences yields the usual form

$$y_{m+k} - y_{m+k-1} = h \sum_{j=0}^{k-1} \beta_j^{(k)} f_{m+j},$$

j	0	1	2
$\beta_j^{(1)}$	1		
$\beta_j^{(2)}$	$-\frac{1}{2}$	$\frac{3}{2}$	
$\beta_j^{(3)}$	$\frac{5}{12}$	$-\frac{16}{12}$	$\frac{23}{12}$

Implicit methods: Interpolation polynomial w.r.t. $x_m, \dots, x_{m+k}$ and $f_m, \dots, f_{m+k}$

$$P(x) = \sum_{\nu=0}^k (-1)^\nu \binom{-t}{\nu} \nabla^\nu f_{m+k}, \quad t = \frac{x - x_{m+k}}{h}, \quad \text{degree of } P = k$$

Method of Adams–Moulton (implicit method)

$$y_{m+k} - y_{m+k-1} = h \sum_{\nu=0}^k \delta_\nu \nabla^\nu f_{m+k}, \quad p = k + 1$$

Coefficients $\delta_0 = 1, \delta_\nu + \frac{1}{2}\delta_{\nu-1} + \frac{1}{3}\delta_{\nu-2} + \cdots + \frac{1}{\nu+1}\delta_0 = 0, \quad \nu = 1, 2, 3, \dots$

ν	0	1	2	3	4	$\delta_\nu = \gamma_\nu - \gamma_{\nu-1}$
δ_ν	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	

Reduction to the usual form

$$y_{m+k} - y_{m+k-1} = h \sum_{j=0}^k \beta_j^{(k)} f_{m+j}$$

$$k = 1 : \quad y_{m+1} - y_m = \frac{h}{2}(f_{m+1} + f_m) = h(f_{m+1} - \frac{1}{2}\nabla f_{m+1})$$

Trapezoidal rule = Runge–Kutta method

Implementation of implicit methods

$$\begin{aligned} \text{Direct iteration} \quad y_{m+k}^{(\nu+1)} &:= W_{m+k-1} + h\beta_k f(x_{m+k}, y_{m+k}^{(\nu)}), \quad \nu = 0, 1, 2, \dots \\ W_{m+k-1} &:= h \sum_{j=0}^{k-1} \beta_j f_{m+j} \end{aligned}$$

Convergence, if $h|\beta_k|L < 1$ (L Lipschitz constant of f w.r.t. y)

In case of a suitable starting value $y_{m+k}^{(0)}$ only a few iteration steps are needed.

Predictor–Corrector methods

$$\left. \begin{array}{ll} \text{Predictor:} & \text{explicit methods} \rightarrow \text{starting value} \\ \text{Corrector:} & \text{implicit method} \rightarrow \text{iteration} \end{array} \right\} \text{suitable pair } P, C$$

Example $P : y_{m+1}^{(0)} = y_m + 2hf_m$

$$E : f_{m+1}^{(0)} = f(x_{m+1}, y_{m+1}^{(0)}) \quad \text{PEC method}$$

$$C : y_{m+1}^{(1)} = y_m + \frac{1}{2}h\{f_{m+1}^{(0)} + f_m\}$$

$$E : f_{m+1}^{(1)} = f(x_{m+1}, y_{m+1}^{(1)})$$

$$C : y_{m+1}^{(2)} = y_m + \frac{1}{2}h\{f_{m+1}^{(1)} + f_m\}$$

$$\rightarrow \text{PEC}, \underbrace{\text{PECEC} \dots \text{EC}}_{\ell \times}$$

P and C with order of consistency $p = 2$.

17 Asymptotic Stability and Convergence

The behaviour of discretization methods in the limit case $h \rightarrow 0$

Stability: Insensibility against perturbations if $h \rightarrow 0$

Convergence: Global discretization error $e_h(x) \rightarrow 0$ if $h \rightarrow 0$

One-step method

Approximation $u_h : u_h(x+h) = u_h(x) + h\phi(x, u_h(x), h), \text{ IV } u_h(x_0)$

Perturbed approx. $v_h : v_h(x+h) = v_h(x) + h[\phi(x, v_h(x), h) + r_h(x)], \text{ IV } v_h(x_0)$

Perturbation $\psi_h(v_h - u_h) := \max\{|v_h(x_0) - u_h(x_0)|, |r_h(x)|, x \in I_h\}$

Definition A one-step method is called *asymptotically stable*, if there exist positive numbers δ_0, h_0, K such that for each $\delta \in [0, \delta_0]$ the perturbed approximation v_h with

$$|\psi_h(v_h - u_h)| \leq \delta$$

satisfies uniformly for all $h \in [0, h_0]$ the condition

$$|v_h(x) - u_h(x)| \leq K \cdot \delta \quad \text{f.a. } x \in I_h.$$

ϕ **Lipschitz continuous**, i.e., $|\phi(x, y, h) - \phi(x, z, h)| \leq L|y - z|$
for all $(x, y, h), (x, z, h) \in D \times [0, h_0]$

Lemma 17.1 *The elements of the series $\{\xi_n\}_{n \geq 0}$ satisfy the inequality*

$$|\xi_{n+1}| \leq a|\xi_n| + b \quad \text{with } a, b \geq 0.$$

Then it follows

$$|\xi_n| \leq a^n |\xi_0| + \begin{cases} \frac{a^n - 1}{a - 1} b, & a \neq 1 \\ nb, & a = 1 \end{cases}.$$

Proof By induction

Theorem 17.2 *A one-step method is asymptotically stable, if the increment function ϕ is Lipschitz continuous.*

Proof

Assumption $\psi_h(v_h - u_h) \leq \delta$, i.e., $|v_h(x_0) - u(x_0)| \leq \delta, |r_h(x)| \leq \delta, x \in I_h$

Lipschitz constant $L > 0$

$$v_h(x+h) - u_h(x+h) = v_h(x) - u_h(x) + h[\phi(x, v_h(x), h) - \phi(x, u_h(x), h)] + hr_h(x)$$

$$|v_h(x+h) - u_h(x+h)| \leq (1+hL)|v_h(x) - u_h(x)| + h\delta$$

Lemma 17.1: $x := x_0 + nh, \xi_n := |v_h(x) - u_h(x)|$

$$|\xi_{n+1}| \leq \underbrace{(1+hL)}_{=:a \neq 1} |\xi_n| + \underbrace{h\delta}_{=:b}$$

$$|\xi_n| \leq (1+hL)^n \delta + \frac{(1+hL)^n - 1}{hL} h \cdot \delta \leq (1+hL)^n \left(\frac{1+L}{L}\right) \cdot \delta$$

$$(1+hL)^n \leq e^{nhL} \leq e^{(b-x_0)L} \quad (\text{independent of } h)$$

Hence $|v_h(x) - u_h(x)| \leq K \cdot \delta$ with $K = \frac{1+L}{L} e^{(b-x_0)L}$. $\square$

Runge–Kutta methods

Smoothness properties of f are transferred to ϕ :

$f \in F_1(I)$, then ϕ continuous w.r.t. x, y, t and Lipschitz continuous w.r.t. y in $D \times [0, h_0]$.

Example: Improved Euler method

$$\phi(x, y, h) = \frac{1}{2}f(x, y) + \frac{1}{2}f(x+h, y+hf(x, y))$$

$$\begin{aligned} \phi(x, y, h) - \phi(x, z, h) &= \frac{1}{2}[f(x, y) - f(x, z)] \\ &\quad + \frac{1}{2}[f(x+h, y+hf(x, y)) - f(x+h, z+hf(x, z))] \end{aligned}$$

$$\begin{aligned} |\phi(x, y, h) - \phi(x, z, h)| &\leq \frac{1}{2}M|y - z| + \frac{1}{2}M|y - z + h[f(x, y) - f(x, z)]| \\ &\leq M|y - z| + \frac{h}{2}M^2|y - z| = M(1 + \frac{1}{2}hM)|y - z| \end{aligned}$$

Lipschitz constant $L := M(1 + \frac{1}{2}h_0M)$

Rule of thumb $hM < 1$ (implicit multistep methods $h|\beta_k|M < 1$), i.e., $L \approx M$

Note that consistency is not necessary for asymptotic stability!

Example: $(0, 0, \frac{1}{2}) : \phi(x, y, h) = \frac{1}{2}f(x, y)$

Lipschitz continuous, hence asymptotically stable, but not consistent!

Not convergent: $y' = 1, y(0) = 0 : u(x) = x, u_h(x) = \frac{1}{2}x$

Definition: A one-step method is called *convergent*, if

$$\lim_{\substack{h \rightarrow 0 \\ \frac{x-x_0}{h} \in \mathbb{N}}} u_h(x) = u(x) \quad \text{for all } x \in I \text{ and for all } f \in F_1(I).$$

Rate of convergence

Order of convergence p if and only if $u_h(x) - u(x) = O(h^p)$ ($h \rightarrow 0$)

Estimate of the **global discretization error**

$$e_h(x) := u_h(x) - u(x), \quad x \in I_h.$$

Theorem 17.3 *Let the increment function ϕ of a one-step method be continuous with respect to x, y, t and Lipschitz continuous with respect to y in $D \times [0, h_0]$. Then the one-step method is convergent if and only if it is consistent.*

Proof Consistency (of order 1) $\implies$ convergence (of order 1)

$$\begin{aligned} e_h(x+h) &= u(x) + h\Delta(x, u(x), h) - [u_h(x) + h\phi(x, u_h(x), h)] \\ &= e_h(x) + h \underbrace{[\Delta(x, u(x), h) - \phi(x, u(x), h)]}_{\substack{\text{local DE} \\ \tau(x, u(x), h) = O(h)}} + h \underbrace{[\phi(x, u(x), h) - \phi(x, u_h(x), h)]}_{\substack{\text{error transport} \\ \text{Lipschitz condition}}} \\ |e_h(x+h)| &\leq \underbrace{(1+hL)}_{=: \hat{a} \neq 1} \underbrace{|e_h(x)|}_{=: \xi_n} + \underbrace{h|\tau(x, u(x), h)|}_{=: \hat{b}}, \quad \underbrace{e_h(x_0)}_{=: \xi_0} = 0 \end{aligned}$$

$$\text{Lemma 17.1} \quad \curvearrowright \quad |e_h(x)| \leq \frac{(1+hL)^n - 1}{L} |\tau(x, u(x), h)|$$

$$\boxed{|e_h(x)| \leq K \cdot |\tau(x, u(x), h)|}$$

Condition number $K = \frac{1}{L}(e^{(b-x_0)L} - 1)$ (rough estimation)

Hence $|e_h(x)| \leq hMK \rightarrow 0$ for $h \rightarrow 0$ uniformly f.a. $x \in I$ ($\frac{x-x_0}{h} \in \mathbb{N}$). $\square$

Remark: Estimation of the global discretization error

locale discretization error $|\tau(x, y, h)| \leq h^p M$ implies $|e_h(x)| \leq h^p \cdot MK$

The order of convergence is at least the order of consistency.

Runge–Kutta methods:

Smoothness assumptions of f are transfered to ϕ !

$\Sigma b_j = 1$ is the natural assumption for quadrature formulas!

Result 17.4 *By its nature a Runge–Kutta method is consistent, convergent and asymptotically stable.*

Linear multistep method $\rho(E)y_m = h\sigma(E)y'_m$

Test equation $y' = 0, y(0) = 0 : \rho(E)y_m = 0$

Homogeneous linear difference equations with constant coefficients:

$$\Sigma \alpha_j y_{m+j} = 0, \quad m = 0, 1, 2, \dots \quad (\text{infinite system})$$

<i>Solution</i>	Series $\{y_m\}_{m \geq 0}$
<i>Initial values</i>	$y_0, y_1, \dots, y_{k-1}$
<i>Recurrence formula</i>	$y_{m+k} = - \sum_{j=0}^{k-1} \alpha_j y_{m+j}$
	$\rightarrow$ unique solution $\{y_m\}_{m \geq 0}$

Characteristic polynomial $\rho(\xi) = \sum_{j=0}^k \alpha_j \xi^j,$

Zeros $\xi_1, \dots, \xi_k$ (compare with linear differential equations)

$\{\xi_\nu^m\}_{m \geq 0}$ is a *solution* of the homogeneous DE $\Sigma \alpha_j \xi_\nu^{m+j} = \xi_\nu^m \rho(\xi_\nu) = 0$

$\xi_1, \dots, \xi_k$ distinct: *fundamental system* $\{\xi_1^m\}_{m \geq 0}, \dots, \{\xi_n^m\}_{m \geq 0}$

ξ zero of multiplicity $\ell (> 1)$:

$\{\xi^m\}_{m \geq 0}, \{m\xi^m\}_{m \geq 0}, \dots, \{m(m-1)\dots(m-\ell+2)\xi^m\}_{m \geq 0}$

ℓ linear independent solutions!

Theorem 17.5 *The homogeneous difference equation $\rho(E)y_m = 0$ with arbitrary initial values $y_0, \dots, y_{k-1}$ has exactly one solution $\{y_m\}_{m \geq 0}$. The components y_m can be represented by a linear combination of the fundamental solutions which are defined by the zeros of the characteristic polynomial ρ .*

Examples

- 1) $y_{m+2} - y_m = 0$, IV $y_0 = \alpha_0, y_1 = \alpha_1 \parallel y_0 = 0, y_1 = \delta$
 zeros of $\rho: \xi_{1/2} = \pm 1$, solution $\{y_m\}_{m \geq 0}$ with $y_m = c_1 1^m + c_2 (-1)^m$
 IV: $\left. \begin{array}{l} y_0 = c_1 + c_2 = \alpha_0 \\ y_1 = c_1 - c_2 = \alpha_1 \end{array} \right\} \curvearrowright c_1 = \frac{\alpha_0 + \alpha_1}{2}, c_2 = \frac{\alpha_0 - \alpha_1}{2} \parallel y_m = \frac{\delta}{2}(1 + (-1)^{m+1})$
- 2) $y_{m+2} - 2y_{m+1} + y_m = 0$, $m = 0, 1, 2, \dots$
 zeros of $\delta: \xi_{1/2} = 1$
 general solution $\{y_m\}_{m \geq 0}$ with $y_m = c_1 \cdot 1^m + c_2 \cdot m \cdot 1^m$
 $y_0 = 0, y_1 = \delta \curvearrowright \begin{array}{l} y_0 = c_1 + c_2 \cdot 0 = 0 \\ y_1 = c_1 + c_2 = \delta \end{array} \curvearrowright c_1 = 0, c_2 = \delta$
 solution $\{y_m\}_{m \geq 0}$ with $y_m = m \cdot \delta \rightarrow \infty$ for $m \rightarrow \infty$ *Instability*
- 3) $y_{m+2} - 4y_{m+1} + 3y_m = 0$, IV $y_0 = 0, y_1 = \delta$
 zeros of $\rho: \xi_1 = 1, \xi_2 = 3$, solution $\{y_m\}_{m \geq 0}$
 with $y_m = c_1 1^m + c_2 3^m = \frac{3^m - 1}{2} \delta \rightarrow \infty$ *Instability*

Growth behaviour of the solution $\{y_m\}_{m \geq 0}$

$$\boxed{\lim_{m \rightarrow \infty} \frac{y_m}{m} = 0} \Leftrightarrow |y_m| \leq K \iff \text{Stability}$$

Zeros of ρ : ξ single zero then $y_m = \xi^m$, $\lim_{m \rightarrow \infty} \frac{y_m}{m} = 0 \Leftrightarrow |\xi| \leq 1$

ξ of multiplicity ℓ then $y_m = m(m-1)\dots(m-\mu+2)\xi^m$,
 $m = 2, \dots, \ell$ ($\ell \geq 2$)

$$\lim_{m \rightarrow \infty} \frac{y_m}{m} = \lim_{m \rightarrow \infty} (m-1)\dots(m-\mu+2)\xi^m = 0 \Leftrightarrow |\xi| < 1$$

Definition The polynomial ρ with the zeros $\xi_1, \dots, \xi_k$ satisfies the **root condition**, if:

$$|\xi_j| \leq 1, \quad j = 1, \dots, k,$$

$|\xi_j| = 1$ implies ξ_j single zero.

Asymptotic stability of linear multistep methods

Analogous theory as for one-step methods:

approximation	$u_h : \rho(E)u_h(x) = h\sigma(E)u'_h(x)$
perturbed approximation	$v_h : \rho(E)v_h(x) = h[\sigma(E)v'_h(x) + r_h(x)]$
perturbation	$\psi_h(v_h - u_h) = \max\{ v_h(x_j) - u_h(x_j) , j = 0, 1, \dots, k-1, r_h(x) , x \in I_h\}$

Definition A linear multistep method is called *asymptotically stable*, if there exist positive numbers δ_0, h_0, K such that for each $\delta \in [0, \delta_0]$ the perturbed approximation v_h with

$$\psi_h(v_h - u_h) \leq \delta$$

satisfies uniformly for all $h \in [0, h_0]$ the condition

$$|v_h(x) - u_h(x)| \leq K \cdot \delta \quad \text{for all } x \in I_h.$$

Theorem 17.6 *A linear multistep method is asymptotically stable if and only if the characteristic polynomial ρ satisfies the root condition.*

Idea of the proof: root condition $\implies$ asymptotic stability

error $\varepsilon_m := v_h(x_m) - u_h(x_m)$, u_h approximation, v_h perturbed approximation

homogeneous difference equation $\rho(E)\varepsilon_m = 0$ ($\alpha_k = 1$)

solution $\{\bar{\varepsilon}_m\}_{m \geq 0}$, $|\bar{\varepsilon}_m| \leq K \cdot \delta$ independent of the initial values

Inhomogeneous difference equation $\rho(E)\varepsilon_m = c_m$

where $c_m = h\sigma(E) \underbrace{[f(x_m, v_h(x_m)) - f(x_m, u_h(x_m))]}_{|\cdot| \leq L|v_h(x_m) - u_h(x_m)|} + h \underbrace{r_h(x_m)}_{|\cdot| \leq \delta}$

and $|c_m| \leq h \cdot \hat{K}$

Solution $\{\hat{\varepsilon}_m\}$ = solution of the homogeneous DE + particular solution

Particular solution $\begin{aligned} \tilde{\varepsilon}_{m+k-1} &= \{\Sigma \text{ of } m \text{ terms } c_\nu \varepsilon_\mu\} \\ |\tilde{\varepsilon}_{m+k-1}| &\leq m \cdot \max |c_\nu| \max |\varepsilon_\mu| \\ &\leq m \cdot h\hat{K} \cdot K \cdot \delta \leq (b - x_0)\hat{K}K \cdot \delta. \end{aligned}$

Solution $|\hat{\varepsilon}_m| \leq |\varepsilon_m| + |\tilde{\varepsilon}_m| \leq M \cdot \delta \Rightarrow$ asymptotic stability $\square$

Remark: Mostly the root condition is used as definition of the asymptotic stability (to avoid the previous proof).

Definition (alternative) A linear multistep method is called *asymptotically stable*, if its characteristic polynomial $\rho(\xi)$ satisfies the root condition.

Zeros of ρ : $\xi_1, \dots, \xi_k$

Consistency $\rho(1) = 0, \rho'(1) = \sigma(1) (\neq 0 \text{ because of irreducibility})$

Main zero $\xi_1 = +1$

Parasitic zeros $\xi_2, \dots, \xi_k$

Examples: IVP $y' = qy, y(0) = 1 : u(x) = e^{qx}, u(x_m) = (e^{qh})^m$

1) $y_{m+2} - y_{m+1} = \frac{h}{2}(3f_{m+1} - f_m)$ (Adams–Bashforth, $p = 2$)

zeros of ρ : $\xi_1 = 1, \xi_2 = 0$

difference equation $y_{m+2} - (1 + \frac{3}{2}qh)y_{m+1} + \frac{1}{2}qhy_m = 0$

characteristic polynomial $\rho(\xi) - qh\sigma(\xi) = \xi^2 - (1 - \frac{3}{2}qh)\xi + \frac{1}{2}qh$

zeros $\xi_{1/2}(qh) = \frac{1}{2}(1 + \frac{3}{2}qh \pm \sqrt{1 + qh + \frac{9}{4}(qh)^2})$

$\xi_1(qh) = 1 + qh + \frac{1}{2}(qh)^2 + O(h^3), \xi_1(0) = \xi_1$ main zero

$\xi_2(qh) = \frac{1}{2}qh + O(h^2)$

solution $\{y_m\}_{m \geq 0}$ with $y_m = \alpha_1 \xi_1^m(qh) + \alpha_2 \xi_2^m(qh)$

Main solution $\xi_1(qh)$ approximates the wanted solution $u(qh)$

Parasitic solution $\xi_2(qh)$ has nothing to do with the solution (caused by the method) $\rightarrow \xi_j(0) = \xi_2$

If $|\xi_j(0)| > 1, j = 2, \dots, k$, then parasitic solutions cause instability!

2) $y_{m+2} - y_m = 2hf_{m+1}$ ($p = 2$), zeros of $\rho : \pm 1$
characteristic polynomial $\xi^2 - 2qh\xi - 1 = 0$

zeros $\xi_1(qh) = 1 + qh + \frac{1}{2}(qh)^2 + O(h^3)$

$\xi_2(qh) = -1 + qh - \frac{1}{2}(qh)^2 + O(h^3)$

solution $\{y_m\}$ with $y_m = \alpha_1 \xi_1^m(qh) + \alpha_2 \xi_2^m(qh)$

$q < 0$: there is no stepsize $h > 0$ with $|\xi_2(qh)| \leq 1$

Several zeros of ρ on the unit circle are disturbing!

(The parasitic solution has equal weight)!

Definition An asymptotically stable linear multistep method is called *strong asymptotically stable*, if the parasitic zeros of ρ satisfy $|\xi_j| < 1$ ($j = 2, \dots, k$).

Example: *Methods of Adams* $\rho(\xi) = \xi^{k-1}(\xi - 1)$
main zero $\xi_1 = 1$, parasitic zeros $\xi_2 = \dots = \xi_k = 0$

Remark: One-step methods have no parasitic solutions ($\rightarrow$ more stable).

Definition A linear multistep method is called **convergent**, if

$$\lim_{h \rightarrow 0} u_h(x) = u(x)$$

$$\frac{x - x_0}{h} \in \mathbb{N}$$

f.a. $x \in I$, f.a. consistent starting values and f.a. $f \in F_1(I)$.

Rate of convergence

Order of convergence p if and only if $u_h(x + kh) - u(x + kh) = O(h^p)$ ($h \rightarrow 0$)

Estimation of the **global discretization error** (for Adams–Bashforth methods)

$$e_h(x) := u_h(x) - u(x), \quad x \in I_h$$

$$|e_h(x + kh)| \leq |e_h(x + (k-1)h)| + hL \sum_{j=0}^{k-1} |\beta_j| |e_h(x + jh)| + h|\tau(x, u(x), h)|$$

$$\boxed{|e_h(x)| \leq K_1 \cdot |\tau(x, u(x), h)| + K_2 \cdot \max\{|e_h(x_0 + jh)|, j = 0, 1, \dots, k-1\}}$$

Theorem 17.7 A linear multistep method is convergent if and only if it is consistent and asymptotically stable.

Proof: See Stoer–Bulirsch

Result 17.8 *A linear multistep method satisfying*

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1)$$

and

$$\text{root condition for } \rho$$

is consistent, convergent und asymptotically stable.

Compare with Runge–Kutta methods (Result 17.4):

$$\Sigma \text{ weights} = 1$$

The maximal order of consistency

$$k\text{-step methods} \quad p^*(k) = 2k$$

$$\text{asymptotically stable} \quad p^*(k) = k + 2$$

$$\text{strong asymptotically stable} \quad p^*(k) = k + 1$$

The root condition for ρ is a strong restriction!

Examples

Kepler method: $y_{m+2} - y_m = \frac{h}{3}(f_{m+2} + 4f_{m+1} + f_m)$, zeros $\xi_1 = 1, \xi_2 = -1$
 $k = 2, p = 4 = 2k = k + 2$

Adams–Moulton: strong asymptotically stable, order $p(k) = k + 1$ (maximal),
 root condition ideally satisfied $\xi_1 = 1, \xi_2 = \dots = \xi_k = 0$.

18 Absolute Stability

Asymptotic stability $h \rightarrow 0$, finite interval $[x_0, b]$, constant in $K \cdot \delta$

Test equation $y' = qy$, $y(0) = 1$, $\operatorname{Re} q < 0$ (real part of q)

Solution $u(x) = e^{qx}$, $x \geq 0$, “one step” $u(x+h) = e^{qh}u(x)$

Stability

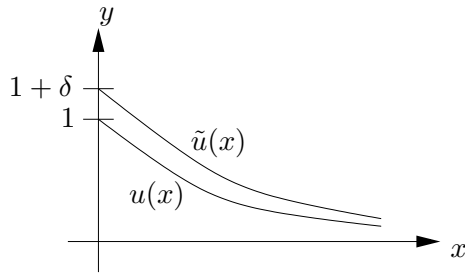
Perturbed initial value $\tilde{y}(0) = 1 + \delta$

perturbed solution $\tilde{u}(x) = (1 + \delta)e^{qx}$

$|\tilde{u}(x) - u(x)| \leq 1 \cdot \delta$ f.a. $x \geq 0$

or $|\tilde{u}(x+h) - u(x+h)| \leq |\tilde{u}(x) - u(x)|$

f.a. $x \geq 0 \rightarrow$ **contractivity**



Or: Compare $u(x)$ with the zero-solution, i.e., $|u(x+h)| \leq |u(x)|$ f.a. $x \geq 0$

Discretization methods

Approximation $u_h(x)$ and *perturbed approximation* $\tilde{u}_h(x)$

Stability $|\tilde{u}_h(x) - u_h(x)| \leq K \cdot \delta$ f.a. $x \in I_h$ (infinite)

Contractivity $|\tilde{u}_h(x+h) - u_h(x+h)| \leq |\tilde{u}_h(x) - u_h(x)|$ f.a. $x \in I_h$

Question: For which stepsizes h is **stability** resp. **contractivity** satisfied?

Only for special test equations an answer is possible!

Test equation $y' = qy$

$y' = f(x, y) \xrightarrow{\text{Linearization}} y' = Jy \xrightarrow{\text{Diagonalization}} y' = qy$

In the other direction we hope that a method, effective for the test equation, will also be effective for the nonlinear system.

Euler polygon method

$$u_h(x+h) = u_h(x) + qh u_h(x) = (1 + qh)u_h(x) = (1 + qh)^{m+1} \quad (x = x_m)$$

Comparison with the exact solution:

e^{qh} is approximated by the Taylor polynomial $1 + qh$ (order 1).

Stability

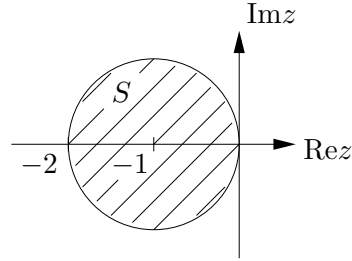
Perturbation $|\tilde{u}_h(0) - u_h(0)| \leq \delta$ implies the effect

$$\begin{aligned} |\tilde{u}_h(x) - u_h(x)| &\leq \underbrace{|1 + qh|^m}_{\leq K} \cdot \delta \quad \text{f.a. } x \in I_h \quad (\text{i.e., } m \rightarrow \infty) \\ &\leq K \Leftrightarrow |1 + qh| \leq 1 \quad \text{and } K = 1 \end{aligned}$$

Stability

if qh satisfies $|1 + qh| \leq 1$, i.e., $qh \in S$.
The parameter q is given!

We prefer $<$ and consider
= as the limit case!



Stability region $A := \{z \in \mathbb{C} \mid |1 + z| < 1\}$

Stability means contractivity

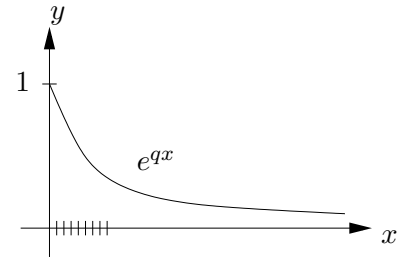
$$|\tilde{u}_h(x+h) - u_h(x+h)| \leq \underbrace{|1 + qh|}_{\leq 1} |\tilde{u}_h(x) - u_h(x)|$$

Error:

local DE $h \cdot \tau(x, y, h) = \frac{1}{2}h^2 u''(x) + O(h^3) \quad (h \rightarrow 0)$

With a reasonable accuracy h is small enough
and $|1 + qh| \leq 1$ is also satisfied.

Result: Treating the test equation accuracy
and stability go hand in hand.



Example: linear system $y' = Fy$

$$F = \begin{pmatrix} -\frac{101}{2} & \frac{99}{2} \\ \frac{99}{2} & -\frac{101}{2} \end{pmatrix}, \quad y(0) = \begin{pmatrix} 3 \\ 1 \end{pmatrix}, \quad \text{eigenvalues } q = -1, -100$$

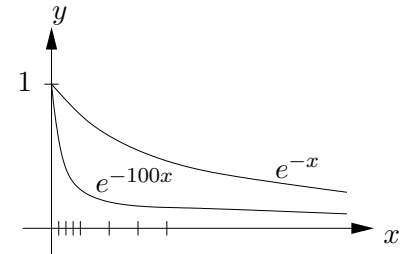
$$\text{solution } u(x) = 2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} e^{-x} + \begin{pmatrix} 1 \\ -1 \end{pmatrix} e^{-100x}$$

Stiff system: strongly distinct time constants!

Euler polygon method $u_h(x+h) = (I + hF)u_h(x)$

Stability/contractivity: spectral radius $\rho(I + hF) < 1$

$$\text{i.e., } |1 + q_i h| < 1, \quad -100h \in S_{\text{Euler}}, \quad h < 0.02$$



Numerical solution $u_h(x) = 2 \begin{pmatrix} 1 \\ 1 \end{pmatrix} (1-h)^m + \begin{pmatrix} 1 \\ -1 \end{pmatrix} (1-100h)^m \quad (x = x_m)$

Initial phase: h small (< 0.02) because of accuracy with respect to e^{-100x}

Afterwards: Choose h for accuracy with respect to e^{-x}

$$\text{Error } \varepsilon_1(h) := e^{-h} - (1-h), \quad \varepsilon_2(h) := e^{-100h} - (1-100h)$$

h	0.5	0.2	0.1	0.05	0.02	0.01	0.005	0.002	0.001
$\varepsilon_1(h)$	0.1	0.02	0.005	0.001	0.0002	$5 \cdot 10^{-5}$	$2 \cdot 10^{-5}$	$2 \cdot 10^{-6}$	$5 \cdot 10^{-7}$
$\varepsilon_2(h)$			9.0		1.2	0.4	0.1	0.02	0.005

If $h > 0.02$ then the already disappeared exponential terms regrow!

Wanted: Methods having a large stability region S

best possible with $S \supseteq \mathbb{C}_-$ (left half plane) $\leadsto$ **absolute stability**

Then choose h only with respect to **accuracy**!

Definition: A linear system $y' = Fy$ is called *stiff*, if it satisfies that all eigenvalues $q_j(F) \in \mathbb{C}_-$ and

$$\frac{\max |\operatorname{Re} q_i(F)|}{\min |\operatorname{Re} q_i(F)|} \geq 50.$$

A nonlinear system $y' = f(x, y)$ is called *stiff*, if the linearized system $y' = J(x)y$ is stiff for each admissible x .

One-step methods: Runge-Kutta (RK) methods (c, A, b)

$$U'_j = f(y + h \sum a_{jk} U'_k) \quad (y = u_h(x)) \quad j = 1, \dots, s,$$

$$u_h(x+h) = y + h \sum b_j U'_j$$

Application to the test equation

$$u_h(x+h) = W(z)u_h(x), \quad z := qh$$

Because

$$\begin{pmatrix} U'_1 \\ \vdots \\ U'_s \end{pmatrix} = eqy + qhA \begin{pmatrix} U'_1 \\ \vdots \\ U'_s \end{pmatrix} \leadsto (I - qhA) \begin{pmatrix} U'_1 \\ \vdots \\ U'_s \end{pmatrix} = eqy, \quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

$$u_h(x+h) = y + hb^T \begin{pmatrix} U'_1 \\ \vdots \\ U'_s \end{pmatrix} = (1 + qhb^T(I - qhA)^{-1}e) y$$

Stability function $W(z) := 1 + zb^T(I - zA)^{-1}e = \frac{\det(I+z\tilde{A})}{\det(I-zA)}$ appr. $\exp(z)$

Explicit RK method: $W(z)$ = polynomial of degree $\leq s$

$$s = p: \quad W(z) = 1 + z + \dots + \frac{1}{p!}z^p \quad \text{Taylor polynomial}$$

Implicit RK method: $W(z) = \frac{N(z)}{D(z)}$, degree $N \leq s$, degree $D \leq s$

Implicit Euler method $W(z) = \frac{1}{1-z} \quad (1, 1, 1)$

Trapezoidal method $W(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} \quad \left(0 \left| \begin{array}{cc} 0 & 0 \end{array} \right| \frac{1}{2} \right. \\ \left. 1 \left| \begin{array}{cc} \frac{1}{2} & \frac{1}{2} \end{array} \right| \frac{1}{2} \right)$

Gaussian method ($s = 1$) $W(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} \quad (\frac{1}{2}, \frac{1}{2}, 1)$

Gaussian RK methods $W(z) = \frac{N(z)}{D(z)}$, degree $N =$ degree $D = s$, order $p = 2s$

Padé approximation

$$\exp(z) - W(z) = O(z^{2s+1}) \quad (z \rightarrow 0) \quad (\text{order } p = 2s)$$

$$\text{Stability region } S = \{z \in \mathbb{C} \mid |W(z)| < 1\}$$

Stability: Choose h such that $qh \in S$.

Absolute stability: Stability independent of $h > 0$, i.e., $\mathbb{C}_- \subseteq S$.

Definition A RK method (c, A, b) is called *A-stable* if it satisfies $\mathbb{C}_- \subseteq S$.

Theorem 18.1 *An explicit RK method cannot be A-stable.*

Proof $W(z)$ = polynomial and hence unbounded in the left half plane. □

Figure: Stability regions

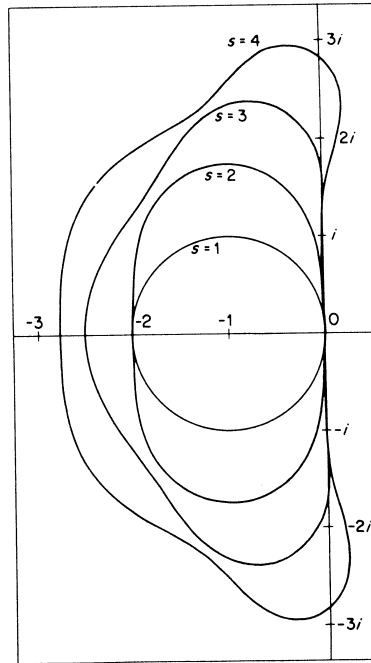
Explicit RK methods with
the largest stability region:

$$s = p = 1, \dots, 4$$

$$W(z) = 1 + z + \dots + \frac{1}{p!} z^p$$

Stability interval

$$(r, 0) \subseteq S \text{ (maximal)}$$



Theorem 18.2 *The implicit Euler and the trapezoidal method are A-stable.*

Proof: $W(z) = \frac{1}{1-z}$ resp. $w(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$

The maximum principle for holomorphic functions

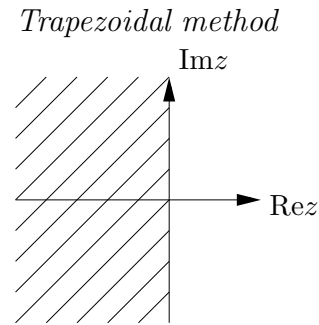
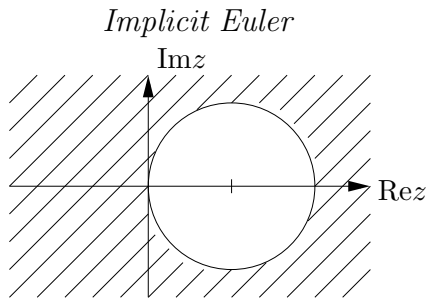
1. There is no pole in the left half plane (no zeros of the denominator polynomial in $\mathbb{C}_-$)
2. $|W(iy)| \leq 1$ f.a. $y \in \mathbb{R}$ (the boundary of the left half plane)

There is no pole in $\mathbb{C}_-$: That's clear!

The boundary

$$|W(iy)|^2 = \frac{1}{|1-iy|^2} = \frac{1}{1+y^2} \leq 1 \quad \text{f.a. } y \in \mathbb{R}$$

$$|W(iy)|^2 = \frac{|1+\frac{1}{2}iy|^2}{|1-\frac{1}{2}iy|^2} = \frac{1+\frac{1}{4}y^2}{1+\frac{1}{4}y^2} = 1 \quad \text{f.a. } y \in \mathbb{R} \quad \square$$



Theorem 18.3 *The Gaussian RK methods with $p = 2s$ are A-stable.*

Idea of the proof:

Stability function $W(z) = \frac{N(z)}{D(z)}$, degree $N = \text{degree } D = s$, Padé approximation

$$N(z) = 1 + \alpha_1 z + \dots + \alpha_s z^s \quad \text{with } \alpha_j > 0, j = 1, \dots, s$$

$$D(z) = 1 - \alpha_1 z + \dots + (-1)^s z^s$$

$D(z)$ has no zeros in the left half plane: Routh–Hurwitz criterion

$$|W(iy)| = 1 \quad \text{for all } y \in \mathbb{R} \quad \square$$

Gaussian RK methods: Stability region $S = \mathbb{C}_-$

Linear multistep methods $\rho(E)y_m = h\sigma(E)f_m \rightarrow$ test equation

Difference equation $\rho(E)y_m = qh\sigma(E)y_m \quad (z := qh)$

Characteristic polynomial $\Pi(\xi, z) := \rho(\xi) - z\sigma(\xi) \rightarrow$ stability polynomial

Zeros $\xi_j(z), \quad j = 1, \dots, k$

Growth behaviour $\frac{y_m}{m} \rightarrow 0$ für $m \rightarrow \infty \rightarrow$ root condition for $\Pi(\xi, z)$

Main zero $\xi_1(z) = \sum_{\nu=1}^p \frac{1}{\nu!} z^\nu + O(z^{p+1}) = e^z + O(z^{p+1}) \quad (|z| \rightarrow 0)$

Parasitic zeros $\xi_2(z), \dots, \xi_k(z)$ are analytic functions of z

Stability region $S := \{z \in \mathbb{C} \mid |\xi_j(z)| < 1, j = 1, \dots, k\}$

Stability interval $I := (r, 0)$ with $r = \max_{\xi < 0} \{-\xi \mid (\xi, 0) \in S\}$

Stability: Choose h such that $z = qh \in S$

Wanted: Methods with a large stability region!

Definition A consistent linear multistep method satisfying the root condition of ρ is called *A-stable* if $\mathbb{C}_- \subseteq S$.

Examples

Euler poygon method ($k = 1$) : $\xi_1(z) = 1 + z$

Implicit Euler method ($k = 1$) : $\xi_1(z) = \frac{1}{1-z}$

Trapezoidal method ($k = 1$) : $\xi_1(z) = \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}$

Midpoint method ($k = p = 2$) : $y_{m+2} - y_m = 2h f_{m+1}$

zeros of ρ : $\xi_1 = 1$, $\xi_2 = -1$ not asymptotically stable

Stability polynomial $\Pi(\xi, z) = \rho(\xi) - z\sigma(\xi) = \xi^2 - 2z\xi - 1$

Zeros $\xi_{1/2}(z) = z \pm \sqrt{z^2 + 1}$

$\xi_1(z) = 1 + z\frac{1}{2}z^2 + O(z^3)$ ($|z| \rightarrow 0$) approximates $\exp(z)$

$\xi_2(z) = -1 + z - \frac{1}{2}z^2 + O(z^3)$ has nothing to do with $\exp(z)$

$$\left. \begin{array}{l} |\xi_1(z)| < 1 \text{ if } z < 0 \\ |\xi_2(z)| < 1 \text{ if } z > 0 \end{array} \right\} \text{ implies } S = \emptyset$$

Adams methods ($k = 2$) : $y_{m+2} - y_{m+1} = \frac{h}{2} (3f_{m+1} - f_m)$

$$\Pi(\xi, z) = \xi^2 - (1 + \frac{3}{2}z)\xi + \frac{1}{2}z$$

$$\xi_{1/2}(z) = \frac{1}{2} \left(1 + \frac{3}{2}z \pm \sqrt{1 + z + \frac{9}{4}z^2} \right)$$

The main solution $\xi_1(z) = 1 + z + \frac{1}{2}z^2 + O(z^3)$ ($|z| \rightarrow 0$) approximates $\exp(z)$

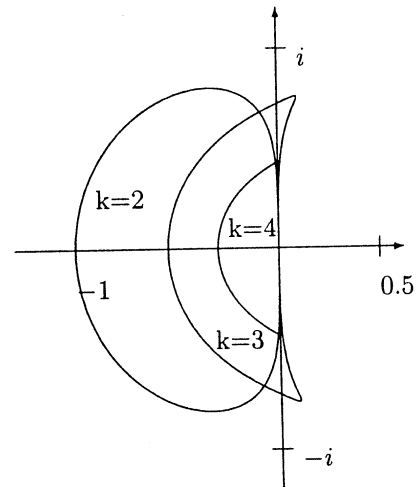
the parasitic solution $\xi_2(z) = \frac{1}{2}z - \frac{1}{2}z^2 + O(z^3)$ has nothing to do with $\exp(z)$

Stability Choose $z = qh$ s.t. $|\xi_i(z)| < 1$

$$|\xi_1(z)| < 1 \text{ for all } z < 0$$

$$|\xi_2(z)| < 1 \text{ for all } -1 < z < 0$$

Figure: Stability region of Adams–Bashforth methods ($k=2,3,4$)



Remark: Parasitic solutions reduce the stability region substantially!

Stability interval $(-r, 0)$ of the Adams methods

Adams–Bashforth methods ($p = k$) (explicit)

k	1	2	3	4
$-r$	-2	-1	$-\frac{6}{11}$	$-\frac{3}{10}$

Adams–Moulton methods ($p = k + 1$) (implicit)

k	1	2	3	4
$-r$	$-\infty$	-6	-3	$-\frac{30}{49}$

Remarks: The stability interval becomes smaller and smaller as k increases, because the number of parasitic solutions increases!

The trapezoidal method is the best method $\rightarrow$ Runge–Kutta method

Theorem 18.4 (Dahlquist barrier) *An A -stable linear multistep method has at most the order of consistency $p = 2$.*

Definition A linear multistep method is called $A(\alpha)$ -stable if

$$\{z | (\arg(-z)) < \alpha, z \neq 0\} \subset S.$$

Backward Difference Formulas (BDF)

$$\sum_{\nu=1}^k \frac{1}{\nu} \nabla^{\nu} y_{m+k} = h f_{m+k}, \quad p = k$$

The root condition for ρ is satisfied if $k \leq 6$.

$k = 1$: Implicit Euler method

$$k = 2 : \frac{3}{2}y_{m+2} - 2y_{m+1} + \frac{1}{2}y_m = h f_{m+2}, \quad \xi_{1/2}(z) = \frac{2 \pm \sqrt{1+2z}}{3-2z}$$

$$k = 3 : \frac{11}{6}y_{m+3} - 3y_{m+2} + \frac{3}{2}y_{m+1} - \frac{1}{3}y_m = h f_{m+3}$$

$k = 1, 2$: A -stable, $k = 3, \dots, 6$: “almost” A -stable

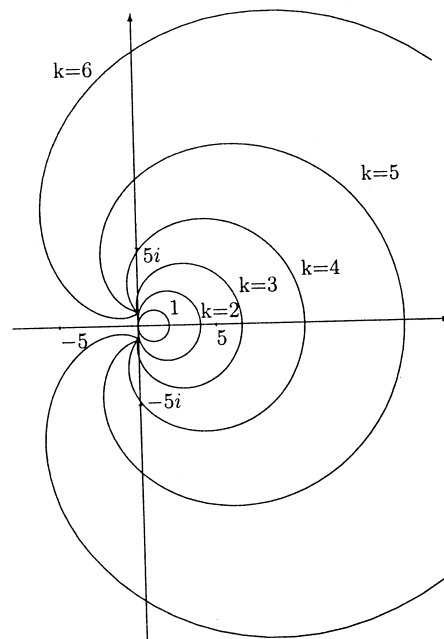
BDF $A(\alpha)$ -stable

k	1	2	3	4	5	6
α	90°	90°	86°	73°	51°	17°

For treating stiff systems A -stable or $A(\alpha)$ -stable methods are necessary!

Figure : Stability regions
of the BDF ($k=1, \dots, 6$)

Stability regions
“left” or “outside”
of the contours
symmetric to
the real axis



Program packages

BDF

→ Gear package

Radau IIA formulas $p = 2s - 1$

→ Hairer–Wanner (volume 2)

(especially Radau 5)