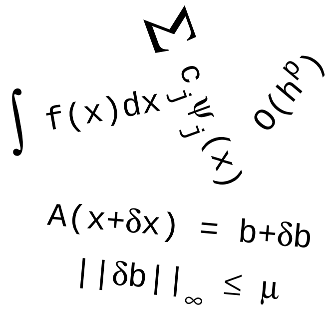

Formulas from

Lars Eldén, Linde Wittmeyer-Koch, Hans Bruun Nielsen

Introduction to
Numerical Computation

– analysis and MATLAB® illustrations

January 9, 2004

Contents

Notation

R_{T} truncation error

R_{XF} error in the result, coming from errors in the function values used

$\ll$ "much smaller than"

$\simeq$ "approximately equal to"

$\lesssim$ "less than or approximately equal to"

2. Error Analysis and Computer Arithmetic

Let a denote an *exact value*, and $\overline{a}$ an *approximation* of a

$$\textit{Absolute error}: \quad \Delta a = \overline{a} - a \ .$$

$$\underset{(a \neq 0)}{\textit{Relative error}}: \quad \frac{\Delta a}{a} \qquad (\simeq \Delta a/\overline{a} \ \text{ if } \ |\Delta a| \ll \overline{a}) \ .$$

Maximal error bound

Let $\Delta f = f(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n) - f(x_1, x_2, \ldots, x_n)$.

$$|\Delta f| \ \lesssim \ \sum_{k=1}^{n} \left| \frac{\partial f}{\partial x_k}(\overline{x}) \ \Delta x_k \right| \ .$$

Floating point representation

Normalized floating point number with $t{+}1$ digits and base β :

$$\begin{aligned} & x = \pm d_0.d_1 d_2 d_3 \ldots d_t \cdot \beta^e \ , \\ & 1 \leq d_0 \leq \beta - 1 \ , \\ & 0 \leq d_i \leq \beta - 1, \quad i = 1, 2, \ldots, t \ , \end{aligned}$$

and e is an integer.

Let x be the representation of the real number X, obtained by rounding. Then

$$\frac{|x - X|}{|X|} \leq \mu, \qquad \mu = \tfrac{1}{2}\beta^{-t} \ .$$

μ is called the *unit roundoff.*

Let $\odot$ denote any of the arithmetic operators $+, -, *$ and $/$, and let $fl[x \odot y]$ denote the computed result of $x \odot y$. If $x \odot y \neq 0$, then

$$\left| \frac{fl[x \odot y] - x \odot y}{x \odot y} \right| \leq \mu \ ,$$

or, equivalently,

$$fl[x \odot y] = (x \odot y)(1 + \epsilon) \ ,$$

for some ϵ that satisfies $|\epsilon| \leq \mu$.

3. Function Evaluation

Remainder Term Estimates

Notation: $S = \sum_{n=1}^{\infty} a_n, \qquad S_N = \sum_{n=1}^{N} a_n, \qquad R_N = S - S_N = \sum_{n=N+1}^{\infty} a_n \ .$

Alternating series.

$$|R_N| \leq |a_{N+1}| \ .$$

Estimation by an integral. Assume that $a_n = f(n)$ and that $f(x)$ is positive and monotonically decreasing for $x > N$. Then

$$R_N = \sum_{n=N+1}^{\infty} f(n) \leq \int_N^{\infty} f(x)\, dx \ .$$

Comparison with a known series. Assume that

$$0 \leq a_n \leq b_n \ , \quad n \geq N+1 \ ,$$

and that $T_N = \sum_{n=N+1}^{\infty} b_n$ is known. Then

$$R_N \leq T_N \ .$$

4. Nonlinear Equations

Iteration methods for the solution of $f(x) = 0$ with a simple root x^*.

Fixed point method. Reformulate $f(x) = 0$ to $x = \varphi(x)$ and iterate:

$$x_{k+1} = \varphi(x_k) \ .$$

Converges if $|\varphi'(x)| \leq m < 1$ for x close to the root x^*.

Newton-Raphson's method.

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \ .$$

Converges if x_0 is chosen sufficiently close to x^*.

The secant method.

$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \ .$$

Order of convergence. A convergent sequence $x_0, x_1, x_2, \ldots$ has the order of convergence p if $p \geq 1$ is the largest positive number such that

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_{k+1} - x^*|^p} = C < \infty \ .$$

C is called the *asymptotic error constant.*

For $p=1$ and $p=2$ the convergence is said to be *linear* and *quadratic*, respectively.

Method-independent error estimate. Let $\overline{x}$ be an approximation to a simple root x^* and $\widetilde{f}(\overline{x})$ be an approximation to $f(\overline{x})$. Then

$$|\overline{x} - x^*| \leq \frac{|\widetilde{f}(\overline{x})| + \delta}{M} \ ,$$

where $|\widetilde{f}(\overline{x}) - f(\overline{x})| \leq \delta$ and $|f'(x)| \geq M$ for all x in a neighbourhood of x^* that includes $\overline{x}$.

Systems of nonlinear equations. *Newton-Raphson's method:*

$$x^{[k+1]} = x^{[k]} - \big(J(x^{[k]})\big)^{-1} f(x^{[k]}), \qquad (J(x))_{ij} = \frac{\partial f_i}{\partial x_j}(x) \ .$$

J is the so-called *Jacobian* of f.

5. Interpolation

Problem: Given function values $f_i = f(x_i)$ at $n{+}1$ distinct points $x_0, x_1, \ldots, x_n$. Seek a polynomial $P(x)$ of degree $\leq n$ such that $P(x_i) = f_i$, $i = i = 0, 1, \ldots, n$.

Newton's interpolation formula

$$P_n(x) = f_0 + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ + \cdots + f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\cdots(x - x_{n-1}) \ ,$$

where $f[x_0, x_1, \ldots, x_k]$ is the kth *divided difference* of f with respect to the points $x_0, x_1, \ldots, x_k$, given by

$$f[x_i] = f(x_i) \ ,$$
$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, x_2 \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0} \ .$$

Lagrange's Interpolating Polynomial

$$P(x) = f_0 L_0(x) + f_1 L_1(x) + \cdots + f_n L_n(x) \ ,$$

$$L_i(x) = \frac{(x - x_0)\cdots(x - x_{i-1})(x - x_{i+1})\cdots(x - x_n)}{(x_i - x_0)\cdots(x_i - x_{i-1})(x_i - x_{i+1})\cdots(x_i - x_n)} \ .$$

Truncation error

$$R_{\mathrm{T}}(x) = f(x) - P(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\,(x - x_0)(x - x_1)\cdots(x - x_n) \ .$$

Truncation error with Newton's interpolation formula,

$$|R_{\mathrm{T}}| \lesssim |\text{ first neglected term }| \ .$$

Linear interpolation

$$P(x) = f_0 + \frac{x - x_0}{x_1 - x_0}(f_1 - f_0) \ .$$

If $\{\overline{f}_i\}$ are given approximations of $\{f(x_i)\}$ and $\max_{i=0,1} |\overline{f}_i - f_i| = \epsilon$, then

$$|R_{\mathrm{XF}}(x)| = |f(x) - P(x)| \leq \epsilon \quad \text{for} \quad x_0 \leq x \leq x_1 \ .$$

Cubic spline interpolation

A cubic spline s with *knots* $x_0 < x_1 < \cdots < x_n$ satisfies

1. s is a polynomial of degree ≤ 3 in each knot interval $[x_{i-1}, x_i]$, $i=1,\ldots,n$,
2. s, s' and s'' are continuous in $[x_0, x_n]$.

For $x_{i-1} \leq x \leq x_i$ we let $s(x) = s_i(x)$, expressed by

$$s_i(x) = a_i + b_i\left(\frac{x-x_{i-1}}{h_i}\right) + c_i\left(\frac{x-x_{i-1}}{h_i}\right)^2 + d_i\left(\frac{x-x_{i-1}}{h_i}\right)^3 ,$$

where

$$h_i = x_i - x_{i-1} .$$

A cubic spline that interpolates $(x_0, f_0),\ (x_1, f_1), \ldots,\ (x_n, f_n)$ is determined by

$$\left.\begin{aligned} a_i &= f_{i-1} , \\ b_i &= h_i s'_{i-1} , \\ c_i &= 3(f_i - f_{i-1}) - h_i(2s'_{i-1} + s'_i) , \\ d_i &= 2(f_{i-1} - f_i) + h_i(s'_{i-1} + s'_i) , \end{aligned}\right\} \quad i = 1, 2, \ldots, n ,$$

where the s'_i satisfy the linear system of equations

$$h_{i+1}s'_{i-1} + 2(h_i+h_{i+1})s'_i + h_i s'_{i+1} = 3\big(h_{i+1}\frac{f_i - f_{i-1}}{h_i} + h_i\frac{f_{i+1} - f_i}{h_{i+1}}\big), \qquad i = 1, 2, \ldots, n-1 ,$$

supplied with two extra conditions. Either

"Natural spline": $\quad 2s'_0 + s'_1 = 3\,\dfrac{f_1 - f_0}{h_1}, \quad s'_{n-1} + 2s'_n = 3\,\dfrac{f_n - f_{n-1}}{h_n}$,

or

"Correct boundary conditions": $\quad s'_0 = f'(x_0)$, $\quad s'_n = f'(x_n)$.

Local truncation error

$$\max_{x_{i-1}\leq x\leq x_i} |f(x) - s_i(x)| \leq \tfrac{1}{384}M_i h_i^4 + \tfrac{1}{4}E'_i h_i ,$$

where

$$M_i = \max_{x_{i-1}\leq x\leq x_i} |f^{(4)}(x)| , \quad E'_i = \max_{j=i-1,i} |f'(x_j) - s'(x_j)| .$$

Global truncation error. If the spline s satisfies the correct boundary conditions, then

$$\max_{x_0\leq x\leq x_n} |s(x) - f(x)| < \tfrac{5}{384}\,h^4 M, \quad h = \max_i h_i, \quad M = \max_i M_i .$$

6. Differentiation and Richardson Extrapolation

Forward difference approximation of the first derivative:

$$f'(x) = \frac{f(x+h) - f(x)}{h} + R_{\mathrm{T}}, \qquad R_{\mathrm{T}} = a_1 h + a_2 h^2 + a_3 h^3 + \cdots \ .$$

If $\overline{f}(x)$ and $\overline{f}(x+h)$ are approximations to $f(x)$ and $f(x+h)$ with $\max\{|\overline{f}(x) - f(x)|, |\overline{f}(x+h) - f(x+h)|\} \le \epsilon$, then

$$|R_{\mathrm{XF}}| \le \frac{2\epsilon}{h} \ .$$

Central difference approximation of the first derivative:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + R_{\mathrm{T}}, \qquad R_{\mathrm{T}} = b_1 h^2 + b_2 h^4 + b_3 h^6 + \cdots \ .$$

$$|R_{\mathrm{XF}}| \le \frac{\epsilon}{h} \ .$$

Second derivative:

$$f''(x) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2} + R_{\mathrm{T}}, \qquad R_{\mathrm{T}} = c_1 h^2 + c_2 h^4 + c_3 h^6 + \cdots \ .$$

$$|R_{\mathrm{XF}}| \le \frac{4\epsilon}{h^2} \ .$$

Richardson extrapolation

Assume that

$$F_1(h) = F(0) + a_1 h^{p_1} + a_2 h^{p_2} + \cdots \ ,$$

with known exponents $p_1, p_2, \ldots$, but unknown $a_1, a_2, \ldots$. We want to compute $F(0)$. Further, assume that F_1 has been computed for arguments $\ldots, q^3h, q^2h, qh, h$, where $q > 1$.

The first term in the expansion of the truncation error can be eliminated by putting

$$F_2(h) = F_1(h) + \frac{1}{q^{p_1} - 1}\left(F_1(h) - F_1(qh)\right) \ .$$

Then

$$F_2(h) = F(0) + \widetilde{a}_2 h^{p_2} + \widetilde{a}_3 h^{p_3} + \cdots \ .$$

Repeated extrapolation

$$F_{k+1}(h) = F_k(h) + \frac{1}{q^{p_k} - 1}\big(F_k(h) - F_k(qh)\big) \quad k = 1, 2, \ldots \ .$$

Extrapolation scheme

$$\begin{array}{ccccc}
F_1(q^3h) & & & & \\
F_1(q^2h) & F_2(q^2h) & & & \\
F_1(qh) & F_2(qh) & F_3(qh) & & \\
F_1(h) & F_2(h) & F_3(h) & F_4(h) & \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

If h is sufficiently small, then the difference between two adjacent values in the same column gives an upper bound for the *truncation error*.

7. Integration

Numerical computation of

$$\int_a^b f(x)\,dx\ .$$

Equidistant points, $x_i = a + ih, \quad i = 0, 1, \ldots, m, \quad h = \dfrac{b-a}{m}$. Let $f_i = f(x_i)$.

Trapezoidal rule

$$T(h) = h\left(\tfrac{1}{2}f_0 + f_1 + \cdots + f_{m-1} + \tfrac{1}{2}f_m\right)\ .$$

Truncation error

$$R_{\mathrm{T}} = \int_a^b f(x)\,dx - T(h) = -\frac{b-a}{12}\,h^2 f''(\eta), \quad a < \eta < b\ ,$$

or

$$R_{\mathrm{T}} = a_1 h^2 + a_2 h^4 + \cdots\ .$$

If $\{\overline{f}_i\}$ are approximations to $\{f_i\}$ with $\max_i |\overline{f}_i - f_i| \le \epsilon$, then

$$|R_{\mathrm{XF}}| \le (b-a)\epsilon\ .$$

Simpson's formula

$$S(h) = \frac{h}{3}\left(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{m-2} + 4f_{m-1} + f_m\right)\ ,$$

where m is even. *Truncation error*

$$R_{\mathrm{T}} = \int_a^b f(x)\,dx - S(h) = -\frac{b-a}{180}\,h^4 f^{(4)}(\eta), \quad a < \eta < b\ ,$$

or

$$R_{\mathrm{T}} = b_1 h^4 + b_2 h^6 + \cdots\ .$$

Romberg's method

Trapezoidal method with repeated Richardson extrapolation, and successive halving of the step length ($q = 2$). Truncation error is estimated as in the general Richardson extrapolation.

Effect of erroneous function values: $|R_{\mathrm{XF}}| \le (b-a)\epsilon$.

8. Linear Systems of Equations

The system

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

can be written in matrix notation

$$Ax = b \ ,$$

where A is the $n{\times}n$ *coefficient matrix* and b is the $n{\times}1$ right hand side vector. We assume that A is nonsingular.

Triangular systems

$$
\begin{aligned}
u_{11}x_1 + u_{12}x_2 + \cdots + u_{1n}x_n &= c_1 \\
u_{22}x_2 + \cdots + u_{2n}x_n &= c_2 \\
&\vdots \\
u_{nn}x_n &= c_n
\end{aligned}
$$

can be solved by *back substitution*:

$$
\begin{aligned}
x_n &= c_n/u_{nn} \\
x_i &= \Big(c_i - \sum_{j=i+1}^{n} u_{ij}x_j\Big)/u_{ii} \ , \quad i = n{-}1, n{-}2, \ldots, 1 \ .
\end{aligned}
$$

Gaussian elimination

The system is transformed to upper triangular form

$$\big(\, A \mid b \,\big) \ \rightarrow \ \big(\, U \mid c \,\big)$$

in a series of $n{-}1$ steps. In the typical step the current system is

$$
\left(\begin{array}{cccccc|c}
a_{11} & a_{12} & & & \cdots & a_{1n} & b_1 \\
 & a_{22} & & & \cdots & a_{2n} & b_2 \\
 & & \ddots & & & \vdots & \vdots \\
 & & & a_{kk} & a_{k,k+1} \ \cdots & a_{kn} & b_k \\
 & & & \vdots & \vdots & \vdots & \vdots \\
 & & & a_{ik} & a_{i,k+1} \ \cdots & a_{in} & b_i \\
 & & & \vdots & \vdots & \vdots & \vdots \\
 & & & a_{nk} & a_{n,k+1} \ \cdots & a_{nn} & b_n
\end{array}\right) .
$$

The elements in the kth column below a_{kk} are zeroed by subtracting multiples of the kth row

$$
\left.
\begin{array}{rcll}
m_{ik} & := & a_{ik}/a_{kk} & \\
a_{ij} & := & a_{ij} - m_{ik}a_{kj}, & j = k{+}1, \ldots, n \\
b_i & := & b_i - m_{ik}b_k &
\end{array}
\right\} \quad i = k{+}1, \ldots, n \ .
$$

After $n{-}1$ steps A and b have been transformed to U and c, respectively, and x is computed by *back substitution*.

Partial pivoting

In each step determine the row index ν such that

$$|a_{\nu k}| = \max_{k \le i \le n} |a_{ik}| \; .$$

If $\nu > k$, then rows k and ν are interchanged, and the elimination proceeds. With partial pivoting the multipliers satisfy $|m_{ik}| \le 1$.

The purpose of pivoting is to avoid that matrix elements become too large during the elimination, with associated loss of accuracy. Pivoting is not needed if

a) A is *symmetric* and *positive definite* (*spd*), ie

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

or

b) A is *diagonally dominant*, ie

$$|a_{ii}| \ge \sum_{j=1, j\neq i}^{n} |a_{ij}| \; , \qquad i = 1, 2, \ldots, n \; ,$$

with strict inequality for at least one i.

LU Factorization

Gaussian elimination with partial pivoting applied to a nonsingular matrix A is equivalent to the factorization

$$P\,A = L\,U \; ,$$

where P is a permutation matrix, L is a unit lower triangular matrix, and U is an upper triangular matrix. L has diagonal elements equal to one and

$$(L)_{ik} = m_{ik} \; ,$$

where the m_{ik} are the multipliers used in the elimination.

If A is *spd*, then we can use the factorization

$$A = LDL^T \; ,$$

where L is a unit lower triangular matrix and D is a diagonal matrix with positive diagonal elements. Alternatively, we can use the *Cholesky factorization*

$$A = C^T C \; ,$$

where C is an upper triangular matrix.

Solution of $Ax^{[k]} = b^{[k]}, \; k = 1, 2, \ldots, K$ when the LU factorization is known:

$$\begin{array}{l} \textbf{for } k = 1, 2, \ldots, K \textbf{ do} \\ \quad \text{solve} \quad Ly^{[k]} = b^{[k]} \\ \quad \text{solve} \quad Ux^{[k]} = y^{[k]} \end{array}$$

Operation count

	Number of flops (floating point operations)
Transformation to triangular form (computation of the LU factorization)	$\frac{2}{3}\,n^3$
Computation of the LDL$^{\mathrm{T}}$or the Cholesky factorization of an *spd* matrix	$\frac{1}{3}\,n^3$
Solution of a triangular system	n^2
Matrix-vector multiplication	$2n^2$
Computation of A^{-1}	$2n^3$
Solution of a tridiagonal system (without pivoting)	$8n$

Vector and Matrix Norms

Vector norms

$$\text{Euclidean norm}: \quad \|x\|_2 = \left(\, x_1^2 + \cdots + x_n^2 \,\right)^{1/2} = \sqrt{x^T x}\ ,$$

$$\text{maximum norm}: \quad \|x\|_\infty = \max_{1\le i\le n} |x_i|\ .$$

Induced matrix norm

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \;=\; \max_{\|x\|=1} \|Ax\|\ ,$$

where $\|\cdot\|$ is a vector norm.

$$\|A\|_2 = \left(\max_{1\le j\le n} \lambda_j(A^T A)\right)^{1/2}\ , \qquad \text{(the square root of the largest eigenvalue of } A^T A\text{)}$$

$$\|A\|_\infty = \max_{1\le i\le n} \Big\{ \sum_{j=1}^{n} |a_{ij}| \Big\}\ .$$

From the definition it follows that $\quad \|Ax\| \le \|A\| \cdot \|x\|\ .$

Sensitivity analysis

Define the *condition number* of A,

$$\kappa(A) = \|A\| \cdot \|A^{-1}\|\ ,$$

and consider

$$\text{Exact system:} \qquad Ax = b\ ,$$

$$\text{perturbed system:} \qquad (A+\delta A)\overline{x} = b + \delta b\ .$$

If $\tau = \|A^{-1}\| \cdot \|\delta A\| \;=\; \kappa(A)\,\dfrac{\|\delta A\|}{\|A\|} \;<\; 1$, then

$$\frac{\|\overline{x} - x\|}{\|x\|} \;\le\; \frac{\kappa(A)}{1-\tau} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right)\ .$$

Estimate error in "given solution" $\widetilde{x}$.

$$\frac{\|\widetilde{x} - x\|}{\|x\|} \leq \kappa(A) \, \frac{\|r\|}{\|b\|} \, , \qquad r = b - A\widetilde{x} \, .$$

r is called the *residual*.

Rounding Errors in Gaussian Elimination

rule of thumb: If the unit roundoff and the condition number satisfy $\mu \simeq 10^{-d}$ and $\kappa_\infty(A) \simeq 10^q$, then a stable version of Gaussian elimination can be expected to produce a solution $\hat{x}$ that has about $d - q$ correct decimal digits.

Overdetermined Systems

Let A be an $m \times n$ with $m > n$ and linearly independent columns. The *least squares* problem

$$\min \|Ax - b\|_2$$

has a unique solution, which can be found by solving the *normal equations*

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

Alternatively, the least squares solution can be found via orthogonal transformation.

9. Approximation

Problem. Seek a function f^* that has minimum "distance" to either

a given function f on the interval $[a, b]$, (continuous case)

or

a given vector $f_{\mathrm{G}} = \big(f(x_1), f(x_2), \ldots, f(x_m)\big)^T$. (discrete case)

Use a norm to measure "distance".

Maximum norm (also called *Chebyshev norm*)

$$\|f\|_\infty = \begin{cases} \max\limits_{a \le x \le b} |f(x)| & \text{(continuous case)} , \\ \max\limits_{1 \le i \le m} |f(x_i)| & \text{(discrete case)} . \end{cases}$$

Euclidean norm

$$\|f\|_2 = \begin{cases} \Big(\displaystyle\int_a^b w(x) f(x)^2 \, dx\Big)^{1/2} & \text{(continuous case)} , \\ \Big(\displaystyle\sum_{i=1}^m w_i f(x_i)^2\Big)^{1/2} & \text{(discrete case)} . \end{cases}$$

w is a so-called *weight function*, $w(x) > 0$.

Scalar product

$$(f, g) = (g, f) = \begin{cases} \displaystyle\int_a^b w(x) f(x) g(x) \, dx & \text{(continuous case)} , \\ \displaystyle\sum_{i=1}^m w_i f(x_i) g(x_i) & \text{(discrete case)} . \end{cases}$$

In both the continuous and the discrete case

$$\|f\|_2 = (f, f)^{1/2} .$$

φ and ψ are said to be *orthogonal* if $(\varphi, \psi) = 0$.

The sequence $\varphi_0, \varphi_1, \ldots$ is called an *orthogonal system* if $(\varphi_i, \varphi_j) = 0$ for $i \neq j$ and $(\varphi_i, \varphi_i) \neq 0$ for all i. If, in addition, $(\varphi_i, \varphi_i) = 1$ for all i, the sequence is called an *orthonormal system.*

Least Squares Method

Seek a linear combination of the linearly independent functions $\varphi_0, \varphi_1, \ldots, \varphi_n$,

$$f^* = c_0^* \varphi_0 + c_1^* \varphi_1 + \cdots + c_n^* \varphi_n ,$$

such that $\|f - f^*\|_2$ is minimized. f^* is characterized by the *normal equations*

$$(\varphi_0, \varphi_k) c_0^* + (\varphi_1, \varphi_k) c_1^* + \cdots + (\varphi_n, \varphi_k) c_n^* = (f, \varphi_k), \quad k = 0, 1, \ldots, n .$$

If $\varphi_0, \varphi_1, \ldots, \varphi_n$ is an orthogonal system, we get the *orthogonal coefficients* (also called *Fourier coefficients*),

$$c_k^* = \frac{(f, \varphi_k)}{(\varphi_k, \varphi_k)} , \quad k = 0, 1, \ldots, n .$$

Orthogonal Polynomials

Given a scalar product and the leading coefficients $A_0, A_1, \ldots$, the polynomials $P_k(x) = A_k x^k + \cdots$ constructed by the recurrence

$$\begin{aligned} P_0(x) &= A_0 \\ P_1(x) &= (\alpha_0 x - \beta_0) P_0(x) \\ P_{k+1}(x) &= (\alpha_k x - \beta_k) P_k(x) - \gamma_k P_{k-1}(x), \quad k = 1, 2, \ldots \ , \end{aligned}$$

where

$$\begin{aligned} \alpha_k &= \frac{A_{k+1}}{A_k} \ , & k &= 0, 1, 2, \ldots \ , \\ \beta_k &= \frac{\alpha_k (x P_k, P_k)}{(P_k, P_k)} \ , & k &= 0, 1, 2, \ldots \ , \\ \gamma_k &= \frac{\alpha_k (P_k, P_k)}{\alpha_{k-1} (P_{k-1}, P_{k-1})} \ , & k &= 1, 2, \ldots \ , \end{aligned}$$

form an orthogonal system. In the discrete case, with the grid $x_1, x_2, \ldots, x_m$, the last polynomial in the sequence is P_{m-1}.

Transformation of variable between $a \le x \le b$ and $-1 \le t \le 1$,

$$t = \frac{2x - (b + a)}{b - a} \ , \qquad x = \tfrac{1}{2}(b - a)t + \tfrac{1}{2}(a + b) \ .$$

Legendre Polynomials

$$\int_{-1}^{1} P_k(x) P_n(x) \, dx = \begin{cases} 0 & \text{for } k \neq n \ , \\ \dfrac{2}{2n + 1} & \text{for } k = n \ . \end{cases}$$

$$P_n(x) = \frac{1}{2^n \cdot n!} \, \frac{d^n}{dx^n} \, (x^2 - 1)^n \ .$$

Recurrence,

$$\begin{aligned} &P_0(x) = 1 \ , \quad P_1(x) = x \ , \\ &P_{n+1}(x) = \frac{2n + 1}{n + 1} \, x P_n(x) - \frac{n}{n + 1} \, P_{n-1}(x), \quad n = 1, 2, \ldots \ . \end{aligned}$$

First five Legendre polynomials

$$\begin{aligned} &P_0(x) = 1, \qquad P_1(x) = x, \qquad P_2(x) = \frac{1}{2}(3x^2 - 1) \ , \\ &P_3(x) = \frac{1}{2}(5x^2 - 3x), \qquad P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3) \ . \end{aligned}$$

Chebyshev Polynomials

$$\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} T_k(x) T_n(x)\, dx = \begin{cases} 0 & \text{for } k \neq n\ , \\ \frac{1}{2}\pi & \text{for } k = n > 0\ , \\ \pi & \text{for } k = n = 0\ . \end{cases}$$

$$T_n(x) = \cos(n \arccos x)\ .$$

Recurrence,

$$\begin{aligned} &T_0(x) = 1\ , \quad T_1(x) = x\ , \\ &T_{n+1}(x) = 2x\, T_n(x) - T_{n-1}(x), \quad n = 1, 2, \ldots\ . \end{aligned}$$

First five Chebyshev polynomials

$$\begin{aligned} &T_0(x) = 1, \qquad T_1(x) = x, \qquad T_2(x) = 2x^2 - 1\ , \\ &T_3(x) = 4x^3 - 3x, \qquad T_4(x) = 8x^4 - 8x^2 + 1\ . \end{aligned}$$

Zeros of T_n (*Chebyshev nodes*),

$$x_i = \cos\left(\frac{2i-1}{2n}\,\pi\right), \quad i = 1, 2, \ldots, n\ .$$

T_n oscillates between ± 1 in the points

$$\widetilde{x}_k = \cos\left(\frac{k}{n}\,\pi\right), \quad k = 0, 1, \ldots, n\ .$$

Discrete Cosine Transform (DCT)

The functions $\varphi_0, \varphi_1, \ldots, \varphi_{m-1}$, defined by

$$\varphi_k(x) = \alpha_k \cos kx, \qquad \alpha_k = \begin{cases} \sqrt{1/m}\ , & k = 0 \\ \sqrt{2/m}\ , & k > 0\ . \end{cases}$$

form an orthonormal system with respect to the scalar product

$$(u, v) = \sum_{l=1}^{m} u(x_l) \cdot v(x_l)\ , \qquad x_l = \frac{(2l-1)\pi}{2m}\ .$$

Given a *signal*, ie a vector $f_{\mathrm{G}} \in \mathbb{R}^m$. Its DCT is

$$c = (c_0, c_1, \ldots, c_{m-1})^T\ , \qquad c_j = \varphi_{j\mathrm{G}}^T f_{\mathrm{G}}\ .$$

Given the DCT c, the signal can be found by the *inverse discrete cosine transform (IDCT)*

$$f_{\mathrm{G}} = \sum_{j=0}^{m-1} c_j \varphi_{j\mathrm{G}}\ .$$

Minimax (Chebyshev) Approximation

Find the polynomial p^* of degree $\leq n$ such that

$$E_n(f) = \|f - p_n^*\|_\infty \leq \|f - p_n\|_\infty \quad \text{for all polynomials } p_n \text{ of degree } \leq n \ .$$

Alternation property: Assume that $f \in C[a,b]$. p_n^* is the best maximum norm approximation of f if and only if there are points $a \leq \xi_1 < \xi_2 < \cdots < \xi_{n+2} \leq b$ such that

$$|f(\xi_k) - p_n^*(\xi_k)| = \|f - p_n^*\|_\infty, \quad k = 1, 2, \ldots, n{+}2$$

and

$$f(\xi_{k+1}) - p_n^*(\xi_{k+1}) = -\left(f(\xi_k) - p_n^*(\xi_k)\right), \quad k = 1, 2, \ldots, n{+}1 \ .$$

Approximation to p_n^* by *Chebyshev interpolation*: Transform the range $[a,b]$ to $[-1,1]$ and use interpolation points

$$x_i = \cos\left(\frac{2i+1}{2(n+1)}\,\pi\right), \quad i = 0, 1, \ldots, n \ .$$

Maximum error is at most $5E_n(f)$ if $n \leq 100$.

10. Ordinary Differential Equations

Initial Value Problem

$$y' = f(x, y), \qquad y(a) = \alpha \ .$$

Seek the solution on the range $[a, b]$. Introduce a *grid* with *step length* h

$$x_n = a + nh, \quad n = 0, 1, \ldots, N, \qquad h = \frac{b-a}{N} \ .$$

Find approximations y_n to $y(a + nh)$

Local truncation error at x_{n+1} is the difference between the computed value y_{n+1} and the value at x_{n+1} on the solution curve that passes through the point (x_n, y_n).

Global truncation error at x_{n+1} is the difference $R_{\mathrm{T}} = y(x_{n+1}) - y_{n+1}$, where $y(x)$ is the solution of the given initial value problem.

Stability. When the numerical method is applied to the *test problem*

$$y' = \lambda y, \qquad y(0) = 1 \ ,$$

with $\lambda < 0$, the sequence $y_1, y_2, \ldots$ should be decreasing.

Euler's method

$$\begin{aligned} y_0 &= \alpha \ , \\ y_{n+1} &= y_n + hf(x_n, y_n), \quad n = 0, 1, \ldots, N-1 \ . \end{aligned}$$

Local truncation error $O(h^2)$. Global truncation error $|R_{\mathrm{T}}| = O(h)$.
The method is stable for $h < 2/|\lambda|$.

Heun's method

$$\begin{aligned} k_1 &= f(x_n, y_n) \ , \\ k_2 &= f(x_n + h, y_n + hk_1) \ , \\ y_{n+1} &= y_n + \frac{h}{2}(k_1 + k_2) \ . \end{aligned}$$

$|R_{\mathrm{T}}| = O(h^2)$. The method is stable for $h < 2/|\lambda|$.

Classical Runge-Kutta method

$$\begin{aligned} k_1 &= f(x_n, y_n) \ , \\ k_2 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_1) \ , \\ k_3 &= f(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hk_2) \ , \\ k_4 &= f(x_n + h, y_n + hk_3) \ , \\ y_{n+1} &= y_n + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \ . \end{aligned}$$

$|R_{\mathrm{T}}| = O(h^4)$. The method is stable for $h < 2.785/|\lambda|$.

Trapezoidal method (an *implicit* method

$$y_{n+1} = y_n + \tfrac{1}{2}h\big(f(x_n, y_n) + f(x_{n+1}, y_{n+1})\big) \ .$$

$|R_{\mathrm{T}}| = O(h^2)$. Stable for all $h > 0$.

Boundary Value Problems

$$y'' = \psi(x, y, y'), \qquad y(a) = \alpha, \quad y(b) = \beta \ .$$

A difference method. Introduce a grid $x_n = a + nh$, $n = 0, 1, \ldots, N; h = \dfrac{b-a}{N}$, and approximate derivatives by central differences,

$$y''(x_n) \simeq \frac{y(x_{n-1}) - 2y(x_n) + y(x_{n+1})}{h^2} \ , \qquad y'(x_n) \simeq \frac{y(x_{n+1}) - y(x_{n-1})}{2h} \ .$$

Use these in the differential equation for $x = x_1, \ldots, x_{N-1}$; replace "$\simeq$" by "$=$" and $y(x_k)$ by the approximation y_k,

$$\frac{y_{n-1} - 2y_n + y_{n+1}}{h^2} = \psi\big(x_n, y_n, \frac{y_{n+1} - y_{n-1}}{2h}\big), \quad n = 1, \ldots, N-1 \ ,$$

and supply with the boundary conditions: $y_0 = \alpha$, $y_N = \beta$. This is a (possibly nonlinear) system of $N-1$ equations in the $N-1$ unknowns $y_1, \ldots, y_{N-1}$.

Truncation error $O(h^2)$.

A finite element method – Galerkin's method

$$Ly = -y'' + qy = f, \qquad y(a) = y(b) = 0 \ .$$

Let $\mathbb{V}$ be a class of *test functions*, that satisfy the boundary conditions

$$\mathbb{V} = \big\{ v \mid v' \text{ is piecewise continuous and bounded on } [a, b], \\ \text{and } v(a) = v(b) = 0 \ \big\} \ .$$

Weak formulation of the boundary value problem,

$$(v, Ly) = (v', y') + q(v, y) = (v, f) \quad \text{for all } v \in \mathbb{V} \ .$$

Choose $\mathbb{V} = \operatorname{span}\{\varphi_j\}_{j=1}^{N-1}$ and $y^h = \displaystyle\sum_{j=1}^{N-1} c_j \varphi_j$. The coefficients satisfy a linear system $(K_0 + K_1)c = F$, where

$$\big(K_1\big)_{ij} = (\varphi_i', \varphi_j'), \quad \big(K_0\big)_{ij} = q(\varphi i, \varphi j), \quad F_i = (\varphi_i, f) \ .$$

The shooting method

Let $g(\gamma)$ denote the value at $x = b$ obtained by numerical solution of the initial value problem

$$y'' = \psi(x, y, y'), \qquad y(a) = \alpha, \quad y'(a) = \gamma \ .$$

Solve the equation (eg by means of the secant method)

$$g(\gamma) - \beta = 0 \ .$$