

Numerical Mathematics II

SS '03

Prof. Dr. E. P. Stephan

30th June 2003

Contents

A	Numerical methods for ordinary differential equations	3
1	Numerical methods for initial value problems	3
1.1	Single-step methods	3
1.2	Multistep Methods	15
1.3	Convergence and consistency of single-step methods	22
1.4	Numerical stability	26
2	Numerical methods for boundary value problems	31
2.1	Shooting methods and collocation methods	31
2.2	Difference methods	34
2.3	Variational methods, Ritz-methods	36
B	Numerical methods for partial differential equations	47
1	Finite differences for elliptic equations	47
1.1	The finite difference method	47
1.2	Convergence of point iteration methods	48
1.3	An example	51
2	Difference Methods for Parabolic Equations	53
2.1	Parabolic equations	54
C	The CG algorithm revisited	57
1	Conjugate Gradient Method	57
2	The Preconditioned Conjugate Gradient Method	59

D Eigenvalues of symmetric matrices	62
1 The von-Mises method (power method, 1929)	62
2 The Jacobi method (1846)	66
3 The QR method	67
3.1 The QR decomposition	67
3.2 QR decomposition and linear equations systems	69
3.3 Shift	69
3.4 The QR method (1961 Francis & Kublanowskaj)	70
E Approx. of periodic functions with trigonometric polynomials	73
1 Representation theorem	73
2 Fast Fourier Transformation	75
F Discrete approximation problems	79
G Calculus of variations - the Euler differential equation	83
1 Introduction	83
2 Euler's differential equation	84
2.1 Derivation of a necessary condition for the solution of (1.1)	84
2.2 Special cases of Euler's differential equation	87
3 Extensions and generalizations	88
3.1 Natural boundary conditions	88
3.2 Variational problems in parametric representation	89
3.3 Isoperimetric problems	90
3.4 Lagrange's and Hilbert's Problem	92
3.5 Several functions in the basic integral	92
3.6 Higher derivatives in the basic integral	92
3.7 Weighted variational problems	93
3.8 The Ritz method revisited	95
3.9 Variational problems for two independent variables	96

Chapter A

Numerical methods for ordinary differential equations

1 Numerical methods for initial value problems

1.1 Single-step methods

We consider the **initial value problem (IVP)** :

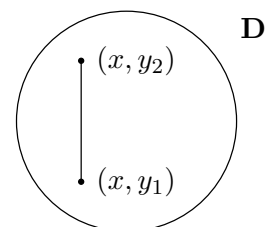
$$(1.1) \quad y' = f(x, y) , \quad y(x_0) = Y_0$$

with $f(x, y) \in C(D)$, $(x_0, Y_0) \in D$, D domain in \mathbb{R}^2 .

Definition 1.1

$Y(x)$ solves (1.1) in $[a, b]$

$$\begin{aligned} \Leftrightarrow \quad \forall a \leq x \leq b : \quad & (1) \quad (x, Y(x)) \in D \\ & (2) \quad Y(x_0) = Y_0 \\ & (3) \quad Y'(x) \text{ exists and } Y'(x) = f(x, Y(x)). \end{aligned}$$



We state the following theorem concerning existence and uniqueness of a solution to (1.1) without proof. Later on (theorem (1.14)) we will give a proof under the restriction that the domain D is bounded.

Theorem 1.2

If $f \in C(D)$, $(x_0, Y_0) \in D$ and f Lipschitz-continuous, i.e.

$$(1.2) \quad \exists K > 0 \quad \forall (x, y_1), (x, y_2) \in D \quad |f(x, y_1) - f(x, y_2)| \leq K |y_1 - y_2| ,$$

then there exists exactly one solution

$$Y(x) \text{ of (1.1) in } I = [x_0 - \alpha, x_0 + \alpha] .$$

Remark 1.3

If $\frac{\partial f}{\partial y}$ is bounded on D then (1.2) holds, since with

$$K := \max_{(x,y) \in D} \left| \frac{\partial f(x,y)}{\partial y} \right|$$

and

$$f(x, y_1) - f(x, y_2) = \frac{\partial f(x, \xi)}{\partial y} (y_1, y_2) \quad , \quad \xi \in (y_1, y_2)$$

one verifies (1.2).

Example 1.4

(a) $y' = 1 + \sin(x, y)$ on $D = \{(x, y) \mid 0 \leq x \leq 1, -\infty < y < \infty\}$.

Then

$$\frac{\partial f}{\partial y} = x \cos(x, y) \quad \Rightarrow \quad K = 1$$

$\Rightarrow \forall (x_0, Y_0), 0 < x_0 < 1 : \exists Y(x)$ solution of (IVP) on $[x_0 - \alpha, x_0 + \alpha] \subset [0, 1]$.

(b) $y' = \frac{2x}{a^2} y^2, y(0) = 1, a > 0 \Rightarrow Y(x) = \frac{a^2}{a^2 - x^2}, -a < x < a.$

Then

$$\frac{\partial f}{\partial y} = \frac{4xy}{a^2}$$

The Lipschitz constant K is bounded if D is bounded :

$-c \leq x \leq c, -b \leq y \leq b \Rightarrow$ Then (1.2) yields :

\exists solution $Y(x)$ in $-\alpha \leq x \leq \alpha$ with $\alpha \leq c.$

1.1.1 Conditions of stability

We now consider the **perturbed problem** :

$$(1.3) \quad \begin{aligned} y' &= f(x, y) + \delta(x), \quad \delta \in C^0 \\ y(x_0) &= Y_0 + \epsilon. \end{aligned}$$

The next theorem gives information about the stability of (1.3) depending on the perturbation δ .

Theorem 1.5

Assumptions as in theorem (1.2)

$\Rightarrow \exists$ solution $Y(x, \delta, \epsilon)$ of (1.3) on $[x_0 - \alpha, x_0 + \alpha]$ uniformly in ϵ and $\delta(x)$ with $|\epsilon| \leq \epsilon_0, \|\delta\|_\infty \leq \epsilon_0, \epsilon_0$ sufficiently small.

Furthermore

$$\max_{|x-x_0| \leq \alpha} |Y(x) - Y(x, \delta, \epsilon)| \leq k [|\epsilon| + \alpha \|\delta\|_\infty]$$

with $k = \frac{1}{1-\alpha K}$, where Y solves (1.1). (K is the Lipschitz constant in (1.2).)

The previous theorem can be applied analogously to **first order systems of differential equations** :

$$\left. \begin{array}{l} y_1' = f_1(x, y_1, \dots, y_m), \quad y_1(x_0) = Y_{1,0} \\ \vdots \\ y_m' = f_m(x, y_1, \dots, y_m), \quad y_m(x_0) = Y_{m,0} \end{array} \right\} \mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{Y}_0$$

where

$$\mathbf{y}(x) = \begin{pmatrix} y_1(x) \\ \vdots \\ y_m(x) \end{pmatrix}, \quad \mathbf{f}(x, \mathbf{y}) = \begin{pmatrix} f_1(x, \mathbf{y}) \\ \vdots \\ f_m(x, \mathbf{y}) \end{pmatrix}, \quad \mathbf{Y}_0 = \begin{pmatrix} Y_{1,0} \\ \vdots \\ Y_{m,0} \end{pmatrix}.$$

Higher order equations must be transformed into **first order systems** :

$$y^{(m)} = f(x, y, y', \dots, y^{(m-1)}) \quad \text{with initial values (IVs)} \quad \left\{ \begin{array}{l} y(x_0) = Y_0 \\ \vdots \\ y^{(m-1)}(x_0) = Y_0^{(m-1)} \end{array} \right. .$$

$$\text{Define new unknowns} \quad \left\{ \begin{array}{l} y_1 := y \\ y_2 := y' \\ \vdots \\ y_m := y^{(m-1)} \end{array} \right.$$

$$\Rightarrow \left\{ \begin{array}{l} y_1' = y_2 \\ y_2' = y_3 \\ \vdots \\ y_{m-1}' = y_m \\ y_m' = f(x, y_1, \dots, y_m) \end{array} \right. \quad \text{with (IVs)} \quad \left\{ \begin{array}{l} y_1(x_0) = Y_0 \\ y_2(x_0) = Y_0' \\ \vdots \\ y_{m-1}(x_0) = Y_0^{(m-2)} \\ y_m(x_0) = Y_0^{(m-1)} \end{array} \right. .$$

Example 1.6

$$y'' = a_1(x)y' + a_0(x)y + g(x), \quad y(x_0) = \alpha, \quad y'(x_0) = \beta$$

$$\Rightarrow \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \underbrace{\begin{pmatrix} 0 & 1 \\ a_0(x) & a_1(x) \end{pmatrix}}_{A(x)} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \underbrace{\begin{pmatrix} 0 \\ g(x) \end{pmatrix}}_{G(x)}, \quad \underbrace{\begin{pmatrix} y_1(x_0) \\ y_2(x_0) \end{pmatrix}}_{\mathbf{y}(x_0)} = \underbrace{\begin{pmatrix} \alpha \\ \beta \end{pmatrix}}_{\mathbf{Y}_0}$$

$$\Leftrightarrow \mathbf{y}' = A(x)\mathbf{y} + G(x), \quad \mathbf{y}(x_0) = \mathbf{Y}_0.$$

1.1.2 Euler's Method

Given (IVP) :

$$\begin{aligned}y'(x) &= f(x, y(x)) \\ y(x_0) &= Y_0.\end{aligned}$$

Set

$$x_0 < x_1 < x_2 < \dots < x_n < \dots \quad \text{with } x_j = x_0 + jh$$

and

$$y_0 := y(x_0), y_1 := y(x_1), \dots, y_n := y(x_n).$$

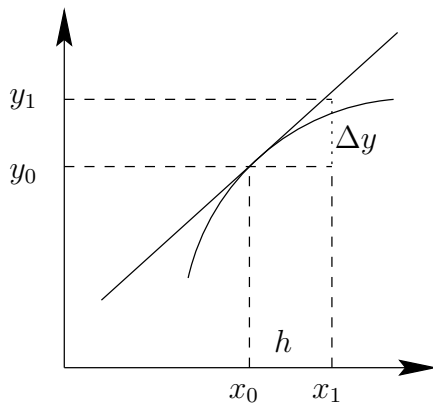
Denote by y_n the approximation obtained via

Euler's Method :

$$(1.4) \quad \begin{aligned}y_{n+1} &= y_n + hf(x_n, y_n), \quad n = 0, 1, 2, \dots \\ y_0 &= Y_0\end{aligned}$$

There are different possible interpretations of (1.4):

Geometrical interpretation :



with

$$\begin{aligned}\frac{\Delta y}{h} &= Y'(x_0) = f(x_0, Y_0) \\ Y(x_1) - Y(x_0) &\simeq \Delta y = hY'(x_0) \\ \Rightarrow Y(x_1) &\simeq Y(x_0) + hf(x_0, Y(x_0))\end{aligned}$$

Taylor series : Expand $Y(x_{n+1})$ around x_n

$$(1.5) \quad Y(x_{n+1}) = Y(x_n) + hY'(x_n) + T_n,$$

with the local discretization error (truncation) at x_{n+1} :

$$(1.6) \quad T_n := \frac{h^2}{2}Y''(\xi_n), \quad x_n \leq \xi_n \leq x_{n+1}.$$

Numerical differentiation :

$$\begin{aligned}\frac{Y(x_{n+1}) - Y(x_n)}{h} &\approx Y'(x_n) = f(x_n, Y(x_n)) \\ \Rightarrow Y(x_{n+1}) &\approx Y(x_n) + hf(x_n, Y(x_n)).\end{aligned}$$

Numerical integration : Integrate $Y'(t) = f(t, Y(t))$ over $[x_n, x_{n+1}]$:

$$Y(x_{n+1}) = Y(x_n) + \underbrace{\int_{x_n}^{x_{n+1}} f(t, Y(t)) dt}_{\approx hf(x_n, Y(x_n))}.$$

Example 1.7

	x	$y_n(x)$	$Y(x)$	$[Y(x) - y_n(x)]$	
h = 0.2	0.4	1.44	1.49	0.05	$y' = y, y(0) = 1$
	0.8	2.07	2.23	0.16	
	1.2	2.99	3.32	0.33	$\Rightarrow \int_1 \frac{dy}{y} = \int_0 dx$
	1.6	4.31	4.95	0.64	$\Rightarrow \ln(y) - \ln(1) = x.$
	2	6.21	7.39	1.18	
h = 0.1	0.4	1.46	1.49	0.03	
	0.8	2.14	2.23	0.09	Hence
	1.2	3.14	3.32	0.18	$Y(x) = e^x.$
	1.6	4.59	4.95	0.36	
	2	6.73	7.39	0.66	

\Rightarrow error reduced by half when h halved.

1.1.3 Convergence

At each step of the Euler method there arises an additional error (1.6). The total error $Y(x) - y_n(x)$ is called **global discretization error**.

Example 1.8

$$y' = 2x, y(0) = 0 \Rightarrow Y(x) = x^2$$

Euler : $y_{n+1} = y_n + 2hx_n, y_0 = 0$

Induction : $y_n = x_{n-1} x_n, n \geq 1$

Error : $Y(x_n) - y_n = x_n^2 - x_n x_{n-1} = x_n \underbrace{(x_n - x_{n-1})}_h = h x_n$

\Rightarrow global error (at each fixed point x) $\approx h$

Lemma 1.9

For all $x \in \mathbb{R}$ there holds : $1 + x \leq e^x$
 and for all $x \in \mathbb{R}, x \geq -1$ there holds : $0 \leq (1 + x)^m \leq e^{mx}$.

Proof:

The assertion follows easily by considering the truncated taylor expansion :

$$e^x = 1 + x + \underbrace{\frac{x^2}{2}}_{>0} e^\xi, 0 < \xi < x.$$

■

General assumption : f satisfies the strong Lipschitz-condition, i.e.

$$(1.7) \quad \exists K > 0 \quad |f(x, y_1) - f(x, y_2)| \leq K |y_1 - y_2|, \quad -\infty < y_1, y_2 < \infty, \quad x_0 \leq x \leq b.$$

Theorem 1.10

Let $Y(x)$ be the solution of (1.1) with $|Y''(x)| \leq c < \infty \quad \forall x \in [x_0, b]$ and let $y_h(x_n)$ ($x_0 \leq x_n \leq b$) solve Euler (1.4). Then there holds

$$(1.8) \quad \max_{x_0 \leq x_n \leq b} |Y(x_n) - y_h(x_n)| \leq e^{(b-x_0)K} |e_0| + \left[\frac{e^{(b-x_0)K} - 1}{K} \right] \tau(h),$$

in which

$$\tau(h) := \frac{h}{2} \|Y''\|_\infty, \quad e_0 := Y_0 - y_h(x_0).$$

If further there holds

$$|Y_0 - y_h(x_0)| \leq c_1 h \quad (h \rightarrow 0) \text{ with } c_1 \geq 0,$$

then there exists $B \geq 0$ such that

$$(1.9) \quad \max_{x_0 \leq x_n \leq b} |Y(x_n) - y_h(x_n)| \leq B h \quad (x_n \leq b, \quad x_{n+1} > b).$$

Proof:

Let

$$e_n := Y(x_n) - y_n(x_n), \quad n \geq 0.$$

Define

$$\tau_n := \frac{h}{2} Y''(\xi_n) \quad (0 \leq n \leq N(h)) \quad \Rightarrow \quad \max_{0 \leq n \leq N-1} |\tau_n| \leq \tau(h)$$

and $Y_n := Y(x_n)$. Then (1.5), (1.1) and (1.4) lead to

$$\begin{aligned} Y_{n+1} &= Y_n + hf(x_n, Y_n) + h\tau_n \\ y_{n+1} &= y_n + hf(x_n, y_n) \quad , \quad 0 \leq n \leq N(h) - 1. \end{aligned}$$

Then

$$\begin{aligned} e_{n+1} &= e_n + h[f(x_n, Y_n) - f(x_n, y_n)] + h\tau_n \\ \stackrel{(1.7)}{\Rightarrow} |e_{n+1}| &\leq |e_n| + hK|Y_n - y_n| + h|\tau_n| \\ \Leftrightarrow |e_{n+1}| &\leq (1 + hK)|e_n| + h|\tau(h)| \quad , \quad 0 \leq n \leq N(h) - 1. \end{aligned}$$

Recursion gives us

$$|e_n| \leq (1 + hK)^n |e_0| + \left[1 + (1 + hK) + \dots + (1 + hK)^{n-1} \right] h\tau(h).$$

The summation formula for the finite geometrical series $1 + r + r^2 + \dots + r^{n-1} = \frac{r^n - 1}{r - 1}$ ($r \neq 1$) yields

$$|e_n| \leq (1 + hK)^n |e_0| + \left[\frac{(1 + hK)^n - 1}{K} \right] \tau(h).$$

Using Lemma 1.9, we now obtain

$$(1 + hK)^n \leq e^{nhK} = e^{(x_n - x_0)K} \leq e^{(b-x_0)K}.$$

All in all, we have proven (1.8).

Set

$$B := c_1 e^{(b-x_0)K} + \left(\frac{e^{(b-x_0)K} - 1}{K} \right) \frac{\|Y''\|_\infty}{2},$$

then (1.9) follows from (1.8). ■

Remark 1.11

It follows from (1.9) that the error is halved (at least) if the step length is halved.

Example 1.12

x_n	$Y_n - y_n$	$hD(x_n)$
0.4	0.00689	0.00670
0.8	0.00920	0.00899
1.2	0.00921	0.00904
1.6	0.00819	0.00808
2.0	0.00682	0.00677

$$y' = -y, \quad y(0) = 1 \quad \Rightarrow \quad Y(x) = e^{-x}$$

From (1.8) we get

$$|Y(x_n) - y_h(x_n)| \leq \frac{h}{2}(e^{x_n} - 1).$$

This is a poor error bound, as it grows exponentially!

$$D'(x) = -D(x) + \frac{1}{2}e^{-x}, \quad D(0) = 0 \quad \Rightarrow \quad D(x) = \frac{1}{2}xe^{-x}$$

$$\Rightarrow \quad Y(x_n) - y_h(x_n) = \frac{h}{2}x_n e^{-x_n}.$$

1.1.4 Asymptotic error analysis (without rounding errors)

For h sufficiently small there holds

$$\left(B(x, h) = O(h^p), \quad p > 0 \quad :\Leftrightarrow \quad \exists c > 0 : |B(x, h)| \leq ch^p, \quad x_0 \leq x \leq b \right)$$

Theorem 1.13

Let $Y \in C^3$ solve (1.1), and let the partial derivatives f_y, f_{yy} be continuous and bounded on $x_0 \leq x \leq b, -\infty < y < \infty$. Further let

$$Y_0 - y_h(x_0) = \delta_0 h + O(h^2)$$

(usually $\delta_0 = 0$). Then the error of Euler's method satisfies the equation

$$(1.10) \quad Y(x_n) - y_h(x_n) = D(x_n)h + O(h^2)$$

where $D(x)$ is the solution of (IVP) :

$$(1.11) \quad D'(x) = f_y(x, Y(x))D(x) + \frac{1}{2}Y''(x), \quad D(x_0) = \delta_0$$

(cf. Example 1.12).

Proof:

Taylor expansion gives

$$Y(x_{n+1}) = Y(x_n) + hY'(x_n) + \frac{h^2}{2}Y''(x_n) + \frac{h^3}{6}Y^{(3)}(\xi_n), \quad x_n \leq \xi_n \leq x_{n+1}.$$

Euler's method (1.4) is

$$Y_{n+1} = Y_n + hf(x_n, Y_n),$$

so we have

$$e_{n+1} = e_n + h[f(x_n, Y_n) - f(x_n, y_n)] + \frac{h^2}{2}Y''(x_n) + \frac{h^3}{6}Y^{(3)}(\xi_n).$$

The Taylor expansion of $f(x_n, y_n)$ as a function of y_n is

$$f(x_n, y_n) = f(x_n, Y_n) + (y_n - Y_n)f_y(x_n, Y_n) + \frac{1}{2}(y_n - Y_n)^2 f_{yy}(x_n, \xi_n), \quad y_n \leq \xi_n \leq Y_n.$$

Inserting this in e_{n+1} yields

$$(1.12) \quad e_{n+1} = [1 + hf_y(x_n, Y_n)]e_n + \frac{h^2}{2}Y''(x_n) + B_n,$$

where

$$(1.13) \quad B_n = \frac{h^3}{6}Y^{(3)}(\xi_n) - \frac{1}{2}hf_{yy}(x_n, \xi_n)e_n^2 \stackrel{(1.9)}{\Rightarrow} B_n = O(h^3).$$

Now let g_n be the dominant part of the error:

$$(1.14) \quad g_{n+1} = [1 + hf_y(x_n, Y_n)]g_n + \frac{h^2}{2}Y''(x_n) \quad \text{mit } g_0 = \delta_0 h.$$

Then there holds $g_n \approx e_n$ (where $e_n = O(h)$). Therefore set $g_n = h\delta_n$:

$$\delta_{n+1} = \delta_n + h[f_y(x_n, Y_n)\delta_n + \frac{1}{2}Y''(x_n)] \quad , \quad x_0 \leq x_n \leq b$$

and with Euler's method for (1.11) there follows from Theorem 1.10:

$$D(x_n) - \delta_n = O(h) \quad , \quad x_0 \leq x_n \leq b \quad \Rightarrow \quad g_n = D(x_n)h + O(h^2).$$

The theorem is proven if we can show that g_n is the main part of the error e_n .

To see this, set $k_n := e_n - g_n$ with $k_0 = e_0 - g_0 = O(h^2)$.

Subtract (1.14) from (1.12) while using (1.13):

$$\begin{aligned} k_{n+1} &= [1 + hf_y(x_n, Y_n)]k_n + B_n \\ |k_{n+1}| &\leq (1 + hK)|k_n| + O(h^3). \end{aligned}$$

This form corresponds to (1.10) with $h\tau(h)$ instead of $O(h^3)$. As in the proof of Theorem 1.10 there follows $|k_n| = O(h^2)$ and

$$e_n = g_n + k_n = [hD(x_n) + O(h^2)] + O(h^2)$$

which proves (1.10). ■

Theorem 1.14 (existence and uniqueness)

Let $D = \{(x, y) \in \mathbb{R}^2 \mid x_0 - \delta_1 < x < x_0 + \delta_2, \quad y_0 - \eta_1 < y < y_0 + \eta_2\}$ for some $\delta_1, \delta_2, \eta_1, \eta_2 \in \mathbb{R}^+$ and $f \in C(D)$ for $(x, y) \in D$ such that

$$\exists M \in \mathbb{R}, M > 0 \quad \forall (x, y) \in D \quad |f(x, y)| < M.$$

With the Lipschitz condition

$$\exists L \in \mathbb{R}, L > 0 \quad \forall (x, y), (x, z) \in D \quad |f(x, y) - f(x, z)| < L|y - z|$$

there follows

$$\frac{1}{\exists} y \in C^1(D) \quad (y' = f(x, y), \quad y(x_0) = y_0) \quad (\text{see (1.1)}).$$

Proof:(a) *Existence* :

We consider the integral form of (1.1):

$$(1.15) \quad y = y_0 + \int_{x_0}^x f(\xi, y) d\xi .$$

As this is a fixed point equation, we would like to show that Picard's iteration (successive approximation) converges towards a solution \tilde{y} of (1.15), which is then also a solution of (1.1). Now we choose an initial approximation $y = y_1(x)$ with $|y_1'| < M$ and $y_1(x_0) = y_0$, for example $y_1(x) := y_0$. Then

$$y_{i+1}(x) = y_0 + \int_{x_0}^x f(\xi, y_i(\xi)) d\xi \quad , \quad i = 1, 2, 3, 4, \dots$$

$$\begin{aligned} \Rightarrow \quad |y_{i+1}(x) - y_i(x)| &\leq \int_{x_0}^x |f(\xi, y_i(\xi)) - f(\xi, y_{i-1}(\xi))| d\xi \\ &\leq L \int_{x_0}^x |y_i(\xi) - y_{i-1}(\xi)| d\xi \quad (i = 1, 2, 3, 4, \dots) . \end{aligned}$$

Suppose $x_0 \leq x \leq X$, $h := X - x_0$. Then

$$|y_2(x) - y_1(x)| = \left| y_0 - y_1(x) + \int_{x_0}^x f(\xi, y_1(\xi)) d\xi \right| \leq |y_0 - y_1| + \int_{x_0}^x M d\xi \leq 2 M h =: N .$$

Successively

$$\begin{aligned} |y_3(x) - y_2(x)| &\leq L \int_{x_0}^x N d\xi = N L (x - x_0) \leq N L h, \\ |y_4(x) - y_3(x)| &\leq L \int_{x_0}^x N L (\xi - x_0) d\xi = N \frac{L^2 (x - x_0)^2}{2!} \leq N \frac{L^2 h^2}{2!}, \\ &\vdots \\ |y_{n+1}(x) - y_n(x)| &\leq L \int_{x_0}^x N \frac{L^{n-2} (\xi - x_0)^{n-2}}{(n-2)!} d\xi = N \frac{L^{n-1} (x - x_0)^{n-1}}{(n-1)!} \leq N \frac{L^{n-1} h^{n-1}}{(n-1)!} . \end{aligned}$$

But

$$\begin{aligned} y_{n+1}(x) &= y_1(x) + \underbrace{[y_2(x) - y_1(x)] + \dots + [y_{n+1}(x) - y_n(x)]}_{\substack{\text{---} \\ \xrightarrow{(n \rightarrow \infty)} e^{Lh}}} \\ &\leq y_1(x) + N \left(\underbrace{1 + Lh + \frac{L^2 h^2}{2!} + \dots + \frac{L^{n-1} h^{n-1}}{(n-1)!}}_{\substack{\text{---} \\ \xrightarrow{(n \rightarrow \infty)} e^{Lh}}} \right) \end{aligned}$$

$$\Rightarrow \quad \tilde{y}(x) := \lim_{(n \rightarrow \infty)} y_{n+1}(x) \quad \text{exists uniformly and } \tilde{y} \in C^0.$$

$$\begin{aligned} \left| \tilde{y}(x) - y_0 - \int_{x_0}^x f(\xi, \tilde{y}(\xi)) d\xi \right| &= \left| \tilde{y}(x) - y_{n+1}(x) - \int_{x_0}^x \{f(\xi, \tilde{y}(\xi)) - f(\xi, y_n(\xi))\} d\xi \right| \\ &\leq |\tilde{y}(x) - y_{n+1}(x)| + L \int_{x_0}^x |\tilde{y}(\xi) - y_n(\xi)| d\xi \quad \xrightarrow{(n \rightarrow \infty)} 0 \\ \Rightarrow \quad \tilde{y}(x) &= y_0 + \int_{x_0}^x f(\xi, \tilde{y}(x)) d\xi. \end{aligned}$$

Differentiation shows that \tilde{y} solves (1.1).

(b) *Uniqueness* :

Assume $z = z(x)$ is another solution. Then

$$z = y_0 + \int_{x_0}^x f(\xi, z(\xi)) d\xi, \quad y_{n+1} = y_0 + \int_{x_0}^x f(\xi, y_n(\xi)) d\xi$$

and

$$|z - y_{n+1}| \leq \int_{x_0}^x |f(\xi, z) - f(\xi, y_n)| d\xi \leq L \int_{x_0}^x |z - y_n| d\xi.$$

But $|z - y_0| \leq Mh = N/2$, hence

$$\begin{aligned} |z - y_1| &\leq \left(\frac{N}{2}\right) L(x - x_0) \leq \left(\frac{N}{2}\right) Lh \\ |z - y_2| &\leq \left(\frac{N}{2}\right) \frac{L^2(x - x_0)^2}{2!} \leq \left(\frac{N}{2}\right) \frac{L^2 h^2}{2!} \\ &\vdots \\ |z - y_{n+1}| &\leq \left(\frac{N}{2}\right) \frac{L^{n+1} h^{n+1}}{(n+1)!} \quad \xrightarrow{(n \rightarrow \infty)} 0 \\ \Rightarrow \quad z(x) &\equiv \tilde{y}(x) \quad (=: y(x)). \end{aligned}$$

■

1.1.5 Single-Step Methods

Euler :

$$(1.16) \quad \begin{aligned} y_{n+1} &= y_n + hf(x_n, y_n), \quad n = 1, 2, \dots \\ y_0 &= y(x_0) \quad (\text{accuracy of } O(h^2)) \end{aligned}$$

Modified Euler : Better approximation of y' by use of an averaged value of derivatives at the beginning and the end of the interval.

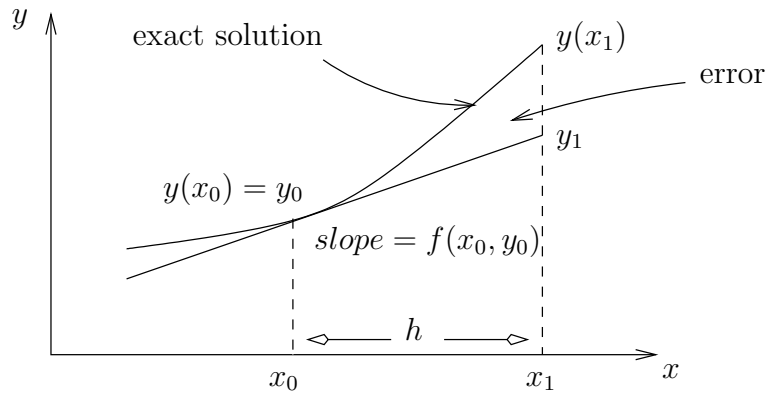


Figure A.1: Euler Method

Use temporary Euler step

$$y_{n+1}^* = y_n + hf(x_n, y_n)$$

to calculate an approximation of the derivative $f(x_{n+1}, y_{n+1}^*)$ at end of interval.

This new derivative is averaged with the initial derivative to obtain a more accurate value for y_{n+1} :

$$y_{n+1} = y_n + \frac{1}{2}h [f(x_n, y_n) + f(x_{n+1}, y_{n+1}^*)].$$

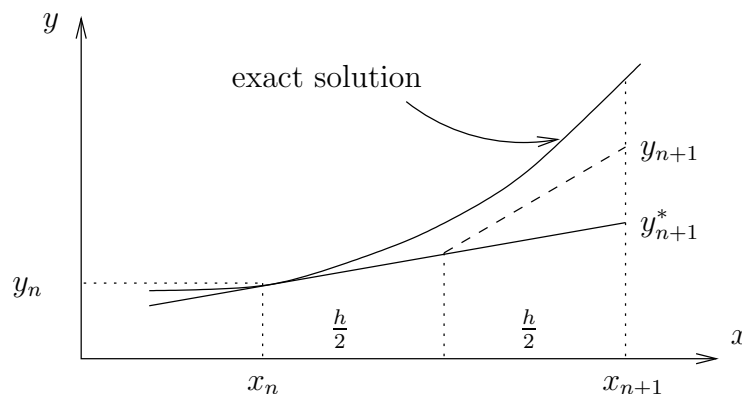


Figure A.2: Modified Euler Method

Improved Euler: $y_{n+1} = y_n + hf(x_n + \frac{h}{2}, y_n + \frac{h}{2}f(x_n, y_n))$.

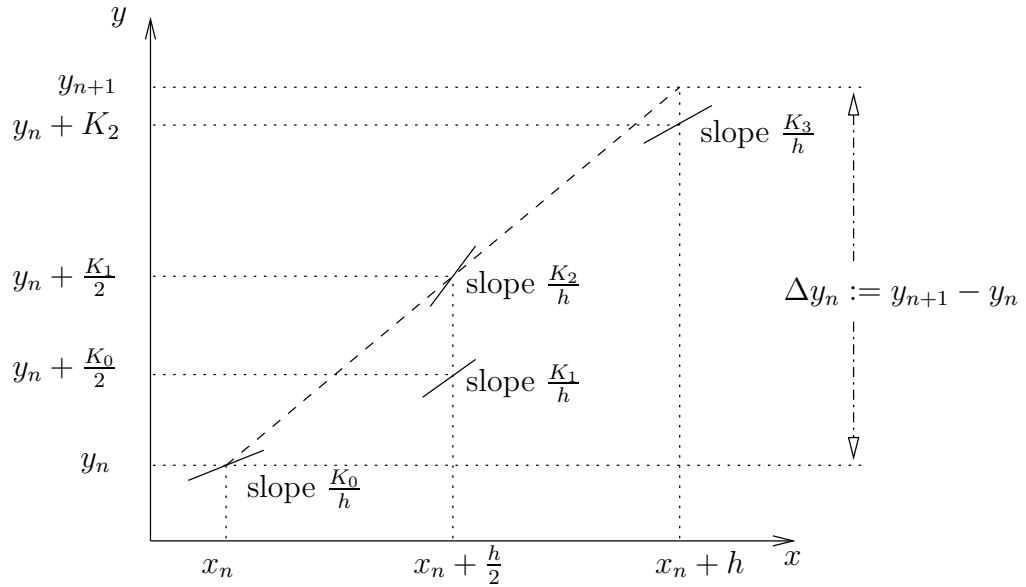
Classical Runge-Kutta (4th order method) : (with Runge's coefficients)

$$(1.17) \quad y_{n+1} = y_n + \frac{(K_0 + 2K_1 + 2K_2 + K_3)}{6}$$

where

$$\begin{aligned} K_0 &= hf(x_n, y_n) \\ K_1 &= hf(x_n + \frac{h}{2}, y_n + \frac{K_0}{2}) \\ K_2 &= hf(x_n + \frac{h}{2}, y_n + \frac{K_1}{2}) \\ K_3 &= hf(x_n + h, y_n + K_2) \end{aligned}$$

i.e. $K_0, K_1, K_2, K_3 = (\text{slopes at different points}) \cdot h$.



Remark 1.15

Figure A.3: Runge-Kutta Method

- (a) Greater accuracy by evaluating $f(x, y)$ at selected points in each subinterval, avoids higher derivatives.
- (b) Taylor algorithm of order $k \iff \text{local error } \epsilon = O(h^{k+1})$.

General Runge-Kutta 2nd order methods :

$$(1.18) \quad y_{n+1} = y_n + ak_1 + bk_2$$

$$\text{where} \quad \begin{aligned} k_1 &:= hf(x_n, y_n) \\ k_2 &:= hf(x_n + \alpha h, y_n + \beta k_1) \end{aligned}$$

and $a, b, \alpha, \beta \in \mathbb{R}$ such that (1.18) agrees with the Taylor expansion of highest possible order.

$$(1.19) \quad \begin{aligned} y(x_{n+1}) &= y(x_n) + hy'(x_n) + \left[\frac{h^2}{2} y''(x_n) \right] + \left[\frac{h^3}{6} y'''(x_n) \right] + \dots \\ &= y(x_n) + hf(x_n, y_n) \\ &\quad + \left[\frac{h^2}{2} (f_x + f_y y') \Big|_{(x_n, y_n)} \right] \\ &\quad + \left[\frac{h^3}{6} (f_{xx} + 2f_{xy} f + f_{yy} f^2 + f_x f_y + f_y^2 f) \Big|_{(x_n, y_n)} \right] \\ &\quad + O(h^4) \end{aligned}$$

with $f_y y' = f_y f$.

$$\begin{aligned} \frac{k_2}{h} &:= f(x_n + \alpha h, y_n + \beta k_1) \\ &= f(x_n, y_n) + \left(\alpha h f_x + \beta k_1 f_y + \frac{\alpha^2 h^2}{2} f_{xx} + \alpha h \beta k_1 f_{xy} + \frac{\beta^2 k_1^2}{2} f_{yy} \right) \Big|_{(x_n, y_n)} + O(h^3). \end{aligned}$$

Substitution into (1.18) yields:

$$y_{n+1} = y_n + (a+b)hf + bh^2(\alpha f_x + \beta f f_y) + \underbrace{bh^3 \left(\frac{\alpha^2}{2} f_{xx} + \alpha\beta f f_{xy} + \frac{\beta^2}{2} f^2 f_{yy} \right)}_{\text{local error } O(h^3)} + O(h^4).$$

Comparison with (1.19) shows (by identifying same powers of h):

$$\begin{aligned} a + b &= 1, \\ b\alpha &= b\beta = \frac{1}{2}. \end{aligned}$$

We have 4 unknowns and 3 equations. So there are many possible Runge-Kutta Methods of 2^{nd} order. We choose $a = b = \frac{1}{2}$, $\alpha = \beta = 1$. This leads to

Runge-Kutta 2^{nd} order :

$$y_{n+1} = y_n + \frac{1}{2}(k_1 + k_2)$$

$$\begin{aligned} \text{where } k_1 &:= hf(x_n, y_n) \\ k_2 &:= hf(x_n + h, y_n + k_1) \end{aligned}$$

This is equal to the **modified Euler**.

Drawback : (a) The local error is difficult to estimate.
(b) The function $f(x, y)$ must be evaluated twice for each step of integration.

Advantage : Higher order method than Euler ($O(h^2)$) \Rightarrow can use larger step size.

1.2 Multistep Methods

Runge-Kutta uses approximation of $y(x_k)$ to obtain approximation of $y(x_{k+1})$ (one-step method). However, if we already have obtained $y(x_k), y(x_{k-1}), y(x_{k-2}), \dots$, why not use them to determine $y(x_{k+1})$?

Notation : A method using n approximation values of $y(x)$ to compute the next value is called an **n-step method**.

n-step method needs values for y_0, y_1, \dots, y_{n-1} to get started. These **starting values** must be computed by one-step methods. This is assumed in the following.

Suppose we already have approximations for y' and y at x_0, x_1, \dots, x_n . Integrate $y' = f(x, y)$ from x_n to x_{n+1} :

$$(1.20) \quad y(x_{n+1}) = y(x_n) + \int_{x_n}^{x_{n+1}} f(x, y(x)) dx.$$

Our first approach leads to the Adams-Bashforth methods.

To integrate $f(x, y(x))$ we approximate $f(x, y(x))$ by a polynomial which interpolates $f(x, y(x))$ at the $(m+1)$ points $x_n, x_{n-1}, x_{n-2}, \dots, x_{n-m}$. Define

$$f_k := f(x_k, y(x_k)) \quad \text{and} \quad s := \frac{x - x_n}{h}.$$

The interpolatory polynomial p_m is given by the

Newton backward formula of degree m :

$$(1.21) \quad p_m(x) = \sum_{k=0}^m (-1)^k \binom{-s}{k} \Delta^k f_{n-k},$$

where the $\Delta^i f_s$ are the **forward differences**

$$(1.22) \quad \Delta^i f_s = \begin{cases} f_s & , \quad i = 0 \\ \Delta(\Delta^{i-1} f_s) = \Delta^{i-1} f_{s+1} - \Delta^{i-1} f_s & , \quad i > 0. \end{cases}$$

With $dx = hds$ we get:

Adams-Bashforth method of degree m :

$$(1.23) \quad \begin{aligned} y_{n+1} &= y_n + h \int_0^1 \sum_{k=0}^m (-1)^k \binom{-s}{k} \Delta^k f_{n-k} ds \\ &= y_n + h \left\{ \gamma_0 f_n + \gamma_1 \Delta f_{n-1} + \dots + \gamma_m \Delta^m f_{n-m} \right\} \end{aligned}$$

with

$$\gamma_k := (-1)^k \int_0^1 \binom{-s}{k} ds$$

$$\Rightarrow \quad \gamma_0 = 1, \quad \gamma_1 = \frac{1}{2}, \quad \gamma_2 = \frac{5}{12}, \quad \gamma_3 = \frac{3}{8}, \quad \gamma_4 = \frac{251}{720}, \quad \dots$$

Remark 1.16

(a) $m = 0 \Rightarrow$ **Euler**

(b) **In general :**

(1.23) requires

$$\begin{aligned} &y' = f \text{ at } m+1 \text{ points : } x_n, x_{n-1}, \dots, x_{n-m} \\ &\text{and } \Delta f_{n-1}, \Delta^2 f_{n-2}, \dots, \Delta^m f_{n-m}. \end{aligned}$$

From this we get $y_{n+1} \cdot x_{n+1} \mapsto x_n$ (relabel), form new differences and repeat the process.

Using the above, we will now compute the **Adams-Bashforth 4-step method:**

Set $\mathbf{m} = \mathbf{3}$:

$$\begin{array}{ccccccc} x_{n-3} & y_{n-3} & f_{n-3} & & & & \\ & & & \Delta f_{n-3} & & & \\ x_{n-2} & y_{n-2} & f_{n-2} & & \Delta^2 f_{n-3} & & \\ & & & \Delta f_{n-2} & & \Delta^3 f_{n-3} & \\ x_{n-1} & y_{n-1} & f_{n-1} & & \Delta^2 f_{n-2} & & \\ & & & \Delta f_{n-1} & & & \\ x_n & y_n & f_n & & & & \end{array}$$

Then there follows with (1.23):

$$(1.24) \quad y_{n+1} = y_n + h \left(f_n + \frac{1}{2} \Delta f_{n-1} + \frac{5}{12} \Delta^2 f_{n-2} + \frac{3}{8} \Delta^3 f_{n-3} \right)$$

with

$$\begin{aligned}\Delta f_{n-1} &:= f_n - f_{n-1} \\ \Delta^2 f_{n-2} &:= f_n - 2f_{n-1} + f_{n-2} \\ \Delta^3 f_{n-3} &:= f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3} \\ \Rightarrow y_{n+1} &= y_n + \frac{h}{24} (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) .\end{aligned}$$

Newton backward formula ($m = 3$) :

The *local error* is

$$h^4 f^{(4)}(\eta) \binom{-s}{4} , \quad \eta \in (x_{n-3}, x_n),$$

and the error of (1.24) is

$$E := h \int_0^1 h^4 f^{(4)}(\eta) \binom{-s}{4} ds .$$

Because $\binom{-s}{4}$ does not change its sign in $[0, 1]$, it follows from the second mean value theorem of calculus that there exists $\xi \in (x_{n-3}, x_{n+1})$ such that

$$E_{AB} = h^5 f^{(4)}(\xi) \int_0^1 \binom{-s}{4} ds = h^5 y^{(4)}(\xi) \frac{251}{720} .$$

Derivation of multistep methods by using numerical integration :

Instead of integrating $f(x, y)$ in (1.20) from x_n to x_{n+1} , we can integrate from x_{n-p} to x_{n+1} for some integer $p \geq 0$. Interpolation at $m + 1$ points $x_n, x_{n-1}, \dots, x_{n-m}$ with Newton's backward formula yields :

$$(1.25) \quad y_{n+1} = y_{n-p} + h \int_{-p}^1 \sum_{k=0}^m (-1)^k \binom{-s}{k} \Delta^k f_{n-k} ds .$$

Example 1.17

$p = 0$: Adams-Bashforth (1.23)

$m = 1, p = 1$:

$$(1.26) \quad y_{n+1} = y_{n-1} + 2hf_n , \quad E = \frac{h^3}{3} y'''(\xi)$$

$m = 3, p = 3$:

$$(1.27) \quad y_{n+1} = y_{n-3} + \frac{4h}{3} (2f_n - f_{n-1} + 2f_{n-2}) , \quad E = \frac{14}{45} h^5 y^{(4)}(\xi) .$$

Remark 1.18

- (a) (1.26) is comparable in simplicity with Euler but has a smaller discretization error.
- (b) (1.27) requires knowledge of $f(x, y)$ at only 3 points; discretization error like Adams-Bashforth.

(c) Let m be odd and $m = p$. Then all methods (1.25) have zero coefficients of m -th difference.

Disadvantages and advantages of multistep formulas

Disadvantages (1) Not self-starting. Adams-Bashforth needs successive values of $f(x, y)$ at equally spaced points before the formula can be used.
 (2) The constants $C \in \mathbb{R}$ in the error estimates are larger than those of the corresponding Runge-Kutta formulas, so Runge-Kutta is more accurate.

Advantage : Only one derivative evaluation per step, in contrast to four when using Runge-Kutta

\Rightarrow multistep methods are faster and easier to implement.

Our next approach to multistep methods makes use of **quadrature formulas**.

Derivation of multistep methods for

$$\begin{aligned} y'(x) &= f(x, y(x)) \quad , \quad a \leq x \leq b \\ y(a) &= \eta \end{aligned}$$

$$(1.28) \quad y(x+t) = y(x) + \int_x^{x+t} f(s, y(s)) ds \quad , \quad a \leq x < x+t \leq b.$$

Replace the integral by a quadrature formula:

(1) **Rectangle rule** ($t = h$):

$$\int_x^{x+h} f(s, y(s)) ds = hf(x, y(x)) + \frac{h^2}{2} f'(\xi, y(\xi)) \quad , \quad \xi \in (x, x+h),$$

and with (1.28) it follows that

$$y(x+h) = y(x) + hf(x, y(x)) + \frac{1}{2}h^2 f'(\xi, y(\xi)).$$

For $x = x_k$, $x+h = x_{k+1}$ we get the formula of **Euler** :

$$(1.29) \quad \boxed{y_{k+1} = y_k + hf(x_k, y_k)}$$

(2) **Trapezoidal rule** ($t = h$):

$$\int_x^{x+h} f(s) ds = \frac{h}{2}(f(x) + f(x+h)) - \frac{f''(\xi)}{12}h^3.$$

This results in the **implicit trapezoidal method**:

$$(1.30) \quad \boxed{y_{n+1} = y_n + \frac{h}{2} [f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \quad , \quad n = 0, 1, \dots}$$

where y_{n+1} appears implicitly. Solving this using fixed-point iteration over y (x_n fixed) after finding starting value for y by using explicit Euler. This brings us to the **predictor-corrector methods**:

$$(1.31) \quad \left\| \begin{array}{l} \text{predictor : (open type, first approximation of } y_{n+1} \text{)} \\ \\ y_{n+1}^{(0)} = y_n + hf(x_n, y_n) \\ \\ \text{corrector : (closed type)} \\ \\ y_{n+1}^{(1)} = y_n + \frac{h}{2} [f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(0)})] \\ \\ \vdots \\ \\ y_{n+1}^{(k)} = y_n + \frac{h}{2} [f(x_n, y_n) + f(x_{n+1}, y_{n+1}^{(k-1)})] \quad , \quad k = 1, 2, \dots \end{array} \right.$$

Iteration terminated when

$$\frac{|y_{n+1}^{(k)} - y_{n+1}^{(k-1)}|}{|y_{n+1}^{(k)}|} < \varepsilon .$$

(3) **Simpson's rule** :

$$\int_{x_0}^{x_2} f(x) dx = \frac{h}{3} (f(x_0) + 4f(x_1) + f(x_2)) - \left[\frac{f^{(4)}(\xi)}{90} \right] h^5$$

with $x_k = x_0 + kh$, $k = 0, 1, \dots$.

Take for these points: $x_0 = x$, $x_1 = x + h$, $x_2 = x + 2h$, $x_2 - x_0 = 2h$. Then there follows for the integral

$$\begin{aligned} & \int_x^{x+2h} f(s, y(s)) ds \\ &= \frac{h}{3} [f(x, y(x)) + 4f(x+h, y(x+h)) + f(x+2h, y(x+2h))] - \frac{1}{90} f^{(4)}(\xi, y(\xi)) h^5, \end{aligned}$$

and with (1.28) we get

$$y(x+2h) = y(x) + \frac{1}{3} h [f(x, y(x)) + 4f(x+h, y(x+h)) + f(x+2h, y(x+2h))] + E,$$

where E is the truncation error.

For $x = x_k$ we get **Milne's method** :

$$(1.32) \quad \boxed{y_{k+2} = y_k + \frac{h}{3} [f(x_k, y_k) + 4f(x_{k+1}, y_{k+1}) + f(x_{k+2}, y_{k+2})]}$$

where y_{k+2} is implicit.

Remark 1.19

Predictor-corrector methods based on Simpson's rule, like Milne's method, are numerically unstable.

Corrector formulas of higher order: Adams-Moulton methods

Adams-Moulton methods are closely related to the Adams-Bashforth methods, only now we also consider the point x_{n+1} when interpolating f . Interpolating $f(x, y)$ at x_{n+1}, \dots, x_{n-m} with the Newton backward formula yields

$$p_{m+1}(s) = \sum_{k=0}^{m+1} (-1)^k \binom{1-s}{k} \Delta^k f_{n+1-k} \quad , \quad s = \frac{x - x_n}{h}$$

$$(1.20) \Rightarrow \int_{x_n}^{x_{n+1}} p_{m+1}(s) dx = y_{n+1} - y_n .$$

With this we get

$$y_{n+1} = y_n + h \left(\gamma'_0 f_{n+1} + \gamma'_1 \Delta f_n + \dots + \gamma'_{m+1} \Delta^{m+1} f_{n-m} \right) ,$$

and for the error we obtain

$$E = \gamma'_{m+2} h^{m+3} y^{(m+3)}(\xi) ,$$

where

$$\gamma'_k = (-1)^k \int_0^1 \binom{1-s}{k} ds \quad (k = 0, 1, \dots, m+1)$$

$$\Rightarrow \quad \gamma'_0 = 1, \quad \gamma'_1 = -\frac{1}{2}, \quad \gamma'_2 = -\frac{1}{12}, \quad \gamma'_3 = -\frac{1}{24}, \quad \gamma'_4 = -\frac{10}{720}, \quad \dots$$

For $m = 2$ we get the **Adams-Moulton-formula** of 4th order:

$$(1.33) \quad y_{n+1} = y_n + \frac{h}{24} (9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}) \quad , \quad E_{AM} = -\frac{19}{720} h^5 y^{(5)}(\xi) .$$

Remark 1.20

- (a) The corrector formula (1.33) is of closed type since $f_{n+1} = f(x_{n+1}, y_{n+1})$ where y_{n+1} is unknown. Solution by iteration.
- (b) Convenient predictor to use for corrector (1.33) : Adams-Bashforth of 4th order.

Algorithm :

For $y' = f(x, y)$, h fixed, $x_n := x_0 + nh$, with (y_0, f_0) , (y_1, f_1) , (y_2, f_2) , (y_3, f_3) given, $n = 3, 4, \dots$

(1) Compute $y_{n+1}^{(0)}$ by

$$y_{n+1}^{(0)} = y_n + \frac{h}{24} (55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) .$$

(2) Compute $f_{n+1}^{(1)} = f(x_{n+1}, y_{n+1}^{(0)})$.

(3) Compute $y_{n+1}^{(k)}$, $k = 1, 2, \dots$ by

$$y_{n+1}^{(k)} = y_n + \frac{h}{24} \left(9f(x_{n+1}, y_{n+1}^{(k-1)}) + 19f_n - 5f_{n-1} + f_{n-2} \right) .$$

(4) Iterate on k until

$$\frac{|y_{n+1}^{(k)} - y_{n+1}^{(k-1)}|}{|y_{n+1}^{(k)}|} < \varepsilon$$

for ε prescribed.

For the error we have with $\xi_1 \neq \xi_2$:

$$\begin{aligned} E_{AB} &= y(x_{n+1}) - y_{n+1}^{(0)} = \frac{251}{720} h^5 y^{(5)}(\xi_1), \\ (1.34) \quad E_{AM} &= y(x_{n+1}) - y_{n+1}^{(1)} = -\frac{19}{720} h^5 y^{(5)}(\xi_2). \end{aligned}$$

Suppose $y^{(5)} \approx \text{const.}$, then there follows

$$\begin{aligned} h^5 y^{(5)} &= \frac{720}{270} (y_{n+1}^{(1)} - y_{n+1}^{(0)}) \\ (1.34) \quad y(x_{n+1}) - y_{n+1}^{(1)} &\approx -\frac{1}{14} (y_{n+1}^{(1)} - y_{n+1}^{(0)}) =: D_{n+1}. \end{aligned}$$

Choose E_1, E_2 : $E_1 \leq \frac{|D_{n+1}|}{h} \leq E_2$ and go on with algorithm with same value of h .

$\frac{|D_{n+1}|}{h} > E_2$: Reduce stepsize to $\frac{h}{2}$ and compute starting values again \Rightarrow algorithm.

$\frac{|D_{n+1}|}{h} < E_1$: Too accurate, save computing time by enlarging the stepsize to $2h$ and compute starting values again \Rightarrow algorithm.

Example 1.21

$$y' = x + y, \quad y(0) = 0, \quad x \in [0, 1], \quad h = \frac{1}{32}, \quad y_{\text{exact}} = e^x - 1 - x$$

with

$$y\left(\frac{1}{32}\right) = 0.49340725 \cdot 10^{-3}, \quad y\left(\frac{2}{32}\right) = 0.19944459 \cdot 10^{-2}, \quad y\left(\frac{3}{32}\right) = 0.45351386 \cdot 10^{-2}$$

and

$$\begin{aligned} f_0 &= x_0 + y_0 = 0, \\ f_1 &= x_1 + y_1 = 0.3125 \cdot 10^{-1} + 0.49340725 \cdot 10^{-3} = 0.317434 \cdot 10^{-1}, \\ f_2 &= x_2 + y_2 = 0.625 \cdot 10^{-1} + 0.199445 \cdot 10^{-2} = 0.644945 \cdot 10^{-1}, \\ f_3 &= x_3 + y_3 = 0.9375 \cdot 10^{-1} + 0.45351 \cdot 10^{-2} = 0.98285 \cdot 10^{-1}. \end{aligned}$$

(I) **Adams-Bashforth** :

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \frac{\mathbf{h}}{24} (55\mathbf{f}_n - 59\mathbf{f}_{n-1} + 37\mathbf{f}_{n-2} - 9\mathbf{f}_{n-3}),$$

and so we get for y_4 :

$$y_4 = y_3 + \frac{h}{24} (55f_3 - 59f_2 + 37f_1 - 9f_0) \approx 0.81484 \cdot 10^{-2} = y_4^{(0)}.$$

(II) **Adams-Moulton** :

$$\mathbf{y}_{n+1}^{(k)} = \mathbf{y}_n + \frac{\mathbf{h}}{24} \left(9\mathbf{f}(\mathbf{x}_{n+1}, \mathbf{y}_{n+1}^{(k-1)}) + 19\mathbf{f}_n - 5\mathbf{f}_{n-1} + \mathbf{f}_{n-2} \right) \quad , \quad \mathbf{k} = 1, 2, \dots$$

$$\begin{aligned} y_4^{(1)} &= y_3 + \frac{h}{24} \left(9f(x_4, y_4^{(0)}) + 19f_3 - 5f_2 + f_1 \right) \\ &= y_3 + \frac{h}{24} \left(9(x_4 + y_4^{(0)}) + 1.867415 - 0.3224725 + 0.0317434 \right) \\ &\approx 0.81484 \cdot 10^{-2} = y_4^{(0)}. \end{aligned}$$

Note: Numerically stable !

(III) **Milne's predictor-corrector** (4th order) :

$$\begin{aligned} \mathbf{y}_{n+1}^{(0)} &= \mathbf{y}_{n-3} + \frac{4\mathbf{h}}{3} (2\mathbf{f}_n - \mathbf{f}_{n-1} + 2\mathbf{f}_{n-2}) \quad , \quad \mathbf{E}_M^0 = \frac{28}{90} \mathbf{h}^5 \mathbf{y}^{(5)}(\xi_1) \\ \mathbf{y}_{n+1}^{(1)} &= \mathbf{y}_{n-1} + \frac{\mathbf{h}}{3} (\mathbf{f}_{n+1}^{(0)} + 4\mathbf{f}_n + \mathbf{f}_{n-1}) \quad , \quad \mathbf{E}_M^1 = -\frac{1}{90} \mathbf{h}^5 \mathbf{y}^{(5)}(\xi_2). \end{aligned}$$

Note: Numerically instable: error introduced at one stage grows exponentially.

Remark 1.22

Runge-Kutta : Self-starting, stable, good accuracy, no estimate on local error: no knowledge of optimal h .

Predictor-corrector : Automatic error estimate at each step \Rightarrow selection of optimal h ; fast, numerically instable, more difficult to implement.

1.3 Convergence and consistency of single-step methods

Let us consider the **IVP** in the domain $G \subseteq \mathbb{R}^{n+1}$:

$$(1.35) \quad \mathbf{y}' = \mathbf{f}(x, \mathbf{y}) \quad \text{with} \quad \mathbf{y}(x_0) = \mathbf{y}_0, \quad (x_0, \mathbf{y}_0) \in G,$$

and let us assume that \mathbf{f} is Lipschitz-continuous, so that for every starting value there exists exactly one solution in G .

Speaking generally, let us define a

numerical method for approximating the solution \mathbf{y} of the IVP (1.35) as a method which, in the interval $[a, b]$,

- a) defines a grid Δ with: $a = x_0 < x_1 < x_2 < \dots < x_m = b$,
- b) computes a grid function $\mathbf{y}_n : \Delta \rightarrow \mathbb{R}^n$ with $(x, \mathbf{y}_n(x)) \in G$, $x \in \Delta$.

Definition 1.23

A numerical method is called **convergent** for the IVP (1.35) on $[a, b]$ if for the global error

$$\varepsilon_h := y(x) - y_h(x) \quad \text{for } x \in \Delta$$

there holds

$$\|\varepsilon_h\| := \max_{x_i} |y_h(x_i) - y(x_i)| \rightarrow 0 \quad \text{for } h \rightarrow 0 \quad (\text{maximum step size}).$$

The method has the **order of convergence** $p > 0$ if

$$\|\varepsilon_h\| = O(h^p).$$

Example 1.24

We consider the IVP

$$(1.36) \quad y' = \lambda y \quad \text{with } y(x_0) = y_0 \in \mathbb{R}, \lambda \in \mathbb{R}.$$

Separation of variables:

$$\int \frac{dy}{y} = \lambda \int dx \quad \Rightarrow \quad (\ln y = \lambda x + \tilde{c} \Rightarrow y = c e^{\lambda x}).$$

Initial value (IV):

$$y(x_0) = y_0 = c e^{\lambda x_0} \quad \Rightarrow \quad c = y_0 e^{-\lambda x_0}.$$

General solution:

$$y = y_0 e^{\lambda(x-x_0)}.$$

Euler's method for $h > 0$:

$$\begin{aligned} y_0 &= y(x_0) \\ y_j &= y_{j-1} + \lambda h y_{j-1} = (1 + h\lambda)y_{j-1} = (1 + h\lambda)^2 y_{j-2} = \dots \\ &= (1 + h\lambda)^j y_0 \end{aligned}$$

Trick:

$$\begin{aligned} e^{h\lambda} &= 1 + h\lambda + \frac{(h\lambda)^2}{2!} + \frac{(h\lambda)^3}{3!} + \dots \\ &= 1 + h\lambda + O(h^2) \quad \Rightarrow \quad 1 + h\lambda = e^{h\lambda} + O(h^2). \end{aligned}$$

Then there follows

$$y_j = (1 + h\lambda)^j y_0 = (e^{h\lambda} + O(h^2))^j y_0 = (e^{jh\lambda} + O(jh^2)) y_0,$$

and thus

$$\begin{aligned}
 \|\varepsilon_h\| &:= \max_{0 \leq j \leq m} |y_j - y(x_0 + jh)| \\
 &= \max_{0 \leq j \leq m} \left| (e^{jh\lambda} + O(jh^2))y_0 - e^{jh\lambda}y_0 \right| \\
 &= \max_{0 \leq j \leq m} \left| O(jh^2)y_0 \right| \\
 &= O(h), \quad \text{da } 0 \leq jh \leq b - a.
 \end{aligned}$$

So Euler's method is convergent for the IVP (1.36) with convergence order of 1.

Order of convergence of the modified Euler's method

for the same IVP (1.36) :

$$y' = \lambda y, \quad y(x_0) = y_0.$$

One obtains

$$\begin{aligned}
 y_j &= y_{j-1} + h\lambda \left(y_{j-1} + \frac{h}{2}\lambda y_{j-1} \right) \\
 &= \left(1 + h\lambda + \frac{(h\lambda)^2}{2} \right) y_{j-1} \\
 &= \left(1 + h\lambda + \frac{(h\lambda)^2}{2} \right)^j y_0, \quad j = 0, 1, \dots, m.
 \end{aligned}$$

the same trick as before yields:

$$\begin{aligned}
 1 + h\lambda + \frac{(h\lambda)^2}{2} &= e^{h\lambda} + O(h^3) \\
 \Rightarrow y_j &= \left[e^{h\lambda} + O(h^3) \right]^j y_0 = \left[e^{jh\lambda} + O(jh^3) \right] y_0 \\
 \Rightarrow \|\varepsilon_h\| &= \max_{0 \leq j \leq m} \left| \left(e^{jh\lambda} + O(jh^3) \right) y_0 - e^{jh\lambda} y_0 \right| = O(h^2).
 \end{aligned}$$

And so the modified Euler's method has the order of convergence 2.

In order to make general statements about the convergence of one-step methods, we will now define what a single-step method is:

Definition 1.25

Given the IPV

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

and a grid Δ :

$$a = x_0 < x_1 < \dots < x_m = b, \quad h_j := x_{j+1} - x_j.$$

Single-step method :

$$\begin{aligned}
 \mathbf{y}_0 &:= \mathbf{y}(x_0) \\
 x_{j+1} &:= x_j + h_j \\
 \mathbf{y}_{j+1} &:= \mathbf{y}_j + h_j \Phi(x_j, \mathbf{y}_j; h_j), \quad j = 0, 1, 2, \dots, m-1.
 \end{aligned}$$

$\Phi(\cdot, \cdot; h) : G \rightarrow \mathbb{R}^n$ is called the **method function** (or **slope estimator**).

Example 1.26

$$\begin{aligned} \text{Euler's method} & : & \Phi(x, y; h) & := f(x, y) \\ \text{Improved Euler's method} & : & \Phi(x, y; h) & := f\left(x + \frac{h}{2}, y + \frac{h}{2}f(x, y)\right) \end{aligned}$$

Definition 1.27

Let y be the solution of the IVP in the interval I . Let x_i and $x_i + h$ be two points in I . Then

$$d(x_i, y(x_i); h) := \frac{1}{h} \underbrace{\left(y(x_i + h) - [y(x_i) + h\Phi(x_i, y(x_i); h)] \right)}_{(*)},$$

(*) is called the **local discretization error of the single-step method at $(x_i, y(x_i))$** . Here

$y(x_i + h)$ is the exact value of the solution y at x_{i+1} .

$[y(x_i) + h\Phi(x_i, y(x_i); h)]$ is the approximation at x_{i+1} , where $y(x_i)$ is the exact value at x_i .

Thus (*) describes the error after one step, if there was no error at x_i (\rightarrow local). The factor $\frac{1}{h}$ is needed so that the order of consistency equals the order of convergence.

Question : What is the connection to convergence?

Definition 1.28 (consistency)

A single-step method is called **consistent** with the IVP if for all nodes x_i , $i = 0, \dots, m$ there uniformly holds :

$$|d(x_i, y(x_i); h)| \rightarrow 0 \quad \text{for } h \rightarrow 0.$$

The single-step method has the **order of consistency** $p > 0$ if there exists a constant $C > 0$ with

$$|d(x_i, y(x_i); h)| \leq C h^p \quad \text{for } h \rightarrow 0$$

uniformly for all nodes x_i .

Example 1.29

(a) **Euler's method** :

Typical trick when proving consistency: Taylor expansion:

$$\begin{aligned} d(x_i, y(x_i); h) & = \frac{1}{h} \left(y(x_i + h) - [y(x_i) + hf(x_i, y(x_i))] \right) \\ & = \frac{1}{h} \left(y(x_i + h) - y(x_i) \right) - f(x_i, y(x_i)) \quad , \quad f(x_i, y(x_i)) = y'(x_i) \\ & \stackrel{\text{Taylor}}{=} \frac{1}{h} \left(\left[y(x_i) + hy'(x_i) + \frac{h^2}{2}y''(x_i) + \dots \right] - y(x_i) \right) - y'(x_i) \\ & = \frac{1}{h} \left(hy'(x_i) + \frac{h^2}{2}y''(x_i) + O(h^3) \right) - y'(x_i) \\ & = \frac{h}{2}y''(x_i) + O(h^2) \\ & = O(h) \quad (\text{holds uniformly on } I \text{ if } f \text{ is sufficiently differentiable}). \end{aligned}$$

Thus Euler's method has the order of consistency 1.

(b) **Improved Euler** :

$$\begin{aligned}\Phi(x, y; h) &= f\left(x + \frac{h}{2}, y + \frac{h}{2}f(x, y)\right) \\ &= f(x, y) + \frac{h}{2}f_x + \frac{h}{2}ff_y + \frac{h^2}{2}\left(\frac{1}{4}f_{xx} + \frac{1}{2}ff_{xy} + \frac{1}{4}f^2f_{yy}\right) + O(h^3)\end{aligned}$$

With Taylor expansion:

$$\begin{aligned}y(x+h) &= y(x) + hf + \frac{h^2}{2}Df + \frac{h^3}{6}D^2f + O(h^4) \quad , \quad D := \frac{d}{dx} \\ &= y(x) + hf + \frac{h^2}{2}(f_x + ff_y) + \frac{h^3}{6}(f_{xx} + 2ff_{xy} + f^2f_{yy} + (f_x + ff_y)f_y) + O(h^4)\end{aligned}$$

there follows

$$\begin{aligned}d(x, y(x); h) &= \frac{1}{h}(y(x+h) - y(x)) - \Phi(x, y(x); h) \\ &= h^2\left[\left(\frac{1}{6} - \frac{1}{8}\right)(f_{xx} + 2ff_{xy} + f^2f_{yy}) + \frac{1}{6}(f_x + ff_y)f_y\right] + O(h^3) \\ &= O(h^2),\end{aligned}$$

therefore order of consistency 2.

The connection between consistency and convergence is given in the following theorem.

Theorem 1.30 (Convergence Theorem)

Given a single-step method for an IVP.

- (i) If the method function $\Phi(x, y; h)$ is Lipschitz continuous with respect to y and
- (ii) if the single-step method is consistent with the IVP

then the method converges and the order of convergence equals the order of consistency. (Proof uses a discrete Gronwall Lemma.)

1.4 Numerical stability

An important question when dealing with numerical methods is the propagation of errors through rounding effects or inexact starting values. A stable method decreases the influence of a rounding error in each following step; an instable method on the other hand may let the effect of such an error become larger and larger with each step, making the computed approximations more and more useless.

Example 1.31

$$y' = -100y + 100 \quad , \quad y(0.05) = e^{-5} + 1 \approx 1.00673 + \varepsilon$$

(a) **exact solution:** $y = e^{-100x} + 1$

(b) explicit Euler method:

$$y_{i+1} = y_i + hf(x_i, y_i)$$

h = 0.05 :			h = 0.02 :		
<i>x</i>	<i>u(x)</i>	<i>y(x) - u(x)</i>	<i>x</i>	<i>u(x)</i>	<i>y(x) - u(x)</i>
0.05	1.006738		0.05	1.006738	
0.10	$9.730482 \star 10^{-1}$	$2.699719 \star 10^{-2}$	0.07	$9.932621 \star 10^{-1}$	$7.649829 \star 10^{-3}$
0.15	1.107807	$-1.078068 \star 10^{-1}$	0.09	1.006738	$-6.614537 \star 10^{-3}$
0.20	$5.687714 \star 10^{-1}$	$4.312286 \star 10^{-1}$	0.11	$9.932621 \star 10^{-1}$	$6.754649 \star 10^{-3}$
0.25	2.724914	-1.724914	0.13	1.006738	$-6.735687 \star 10^{-3}$
0.30	-5.899658	6.899658	0.15	$9.932621 \star 10^{-1}$	$6.738253 \star 10^{-3}$
0.35	$2.859863 \star 10^1$	$-2.759863 \star 10^1$	0.17	1.006738	$-6.737906 \star 10^{-3}$
0.40	$-1.093945 \star 10^2$	$1.103945 \star 10^2$	0.19	$9.932621 \star 10^{-1}$	$6.737953 \star 10^{-3}$
0.45	$4.425781 \star 10^2$	$-4.415781 \star 10^2$	0.21	1.006738	$-6.737946 \star 10^{-3}$
0.50	$-1.765312 \star 10^3$	$1.766312 \star 10^3$	0.23	$9.932621 \star 10^{-1}$	$6.737947 \star 10^{-3}$
0.55	$7.066250 \star 10^3$	$-7.065250 \star 10^3$	0.25	1.006738	$-6.737947 \star 10^{-3}$

(c) implicit Euler method:

$$y_{i+1} = y_i + hf(x_{i+1}, y_{i+1}) \Rightarrow y_{i+1} = \frac{y_i + 100h}{1 + 100h}$$

h = 0.05 :			h = 0.02 :		
<i>x</i>	<i>u(x)</i>	<i>y(x) - u(x)</i>	<i>x</i>	<i>u(x)</i>	<i>y(x) - u(x)</i>
0.05	1.006738		0.05	1.006738	
0.10	1.001123	$-1.077591 \star 10^{-3}$	0.07	1.002246	$-1.334100 \star 10^{-3}$
0.15	1.000187	$-1.868593 \star 10^{-4}$	0.09	1.000749	$-6.252510 \star 10^{-4}$
0.20	1.000031	$-3.119214 \star 10^{-5}$	0.11	1.000250	$-2.328519 \star 10^{-4}$
0.25	1.000005	$-5.199020 \star 10^{-6}$	0.13	1.000083	$-8.092420 \star 10^{-5}$
0.30	1.000001	$-8.665064 \star 10^{-7}$	0.15	1.000028	$-2.742227 \star 10^{-5}$
0.35	1.000000	$-1.444187 \star 10^{-7}$	0.17	1.000009	$-9.201326 \star 10^{-6}$
0.40	1.000000	$-2.407069 \star 10^{-8}$	0.19	1.000003	$-3.075305 \star 10^{-6}$
0.45	1.000000	$-4.012691 \star 10^{-9}$	0.21	1.000001	$-1.026210 \star 10^{-6}$
0.50	1.000000	$-6.693881 \star 10^{-10}$	0.23	1.000000	$-3.422203 \star 10^{-7}$
0.55	1.000000	$-1.109584 \star 10^{-10}$	0.25	1.000000	$-1.140931 \star 10^{-7}$

Analysis of error propogation by comparison with the test problem

(1.37) $y' = \lambda y + y_0 \quad , \quad \lambda \in \mathbb{C}$

since

$$y' = f(x, y) \stackrel{\text{Taylor}}{=} f(x, 0) + \underbrace{\frac{df}{dy}(x, 0)}_{:= \lambda} y + \dots$$

1.4.1 Stability of single-step methods

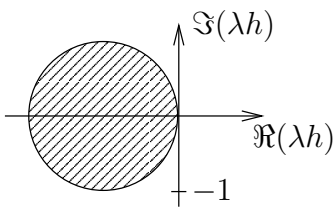
Example 1.32

(1) **Euler's method:**

$$y_{i+1} = y_i + hf(x_i, y_i) = y_i + h\lambda y_i = (1 + \lambda h)y_i$$

Let $\tilde{y}_0 = y_0 + \varepsilon_0$ where ε_0 is a small error in the initial values. It follows that

$$\begin{aligned} \tilde{y}_{i+1} &= y_{i+1} + \varepsilon_{i+1} = \tilde{y}_i + hf(x_i, \tilde{y}_i) \\ \Rightarrow (1 + \lambda h)\tilde{y}_i &= (1 + \lambda h)y_i + (1 + \lambda h)\varepsilon_i \\ \Rightarrow \varepsilon_{i+1} &= (1 + \lambda h)\varepsilon_i = (1 + \lambda h)^2\varepsilon_{i-1} = \dots = (1 + \lambda h)^{i+1}\varepsilon_0. \end{aligned}$$



The method is stable if the influence of ε_0 decreases, i.e. if

$$|1 + \lambda h| < 1,$$

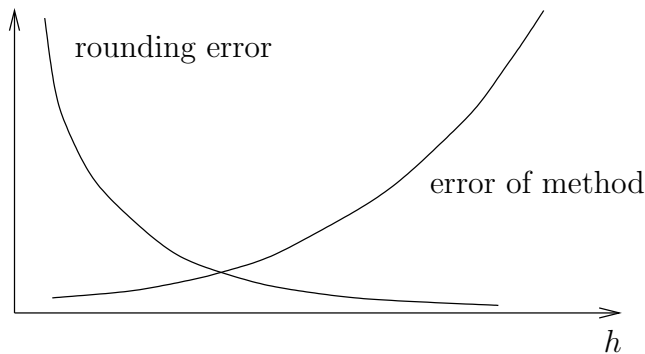
which is a circle in the (λh) -plane with center $(-1, 0)$ and radius 1.

\Rightarrow stability domain of Euler's method.

In Example 1.31 we have $\lambda = -100$, and therefore

$$\begin{aligned} |1 - 100h| < 1 &\Leftrightarrow -1 < 1 - 100h < 1 \\ &\Leftrightarrow \begin{cases} 100h > 0 & , h > 0 \\ 100h < 2 & , h < 0.02 \end{cases} \end{aligned}$$

But for the rounding error there holds



(2) **Implicit Euler method:**

$$y_{i+1} = y_i + hf_{i+1} = y_i + h\lambda y_{i+1} \Rightarrow y_{i+1} = \frac{1}{1 - h\lambda}y_i$$

As in (1):

$$\left| \frac{1}{1 - h\lambda} \right| < 1 \Leftrightarrow |1 - h\lambda| > 1 \Leftrightarrow |h\lambda - 1| > 1$$

which is a circle with center $(1, 0)$ and radius 1.

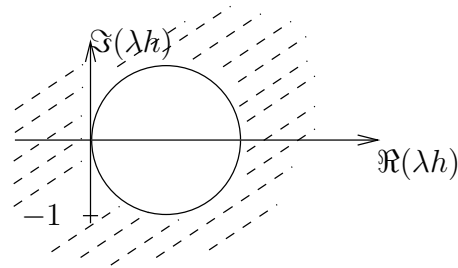


Figure A.4: Domain of stability of the implicit Euler method

In Example 1.31 we have $\lambda = -100$ and therefore

$$\begin{aligned} |100h - 1| > 1 &\Leftrightarrow -1 > 100h - 1 > 1 \\ &\Leftrightarrow \begin{cases} 100h - 1 < -1 &\Rightarrow h < 0 \quad (\text{contradiction}) \\ 100h - 1 > 1 &\Rightarrow h > 0 \end{cases} \end{aligned}$$

(3) **implicit trapezoidal method:**

$$\begin{aligned} y_{i+1} &= y_i + \frac{h}{2} [f_{i+1} + f_i] = y_i + \frac{h}{2} [\lambda y_{i+1} + \lambda y_i] \\ \Rightarrow &\quad \left[1 - \frac{\lambda h}{2}\right] y_{i+1} = \left[1 + \frac{\lambda h}{2}\right] y_i \\ \Leftrightarrow &\quad y_{i+1} = \frac{2 + \lambda h}{2 - \lambda h} y_i \end{aligned}$$

So the condition for stability is

$$\begin{aligned} \left| \frac{2 + \lambda h}{2 - \lambda h} \right| < 1 &\Leftrightarrow \sqrt{\frac{(2 + \lambda h)(\overline{2 + \lambda h})}{(2 - \lambda h)(\overline{2 - \lambda h})}} < 1 \\ &\quad , \text{ mit } \lambda = \lambda_1 + i\lambda_2 \Rightarrow \lambda + \bar{\lambda} = 2\lambda_1 \in \mathbb{R} \\ &\Leftrightarrow \sqrt{\frac{(2 + \lambda h)(2 + \bar{\lambda} h)}{(2 - \lambda h)(2 - \bar{\lambda} h)}} < 1 \\ &\Leftrightarrow \sqrt{\frac{4 + 2\lambda h + 2\bar{\lambda} h + |\lambda|^2 h^2}{4 - 2\lambda h - 2\bar{\lambda} h + |\lambda|^2 h^2}} < 1 \\ &\Leftrightarrow \sqrt{\frac{4 + 4h\lambda_1 + |\lambda|^2 h^2}{4 - 4h\lambda_1 + |\lambda|^2 h^2}} < 1. \end{aligned}$$

This is satisfied if the denominator is larger than the numerator, i.e. if $\Re(\lambda) = \lambda_1 < 0$
 \Rightarrow the domain of stability is the left half-plane.

This means that the trapezoidal method for (1.31) is stable for arbitrary step length h if $\lambda \in \mathbb{R}_{<0}$.

1.4.2 Stability of multistep methods

Usually for fixed x and smaller h the error becomes larger!

Milne: $y_{n+1} = y_{n-1} + 2hf_n$, discretization error $E = \frac{h^3}{3} y'''(\zeta)$.

Euler: $y_{n+1} = y_n + hf_n$, $E = \mathcal{O}(h^2)$

Example 1.33

$y' = -2y + 1$, $y(0) = 1$ (with exact solution $y_{\text{exact}} = \frac{1}{2}e^{-2x} + \frac{1}{2}$).

$f_n := -2y_n + 1$.

Milne: $y_{n+1} + 4hy_n - y_{n-1} = 2h$, $y_0 = 1$.

The solution of this equation has the form $y_n = C_1\beta_1^n + C_2\beta_2^n + \frac{1}{2}$, where $\beta_{1,2} = -2h \pm \sqrt{1 + 4h^2}$ solves $\beta^2 + 4h\beta - 1 = 0$.

Expanding β_1 and β_2 into a Taylor series it follows $\beta_1 = 1 - 2h + \mathcal{O}(h^2)$, $\beta_2 = -(1 + 2h) + \mathcal{O}(h^2)$.

So the solution can be written as

$$y_n = C_1 \left(1 - 2h + \mathcal{O}(h^2)\right)^n + C_2(-1)^n \left(1 + 2h + \mathcal{O}(h^2)\right)^n + \frac{1}{2}.$$

It is $\lim_{\varepsilon \rightarrow 0} (1 + \varepsilon)^{1/\varepsilon} = e$ and for $n = \frac{x_n}{h}$ (x_n fixed) it is

$$\lim_{h \rightarrow 0} (1 + 2h)^n = \lim_{h \rightarrow 0} (1 + 2h)^{\left(\frac{1}{2h}\right)2x_n} = e^{2x_n}$$

So we get

$$\lim_{h \rightarrow 0} (1 - 2h)^n = e^{-2x_n}, \quad h \rightarrow 0: \quad y_n = \underbrace{\left(C_1 e^{-2x_n} + \frac{1}{2}\right)}_{\rightarrow y_{\text{exact}}} + C_2(-1)^n e^{2x_n}$$

It is $C_2 \neq 0$, since we have replaced a first order differential equation by a second order difference equation. From the initial condition it follows that $C_2 = 0$ only if all computations were exact, but there are always round off errors and not exact starting values etc. So there is a small error at each step of integration.

Definition 1.34

A method is called **unstable** \Leftrightarrow The errors introduced in calculation grow exponentially as the computation proceeds.

Remark 1.35

One-step methods (like Runge-Kutta) are numerically stable.

Criteria for stable multistep method:

Take the corresponding difference equation of order k , compute the roots β_i ($i = 1, \dots, k$) of the characteristic equation. For $h \rightarrow 0$ β_1^n converges to the exact solution.

$$\begin{aligned} |\beta_i| < 1, \quad i = 2, \dots, k & \Leftrightarrow \text{the method is strongly stable} \\ \exists i |\beta_i| > 1 & \Rightarrow \text{the error grows exponentially} \end{aligned}$$

Example 1.36

$$f(x, y) = \lambda y$$

Adams-Bashforth:

$$y_{n-1} - y_n - \frac{h\lambda}{24}(55y_n - 59y_{n-1} + 37y_{n-2} - 9y_{n-3}) = 0$$

$$\Rightarrow \beta^4 - \beta^3 - \frac{h\lambda}{24}(55\beta^3 - 59\beta^2 + 37\beta - 9) = 0$$

$$h \rightarrow 0 : \quad \beta^4 - \beta^3 = 0 \quad \Rightarrow \quad \beta_1 = 1, \beta_2 = \beta_3 = \beta_4 = 0$$

$$h \neq 0 : \quad y_n = c_1\beta_1^n + c_2\beta_2^n + c_3\beta_3^n + c_4\beta_4^n$$

$$\Rightarrow \beta_i = \beta_i(h\lambda), i = 1, \dots, 4, \text{ are continuous in } h$$

$$\Rightarrow |\beta_i| < 1, i = 2, 3, 4, \text{ for } h \text{ small enough.}$$

Therefore the Adams-Bashforth method is strongly stable.

Milne:

$$y_{n+1} = y_{n-1} + \frac{h}{3}(f_{n+1} + 4f_n + f_{n-1}).$$

$$y_{n+1} - y_{n-1} - \frac{h\lambda}{3}(y_{n+1} + 4y_n + y_{n-1}) = 0; \quad \underbrace{\rho(\beta)}_{=\beta^2-1} + h\lambda \underbrace{\sigma(\beta)}_{\beta^2+4\beta+1} = 0$$

$$\rho(\beta) = 0 \quad \Rightarrow \quad \beta_1 = 1, \beta_2 = -1.$$

It follows that the Milne method is not strongly stable.

2 Numerical methods for boundary value problems

2.1 Shooting methods and collocation methods

2.1.1 Shooting methods

When solving a differential equation with a numerical method, it may happen that we choose a method that requires other boundary conditions than those given. Then we must “guess” the initial values. This brings us to the so-called shooting methods. Note that these also work for non-linear equations.

Consider

$$y'' - y = 0 \quad , \quad y(0) = 0 \quad , \quad y(1) = 1.$$

Apply initial value methods : These need $y(0)$, $y'(0)$ given. But here $y'(0)$ is unknown.
 $\Rightarrow y'(0) = \alpha$ unknown parameter here, to be determined such that

$$|\tilde{y}(1) - y_{exact}(1)| < \varepsilon.$$

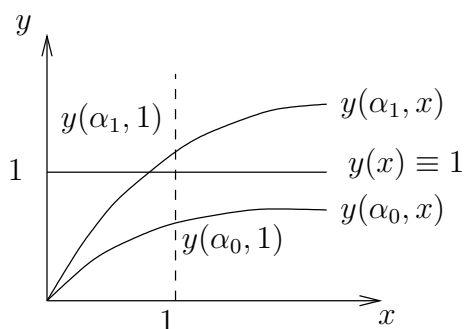
Trick : Guess α and iterate to $y'(0)$.

New initial conditions :

$$y(0) = 0 \quad , \quad y'(0) = \alpha_2 \quad \Rightarrow \quad y(\alpha_2, 1)$$

With the linear interpolation and using α_1 , α_2 we get α_3 and so $y(\alpha_3, 1)$. Repeat this until

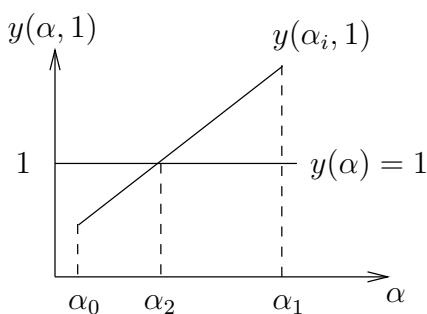
$$|y(\alpha_k, 1) - y(1)| < \varepsilon.$$



With α_0, α_1 we make two guesses for $y'(0)$ and get two solutions $y(\alpha_0, x), y(\alpha_1, x)$ of

$$\begin{aligned} z' &= y & , & & y(0) &= 0 \\ y' &= z & , & & z(0) &= y'(0) = \alpha_0 \text{ or } \alpha_1 . \end{aligned}$$

⇒ Numerical solutions by Runge-Kutta give $y(\alpha_k, x)$.



We obtain the next approximation from linear interpolation

$$(2.38) \quad \alpha_2 = \alpha_0 + (\alpha_1 - \alpha_0) \frac{y(1) - y(\alpha_0, 1)}{y(\alpha_1, 1) - y(\alpha_0, 1)}$$

Example 2.37

Solve

$$y'' - y = 0 \quad , \quad y(0) = 0 \quad , \quad y(1) = 1$$

with the shooting method. Start with initial approximations $\alpha_0 = 0.3$ and $\alpha_1 = 0.4$ for $y'(0)$ and $h = 0.1$.

Use Runge-Kutta (4th order) and the linear interpolation (2.38).

k	α_k	$y(\alpha_k; 1)$
0	0.3	0.35256072
1	0.4	0.47008103
2	0.85091712	0.99999999
3	0.85091712	0.99999999

$$y'_{exact}(x) \Big|_{x=0} = \sinh^{-1}(1) = 0.85091813$$

⇒ very rapid convergence.

2.1.2 Collocation methods

We seek an approximation u_N of the solution y of

$$(2.39) \quad Ly \equiv -y'' + p(x)y' + q(x)y = r(x) \quad , \quad a \leq x \leq b$$

with

$$(2.40) \quad \begin{aligned} a_0 y(a) - a_1 y'(a) &= \alpha, \\ b_0 y(b) + b_1 y'(b) &= \beta, \end{aligned}$$

where $|a_0| + |b_0| \neq 0$.

Let $\{\psi_j(x)\}$, $j = 1, \dots, N$ be a set of linear independent functions and

$$U_N(x) = \sum_{j=1}^N c_j \psi_j(x).$$

One possible idea is to choose the coefficients c_j such that: $\|y - U_N\| \stackrel{!}{=} \min$ in some norm $\|\cdot\|$. We will return to this approach in the next section. Another idea is

Collocation : Choose c_j such that $U_N(x)$ satisfies boundary conditions (2.40) and differential equation (2.39) exactly at selected points interior to $[a, b]$.

$$a_0 U_N(a) - a_1 U_N'(a) = \alpha$$

$$b_0 U_N(b) + b_1 U_N'(b) = \beta$$

$$L U_N(x_i) - r(x_i) = 0$$

for $i = 1, \dots, N - 2$, $x_i \in (a, b)$ distinct.

Choice of basis functions :

(i) $\psi_j(x) \in C^1[a, b]$.

(ii) $\psi_j(x)$ orthogonal over $[a, b]$, i.e.

$$\int_a^b \psi_j(x) \psi_k(x) dx = 0 \quad \text{for } j \neq k.$$

(iii) $\psi_j(x)$ “simple” functions like polynomials or trigonometric functions.

(iv) $\psi_j(x)$ satisfy homogenous boundary conditions (if any).

Example 2.38

(a) $\{\psi_j(x)\} = \{\sin j\pi x\}$, $j = 1, \dots, N$

$$\int_0^1 \sin j\pi x \sin k\pi x dx = 0 \quad (j \neq k), \quad \sin j\pi x = 0 \quad \text{at } x = 0, x = 1 \quad \forall j$$

(b) $\{\psi_j(x)\} = \{P_j(x)\}$, $j = 1, \dots, N$ Legendre polynomials

$$\int_{-1}^1 P_j P_k dx = 0, \quad j \neq k$$

(c) $\psi_j(x)$ piecewise-cubic polynomials

Example 2.39

$$U''(x) - U(x) = 0, \quad U(0) = 0, \quad U(1) = 1$$

$$U_N(x) = c_1 x + c_2 x^2 + c_3 x^3 \quad (\Rightarrow U_N(0) = 0), \quad U_N(1) = c_1 + c_2 + c_3 \stackrel{!}{=} 1.$$

We want $U_N(x)$ to satisfy the differential equation at two points in $(0, 1)$. Take $x_0 = \frac{1}{4}$, $x_1 = \frac{3}{4}$.

$$\begin{aligned}
 U_N''(x) - U_N(x) &= -c_1x + (2 - x^2)c_2 + (6x - x^3)c_3 \\
 \Rightarrow \quad U_N''\left(\frac{1}{4}\right) - U_N\left(\frac{1}{4}\right) &= -\frac{1}{4}c_1 + \frac{31}{16}c_2 + \frac{95}{64}c_3 = 0 \\
 U_N''\left(\frac{3}{4}\right) - U_N\left(\frac{3}{4}\right) &= -\frac{3}{4}c_1 + \frac{23}{16}c_2 + \frac{261}{64}c_3 = 0 \\
 \Rightarrow \quad c_1 &= 0.852237\dots, \quad c_2 = -0.0138527\dots, \quad c_3 = 0.161616\dots \\
 \Rightarrow \quad U_N(x) &= 0.852237x - 0.0138527x^2 + 0.161616x^3
 \end{aligned}$$

and this is an approximation for $U(x)$, $U'(x)$ at any $x \in [0, 1]$.

x	$U_N(x)$	$U(x)$
3.1	0.085247	0.085337
0.25	0.214719	0.214952
0.5	0.424675	0.443409
0.15	0.699567	0.699724
0.9	0.873611	0.873481

In general the convergence of the collocation methods is not assured. A much better method is the Ritz method which is examined in section 2.3.

2.2 Difference methods

For simplicity let us consider the following example

$$(2.41) \quad \boxed{y''(x) + f(x)y' + g(x)y = q(x), \quad y(a) = \alpha, y(b) = \beta, I = [a, b]}$$

and apply finite difference methods by substituting the derivatives by difference quotients. On $I = [a, b]$ we introduce a mesh with $x_0 = a$, $x_i = ih$, $i = 1, 2, \dots, N-1$, $x_N = b$ and use the following central difference approximations

$$y'(x_n) \approx \frac{y(x_{n+1}) - y(x_{n-1}))}{2h}, \quad y''(x_n) \approx \frac{y(x_{n+1}) - 2y(x_n) + y(x_{n-1}))}{h^2}$$

These approximations are of order h^2 .

This yields for (2.41) the system

$$\frac{y_{n-1} - 2y_n + y_{n+1}}{h^2} + \frac{f(x_n)}{2h}(y_{n+1} - y_{n-1}) + g(x_n)y_n = q(x_n), \quad n = 1, 2, \dots, N-1$$

where y_n approximates $y(x_n)$. Writing $f_n := f(x_n)$, $g_n := g(x_n)$ and $q_n := q(x_n)$ this system can be rewritten as

$$(2.42) \quad \left(1 - \frac{h}{2}f_n\right)y_{n-1} + (-2 + h^2g_n)y_n + \left(1 + \frac{h}{2}f_n\right)y_{n+1} = h^2q_n, \quad n = 1, 2, \dots, N-1.$$

This leads to a linear system with $N - 1$ equations and $N - 1$ unknowns

$$(2.43) \quad \begin{aligned} (-2 + h^2 g_1)y_1 + (1 + \frac{h}{2}f_1)y_2 &= h^2 q_1 - (1 - \frac{h}{2}f_1)\alpha \\ (1 - \frac{h}{2}f_2)y_1 + (-2 + h^2 g_2)y_2 + (1 + \frac{h}{2}f_2)y_3 &= h^2 q_2 \\ (1 - \frac{h}{2}f_3)y_2 + (-2 + h^2 g_3)y_3 + (1 + \frac{h}{2}f_3)y_4 &= h^2 q_3 \\ &\dots \quad \dots \quad \dots \\ (1 - \frac{h}{2}f_{N-2})y_{N-3} + (-2 + h^2 g_{N-2})y_{N-2} + (1 + \frac{h}{2}f_{N-2})y_{N-1} &= h^2 q_{N-2} \\ (1 - \frac{h}{2}f_{N-1})y_{N-2} + (-2 + h^2 g_{N-1})y_{N-1} &= h^2 q_{N-1} - (1 + \frac{h}{2}f_{N-1})\beta \end{aligned}$$

We get a tridiagonal system $Ay = b$ with

$$A = \begin{pmatrix} d_1 & c_1 & 0 & & & \\ a_2 & d_2 & c_2 & 0 & & \\ 0 & a_3 & d_3 & c_3 & 0 & \\ & \ddots & \ddots & \ddots & \ddots & 0 \\ & & 0 & a_{N-2} & d_{N-2} & c_{N-2} \\ & & & 0 & a_{N-1} & d_{N-1} \end{pmatrix}$$

Other boundary conditions: $y'(x_0) + \gamma y(x_0) = 0$ and $y(x_N) = \beta$.

We obtain an order h accuracy when we take

$$(2.44) \quad \frac{y(x_0 + h) - y(x_0)}{h} + \gamma y(x_0) = 0 \quad \Rightarrow \quad y_1 + (-1 + \gamma h)y_0 = 0.$$

in (2.42) with $n = 1$. This gives $y_0 = \frac{y_1}{1 - \gamma h}$ and

$$\left[(-2 + h^2 g_1) + \frac{1 - \frac{h}{2}f_1}{1 - \gamma h} \right] y_1 + (1 + \frac{h}{2}f_1)y_2 = h^2 q_1.$$

Replacing the first equation of (2.43) by this one and keeping all other equations of (2.43) unchanged leads also to a tridiagonal system. But (2.44) is only an $\mathcal{O}(h)$ -approximation. So the solution has $\mathcal{O}(h)$ accuracy only.

We obtain order h^2 accuracy using

$$\frac{y(x_0 + h) - y(x_0 - h)}{2h} + \gamma y(x_0) = 0 \quad \Rightarrow \quad y_1 - y_{-1} + 2h\gamma y_0 = 0.$$

But y_{-1} is an exterior point and so we have N unknowns: $y_0, y_1, y_2, \dots, y_{N-1}$. Taking $n = 0$ in (2.42) gives an additional equation

$$(1 - \frac{h}{2}f_0)y_{-1} + (-2 + h^2 g_0)y_0 + (1 + \frac{h}{2}f_0)y_1 = h^2 q_0$$

Eliminating $y_{-1} = y_1 + 2h\gamma y_0$ gives

$$\begin{aligned} \left[2h\gamma(1 - \frac{h}{2}f_0) + (-2 + h^2 g_0) \right] y_0 + 2y_1 &= h^2 q_0, \quad n = 0 \\ (1 - \frac{h}{2}f_1)y_0 + (-2 + h^2 g_1)y_1 + (1 + \frac{h}{2}f_1)y_2 &= h^2 q_1, \quad n = 1 \end{aligned}$$

and the remaining equations of (2.43). So we get N equations for N unknowns.

2.3 Variational methods, Ritz-methods

2.3.1 Variational methods

Consider the differential equation

$$(A.45a) \quad Lu := -\frac{d}{dx} \left(p(x) \frac{du}{dx} \right) + q(x)u(x) = f(x)$$

with boundary values

$$(A.45b) \quad u(a) = u(b) = 0$$

where $p \in C^1([a, b])$, $q, f \in C^0([a, b])$ and $p(x) \geq p_0 > 0$, $q(x) \geq 0 \forall x \in [a, b]$.

Define

$$\mathcal{D} := \{u \in C^2([a, b]) \mid u(a) = u(b) = 0\}.$$

Then (A.45) is equivalent to : Find $u \in \mathcal{D}$ with $Lu = f$.

Theorem 2.40

The operator L is symmetric and positive definite on \mathcal{D} .

Proof:

We must prove $(u, Lv) = (Lu, v) \quad \forall u, v \in \mathcal{D}$:

$$\begin{aligned} (u, Lv) &= \int_a^b u(x) \{-(p(x)v'(x))' + q(x)v(x)\} dx \\ &= \underbrace{-u(x)p(x)v'(x)}\Big|_a^b + \int_a^b p(x)u'(x)v'(x) + q(x)u(x)v(x) dx \\ &= 0, \text{ since } u \in \mathcal{D} \\ &= (v, Lu) \quad (\text{Symmetry in } u, v) \\ &= (Lu, v). \end{aligned}$$

For $u \neq 0$, $u \in [a, b]$ there holds

$$\begin{aligned} (Lu, u) &= \int_a^b \{p(x)(u'(x))^2 + q(x)(u(x))^2\} dx \\ &\geq p_0 \int_a^b (u'(x))^2 dx + \int_a^b q(x)(u(x))^2 dx \\ &\geq \frac{p_0}{(b-a)^2} \int_a^b (u(x))^2 dx > 0, \end{aligned}$$

since Cauchy-Schwarz yields

$$(u(x))^2 = \left(\int_a^x 1 \cdot u'(\xi) d\xi \right)^2 \leq \int_a^x 1^2 d\xi \int_a^x u'(\xi)^2 d\xi \leq (b-a) \int_a^x u'(\xi)^2 d\xi$$

and there thus holds

$$\int_a^b (u(x))^2 dx \leq \int_a^b (b-a) \int_a^b (u'(\xi))^2 d\xi = (b-a)^2 \int_a^b (u'(\xi))^2 d\xi.$$

■

Remark 2.41

$(u, Lv) = (Lu, v)$ not only exists for $u, v \in \mathcal{D}$, but also for piecewise differentiable functions whose first derivatives are square integrable.

Remark 2.41 motivates the following definition:

Definition 2.42

Let H^m be the space of all functions g for which $\|g\|_{H^m} < \infty$, where

$$\|g\|_{H^m}^2 := \int_a^b \sum_{v=0}^m (g^{(v)}(x))^2 dx.$$

Remark 2.43

the IVP (A.45) has exactly one solution $u \in H^2(a, b)$ with $u(a) = u(b) = 0$ for each $f \in L^2(a, b)$. This u depends is continuously on f , i.e.

$$\exists c > 0 : \|u\|_{H^2(a,b)} \leq c \|f\|_{L^2(a,b)}$$

(which justifies (A.45)).

Define the scalar product

$$(u, v) = \int_a^b u(x)v(x) dx \quad \forall u, v \in L^2(a, b)$$

und with this the bilinear form

$$(2.46) \quad [u, v] := (Lu, v) = \int_a^b \{p(x)u'(x)v'(x) + q(x)u(x)v(x)\} dx$$

for all $u, v \in H^2$ with $u(a) = u(b) = 0$. $[\cdot, \cdot]$ is symmetric and positive definite according to Theorem 2.40. For $v(a) = v(b) = 0$ the **potential energy** can be expressed by $I(v) : H^1 \rightarrow \mathbb{R}$, where

$$I(v) := [v, v] - 2(f, v) = \int_a^b \{p(x)(v'(x))^2 + q(x)(v(x))^2 - 2f(x)v(x)\} dx.$$

Theorem 2.44

Let u be the solution of (A.45). Then u minimizes the functional $I(v)$ over $v \in H^1(a, b)$ with $v(a) = v(b) = 0$, i.e. $I(u) < I(v) \quad \forall v \in H^1(a, b), v \neq u$.

Proof:

Take arbitrary $v \in H^1(a, b)$, $v(a) = 0$ and $u(x) \neq v(x)$. Then $Lu = f$ yields

$$\begin{aligned}
 I(v) &= [v, v] - 2(f, v) \\
 &= [v, v] - 2(Lu, v) && ([\cdot, \cdot] \text{ is a bilinear form}) \\
 &= ([v, v] - 2[u, v] + [u, u]) - [u, u] && (\text{extended by zero}) \\
 &= ([v - u, v] - [u, v - u]) - [u, u] \\
 &= [v - u, v - u] - [u, u] \\
 &> -[u, u] \\
 &= [u, u] - 2(Lu, u) \\
 &= I(u).
 \end{aligned}$$

Thus $u = L^{-1}f \in H^1(a, b)$, $u(a) = u(b) = 0$ is a well defined a solution of $\min I(v)$. ■

2.3.2 The Ritz method

Let Φ_j be linear independent functions and let

$$S^h = \left\{ v^h(x) = \sum_{j=1}^N \alpha_j \Phi_j(x), \alpha_j \in \mathbb{R} \right\} \subset H^1(a, b)$$

with $u(a) = 0 = u(b)$.

Restrict $I(v)$ on S^h ($v^h \in S^h$):

$$\begin{aligned}
 I(v^h) &=: \tilde{I}(\alpha_1, \dots, \alpha_N) \\
 &= \int_a^b \left[p(x) \left(\sum_{j=1}^N \alpha_j \Phi_j'(x) \right)^2 + q(x) \left(\sum_{j=1}^N \alpha_j \Phi_j(x) \right)^2 - 2f(x) \sum_{j=1}^N \alpha_j \Phi_j(x) \right] dx.
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial \tilde{I}}{\partial \alpha_i} &= \int_a^b p(x) \left(\Phi_i'(x) \sum_{k=1}^N \alpha_k \Phi_k'(x) + \sum_{j=1}^N \alpha_j \Phi_j'(x) \Phi_i'(x) \right) dx \\
 &\quad + \int_a^b q(x) \left(\Phi_i(x) \sum_{k=1}^N \alpha_k \Phi_k(x) + \sum_{j=1}^N \alpha_j \Phi_j(x) \Phi_i(x) \right) dx - 2 \int_a^b f(x) \Phi_i(x) dx \\
 &= 2 \left\{ \int_a^b \left(p(x) \sum_{j=1}^N \alpha_j \Phi_j'(x) \Phi_i'(x) + q(x) \sum_{j=1}^N \alpha_j \Phi_j(x) \Phi_i(x) \right) dx - \int_a^b f(x) \Phi_i(x) dx \right\} \stackrel{!}{=} 0
 \end{aligned}$$

This results in the linear equation system

$$(2.47) \quad A\alpha = \mathbf{b}, \quad A \in \mathbb{R}^{N \times N}, \quad \alpha, \mathbf{b} \in \mathbb{R}^N$$

$$A = (a_{ij}) \quad \text{mit } a_{ij} := [\Phi_i, \Phi_j] = \int_a^b \left(p(x) \Phi_i'(x) \Phi_j'(x) + q(x) \Phi_i(x) \Phi_j(x) \right) dx,$$

$$\mathbf{b} = (b_1, \dots, b_N)^T \quad \text{mit } b_i := (f, \Phi_i),$$

$$\alpha = (\alpha_1, \dots, \alpha_N)^T.$$

\Rightarrow A is symmetric and positive definite.

Proof:

Let $\mathbf{k} = (k_1, \dots, k_N)^T \neq 0$. Then $0 \neq \omega = \sum_{i=1}^N k_i \phi_i \in S^h$ with

$$0 < [\omega, \omega] = \sum_{i,j=1}^N [\phi_i, \phi_j] k_i k_j = \sum_{i,j=1}^N a_{ij} k_i k_j = \mathbf{k}^T A \mathbf{k}$$

(i.e. A is positive definite). ■

Therefore $A\boldsymbol{\alpha} = \mathbf{b}$ has a unique solution (which can be computed by Cholesky, e.g.).

Let $\boldsymbol{\alpha}^* = (\alpha_1^*, \dots, \alpha_N^*)^T$ be the unique solution. Then

$$I(\boldsymbol{\alpha}^*) = \boldsymbol{\alpha}^{*T} A \boldsymbol{\alpha}^* - 2\boldsymbol{\alpha}^{*T} \mathbf{b} = -\boldsymbol{\alpha}^{*T} A \boldsymbol{\alpha}^*,$$

which gives us

$$\begin{aligned} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T A (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) &= \boldsymbol{\alpha}^T A \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T A \boldsymbol{\alpha}^* + (\boldsymbol{\alpha}^*)^T A \boldsymbol{\alpha}^* \\ &= \boldsymbol{\alpha}^T A \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \mathbf{b} + (\boldsymbol{\alpha}^*)^T A \boldsymbol{\alpha}^* \\ &= I(\boldsymbol{\alpha}) + \boldsymbol{\alpha}^{*T} A \boldsymbol{\alpha}^*. \end{aligned}$$

Because A is positive definite, it is

$$(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T A (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) > 0$$

for $\boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$. Therefore the above considerations yield

$$0 < I(\boldsymbol{\alpha}) - I(\boldsymbol{\alpha}^*) \quad , \quad \boldsymbol{\alpha} \neq \boldsymbol{\alpha}^*$$

$$\Rightarrow \quad v^*(x) = \sum_{j=1}^N \alpha_j^* \Phi_j(x) \in S^h$$

which is the wanted function which minimizes $I(v)$.

$$\Rightarrow \quad \min_{v \in S^h} I(v) = I(v^*).$$

The above method of solving differential equations by solving the corresponding variational equation on a finite dimensional subspace is called the **Ritz** or **Ritz-Galerkin method**.

Theorem 2.45

Let $\bar{\boldsymbol{\alpha}} \in \mathbb{R}^N$ be the solution of (2.47), i.e.

$$u^h(x) = \sum_{j=1}^N \bar{\alpha}_j \Phi_j(x) \in S^h \quad \text{solves} \quad \min_{v \in S^h} I(v).$$

Let $u(x)$ be a solution of (A.45). Then there hold the following equivalent equations:

- (a) $[u - u^h, u - u^h] = \min_{v^h \in S^h} [u - v^h, u - v^h]$
- (b) $[u - u^h, v^h] = 0 \quad \forall v^h \in S^h$
- (c) $[u^h, v^h] = (f, v^h) \quad \forall v^h \in S^h$

Proof:

(a): Take arbitrary $v^h \in S^h$. It is

$$\begin{aligned} [u - v^h, u - v^h] &= [u - v^h, u] - [u - v^h, v^h] = [u, u] - [v^h, u] - [u, v^h] + [v^h, v^h] \\ &= [u, u] - 2[u, v^h] + [v^h, v^h] \\ &= [u, u] + I(v^h) \\ &\geq [u, u] + I(u^h) = [u - u^h, u - u^h]. \end{aligned}$$

(a) \Leftrightarrow (b):

For $v^h \in S^h$, $\varepsilon \in \mathbb{R}$ equation (a) yields:

$$\begin{aligned} [u - u^h, u - u^h] &\leq [u - u^h + \varepsilon v^h, u - u^h + \varepsilon v^h] \\ &= [u - u^h, u - u^h] + 2\varepsilon[u - u^h, v^h] + \varepsilon^2[v^h, v^h]. \end{aligned}$$

This gives us

$$\varepsilon^2[v^h, v^h] \geq 2\varepsilon[u^h - u, v^h],$$

which is only possible for all $\varepsilon \in \mathbb{R}$ if $[u^h - u, v^h] = 0$, which gives (b).

On the other hand, (b) yields for all $v^h \in S^h$:

$$[u - u^h - v^h, u - u^h - v^h] = \underbrace{[u - u^h, u - u^h] + [v^h, v^h]}_{\text{minimal for } v^h \equiv 0}.$$

Consider the left side of the above equation:

$$\begin{aligned} v^h \in S^h &\Rightarrow w^h := u^h + v^h \in S^h \\ &\Rightarrow [u - u^h - v^h, u - u^h - v^h] \text{ is minimal if } w^h \equiv u^h \end{aligned}$$

and thus follows (a).

(b) \Leftrightarrow (c): trivial. ■

Remark 2.46

Theorem 2.45 says that the Ritz solution $u^h \in S^h$ is the projection (with respect to the bilinear form (2.46)) of $\{H^1(a, b), u(a) = u(b) = 0\}$ onto the subspace S^h . A method with this property is called a **projection method**.

Example 2.47

Consider

$$-y'' = \sin x, \quad y(0) = y(\pi) = 0 \quad \Rightarrow \quad p \equiv 1, \quad q = 0, \quad f = \sin x.$$

(a) **One-dimensional approach** :

$$\Phi_1(x) = x(\pi - x), \quad \Phi_1' = \pi - 2x$$

$$a_{11} = \int_0^\pi (\Phi_1'(x))^2 dx = \int_0^\pi (\pi^2 - 4\pi x + 4x^2) dx = \frac{\pi^2}{3},$$

$$\begin{aligned} b_1 &= \int_0^\pi \Phi_1(x) f(x) dx = \int_0^\pi x(\pi - x) \sin x dx = \pi \int_0^\pi x \sin x dx - \int_0^\pi x^2 \sin x dx \\ &= \pi [\sin x - x \cos x]_0^\pi - [2x \sin x + (2 - x^2) \cos x]_0^\pi = \pi^2 + (2 - \pi^2) + 2 = 4 \end{aligned}$$

$$\Rightarrow \frac{\pi^2}{3}\alpha_1 = 4 \Rightarrow \alpha_1 = \frac{12}{\pi^3}.$$

It follows that

$$v^*(x) = \alpha_1 \Phi_1(x) = \frac{12}{\pi^3} x(\pi - x) \approx 0.387x(\pi - x).$$

(b) **Two-dimensional approach :**

$$\begin{aligned} \phi_1(x) &= \sin x \Rightarrow \phi_1'(x) = \cos x \\ \phi_2(x) &= \sin 2x \Rightarrow \phi_2'(x) = 2 \cos 2x \end{aligned}$$

$$a_{11} = \int_0^\pi (\phi_1'(x))^2 dx = \int_0^\pi \cos^2 x dx = \frac{\pi}{2},$$

$$a_{12} = \int_0^\pi \phi_1'(x)\phi_2'(x) dx = 2 \int_0^\pi \cos x \cos 2x dx = 0,$$

$$a_{22} = \int_0^\pi (\phi_2'(x))^2 dx = 4 \int_0^\pi \cos^2 2x dx = 2\pi,$$

$$b_1 = \int_0^\pi \phi_1(x)f(x) dx = \int_0^\pi \sin^2 x dx = \frac{\pi}{2},$$

$$b_2 = \int_0^\pi \phi_2(x)f(x) dx = \int_0^\pi \sin 2x \sin x dx = 0.$$

We obtain the matrix

$$A = \begin{pmatrix} \frac{\pi}{2} & 0 \\ 0 & 2\pi \end{pmatrix}$$

and therefore

$$\frac{\pi}{2}\alpha_1 = \frac{\pi}{2}, \quad 2\pi\alpha_2 = 0 \Rightarrow \alpha_1 = 1, \quad \alpha_2 = 0 \Rightarrow v^*(x) = \sin x.$$

This makes sense, as the exact solution is $y = \sin x$.

2.3.3 Finite elements in one dimension

Example 2.48

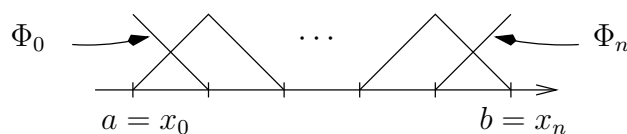
We consider the differential equation (A.45) with $p = 1$ and $q = 1$:

$$\begin{aligned} -u'' + u &= f, \quad x \in (a, b), \\ u(a) &= 0 = u(b). \end{aligned}$$

Choose the **basis functions** :

$$\Phi_i(x) = \begin{cases} \frac{x-a}{h} - i + 1 & , \quad a + (i-1)h \leq x \leq a + ih \\ -\frac{x-a}{h} + i + 1 & , \quad a + ih \leq x \leq a + (i+1)h \\ 0 & , \quad \text{otherwise} \end{cases} \quad (1 \leq i \leq N-1)$$

where $h = \frac{b-a}{n}$.



We obtain the matrix

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & \cdots & 0 \\ -1 & 2 & -1 & \vdots \\ & \ddots & \ddots & \ddots \\ \vdots & & -1 & 2 & -1 \\ 0 & \cdots & & -1 & 2 \end{pmatrix} + \frac{h}{6} \begin{pmatrix} 4 & 1 & \cdots & 0 \\ 1 & 4 & 1 & \vdots \\ & \ddots & \ddots & \ddots \\ \vdots & & 1 & 4 & 1 \\ 0 & \cdots & & 1 & 4 \end{pmatrix} \in \mathbb{R}^{N \times N}.$$

Lemma 2.49 (Friedrich's Lemma)

$$\exists K, \sigma > 0 \quad \forall v \in H^1(a, b) \text{ mit } v(a) = 0 : \quad \sigma \|v\|_{H^1}^2 \leq [v, v] \leq K \|v\|_{H^1}^2.$$

Proof:

(1) There holds

$$[v, v] = \int_a^b (p(v')^2 + qv^2) dx \leq \underbrace{\max_{a \leq x \leq b} (p, q)}_{=: K} \int_a^b (v^2 + (v')^2) dx.$$

This yields the right inequality of the proposition.

(2) First we have

$$\int_a^b p(v')^2 dx \geq p_0 \int_a^b (v')^2 dx.$$

Due to $v(a) = 0$ we also have

$$v(x_0) = \int_a^{x_0} v' dx \quad \forall x_0 \in [a, b].$$

Further, the Cauchy-Schwarz inequality gives us

$$|v(x_0)|^2 = \left| \int_a^{x_0} 1 \cdot v' dx \right|^2 \leq \int_a^{x_0} 1 dx \int_a^{x_0} (v')^2 dx \leq (b-a) \int_a^{x_0} (v')^2 dx.$$

So, integration over x_0 from a to b yields:

$$\int_a^b v^2 dx \leq (b-a) \int_a^b \left[\int_a^{x_0} (v')^2 dx \right] dx_0 \leq (b-a) \int_a^b \int_a^b (v')^2 dx dx_0 \leq (b-a)^2 \int_a^b (v')^2 dx,$$

which results in

$$\begin{aligned}
[v, v] &= \int_a^b (p(v')^2 + qv^2) dx \geq p_0 \int_a^b (v')^2 dx \\
&= \frac{p_0}{1 + (b-a)^2} \left[\int_a^b (v')^2 dx + (b-a)^2 \int_a^b (v')^2 dx \right] \\
&\geq \frac{p_0}{1 + (b-a)^2} \left[\int_a^b (v')^2 dx + \int_a^b v^2 dx \right] = \underbrace{\frac{p_0}{1 + (b-a)^2}}_{=: \sigma} \|v\|_{H^1}^2.
\end{aligned}$$

This is the left inequality. ■

This lemma means that the norm $\|v\|_L := [v, v]^{\frac{1}{2}}$ induced by $[\cdot, \cdot]$ is equivalent to the H^1 -Norm. The next lemma (proof omitted) gives us an estimate for the interpolation error depending on the grid width h .

Lemma 2.50

Let $v_I \in S^h$ be the piecewise interpolate of a function $v \in H^1(a, b)$ with nodes $x_i = a + ih$ ($0 \leq i \leq N$). Then there holds

$$\|v - v_I\|_{H^1} \leq Ch \|v\|_{H^1}.$$

We finally arrive at an estimate for the approximation error. .

Theorem 2.51

Let $u^h(x) \in S^h$ be a Ritz-Galerkin solution with linear finite elements and let $u(x)$ solve (A.45). Then there exists a $M > 0$ independent of u, f such that

$$\|u - u_h\|_{H^1(a,b)} \leq Mh \|f\|_{L^2(a,b)}.$$

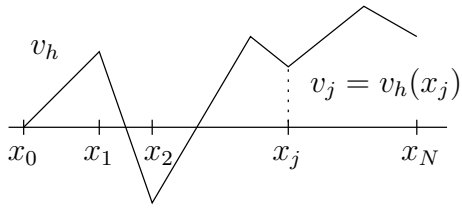
Proof:

Let $u_I \in S^h$ be a piecewise linear interpolate of $u(x)$ with nodes $x_i = a + ih$ ($0 \leq i \leq N$). Then Lemma 2.49, Theorem 2.45 (a) and Lemma 2.50 yield

$$\begin{aligned}
\|u - u_h\|_{H^1}^2 &\leq \sigma^{-1} [u - u_h, u - u_h] \leq \sigma^{-1} [u - u_I, u - u_I] \\
&\leq \sigma^{-1} K \|u - u_I\|_{H^1}^2 \leq \sigma^{-1} KCh^2 \|u\|_{H^1}^2.
\end{aligned}$$

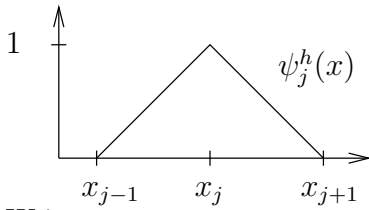
But since there holds $\|u\|_{H^2} \leq \tilde{c} \|f\|_{L^2}$, the statement of the lemma follows with $M = \sigma^{-1} KC^2 \tilde{c}$. ■

We now recapitulate the **Finite Element Method using piecewise linear basis functions**, only now we introduce a slight generalization: We no longer demand $x_j - x_{j-1} = h$ ($j = 1, \dots, N$); instead, we now allow $x_j - x_{j-1}$ to vary.



- $v_h \in S_h \iff$
- (1) v_h is linear in each interval.
 - (2) v_h is continuous ($\|v_h\|_{H^1} < \infty$)
 - (3) $v_h(0) = 0$

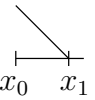
2.3.4 Representation of functions in S_h – a basis for S_h .



$$\psi_j^h(x) := \begin{cases} \frac{(x - x_{j-1})}{(x_j - x_{j-1})} & \text{in } (x_{j-1}, x_j) \\ \frac{(x_{j+1} - x)}{(x_{j+1} - x_j)} & \text{in } (x_j, x_{j+1}) \end{cases}$$

Write

$$v_h(x) = \sum_{j=1}^N v_j \psi_j^h(x) \quad , \quad v_j = v(x_j).$$

Note :  $\psi_0^h(x)$ is **not** included.

Thus $u_h(x) = \sum_{j=1}^N u_j \psi_j^h(x)$ where u_1, \dots, u_N are determined by

$$(2.48) \quad \sum_{k=1}^N u_k a(\psi_k^h, \psi_j^h) = \langle f, \psi_j^h \rangle \quad , \quad i \leq j \leq N.$$

Example 2.52

String vibration satisfies

$$\sum_{k=1}^N u_k \int_0^1 (p \psi_k^h(x) \psi_j^h(x) + q \psi_k^h \psi_j^h) dx = \int_0^1 f(x) \psi_j^h(x) dx.$$

Theorem 2.53

u_h is a solution of (2.48), (i.e. $a(u_h, v) = \langle f, v \rangle \quad \forall v \in S_h$ with given $f \in L^2(0, 1)$) if and only if

$$u_h(x) = \sum_{j=1}^N u_j \psi_j^h(x).$$

Here $\mathbf{u} = (u_1, \dots, u_N)^T$ is the solution of the algebraic system

$$K\mathbf{u} = \mathbf{f},$$

where

$$K = (a_{jk})_{j,k=1}^N \quad \text{with } a_{jk} = a(\psi_j^h, \psi_k^h),$$

$$\mathbf{f} = (f_1, \dots, f_N)^T \quad \text{with } f_j = \langle f, \psi_j^h \rangle.$$

Theorem 2.54

The stiffness matrix K is symmetric and positiv definite (\Rightarrow invertible).

Proof:

Symmetry of K is clear. To show that K is positive definite, first show that if $\mathbf{v} \neq \mathbf{0}$ then $\mathbf{v} \cdot K\mathbf{v} > 0$:

This is clear, since

$$\mathbf{v} \cdot K\mathbf{v} = \int_0^1 \left(p(v_h')^2 + qv_h^2 \right) dx \geq p_{\min} \int_0^1 (v_h')^2 dx.$$

But if $\mathbf{v} \cdot K\mathbf{v} = 0$ then

$$v_h(x) = \sum_{j=1}^N v_j \psi_j^h(x) = 0 \quad , \quad 0 \leq x \leq 1.$$

Since $\{\psi_k^h\}$ is a basis, we have $v_1 = v_2 = \dots = v_N = 0$, so $\mathbf{v} \cdot K\mathbf{v} \neq 0$ if $\mathbf{v} \neq \mathbf{0}$. ■

For the interpolate

$$u_I^h(x) = \sum_{j=1}^N u(x_j) \psi_j^h(x)$$

we have the following result. The following theorem is a generalization of Lemma (2.50).

Theorem 2.55

For any function $u \in H^2(0,1)$ with $u(0) = 0$ define the interpolate $u_I^h \in S_h$ such that $u_I^h(x_j) = u(x_j)$ ($1 \leq j \leq N$). Then there holds

$$\|u - u_I^h\|_s \leq Ch^{2-s} \|u\|_2 \quad (s = 0, 1, \dots)$$

where $C = \text{const} > 0$, independent of h , u and u_I^h .

Proof:

Let $e := u - u_I^h$. We note

$$(2.49) \quad \begin{cases} e(x_{j-1}) = e(x_j) = 0 & (1 \leq j \leq N) \\ e'' = u'' & \text{for } x_{j-1} \leq x \leq x_j \end{cases}.$$

Solve this boundary value problem (2.49) for e (note u'' is given!). There holds

$$e(x) = - \int_{x_{j-1}}^{x_j} G(x,y) u''(y) dy$$

with the **Green's function** :

$$G(x,y) = \begin{cases} \frac{(x - x_{j-1})(x_j - y)}{(x_j - x_{j-1})} & , \quad x_{j-1} \leq x < y \leq x_j \\ \frac{(x_j - x)(y - x_{j-1})}{(x_j - x_{j-1})} & , \quad x_{j-1} \leq y < x \leq x_j \end{cases}.$$

Hence

$$\begin{aligned} \int_{x_{j-1}}^{x_j} |e(x)|^2 dx &\leq \int_{x_{j-1}}^{x_j} \left(\int_{x_{j-1}}^{x_j} G(x, y)^2 dy \right) \left(\int_{x_{j-1}}^{x_j} (u'')^2 dy \right) dx \\ &\leq \underbrace{\left(\int_{x_{j-1}}^{x_j} dx \int_{x_{j-1}}^{x_j} G(x, y)^2 dy \right)}_{O(h^4)} \int_{x_{j-1}}^{x_j} (u'')^2 dy \end{aligned}$$

and

$$\int_{x_{j-1}}^{x_j} |e'(x)|^2 dx \leq \underbrace{\left(\int_{x_{j-1}}^{x_j} dx \int_{x_{j-1}}^{x_j} \left| \frac{\partial G}{\partial x} \right|^2 dy \right)}_{O(h^2)} \int_{x_{j-1}}^{x_j} (u'')^2 dy .$$

Summing up the terms over the whole interval $[x_0, x_N]$ and then taking the square root proves the statement. ■

Chapter B

Numerical methods for partial differential equations

1 Finite differences for elliptic equations

1.1 The finite difference method

$$(1.1) \quad \begin{aligned} u_{xx} + u_{yy} &= f(x, y) && \text{in } \Omega \\ u &= g(x, y) && \text{on } \partial\Omega \end{aligned}$$

where $\Omega := (0 < x < l) \times (0 < y < l)$.

Take $h = \frac{l}{4}$ and the grid points $(x_m, y_n) = (mh, nh)$, $0 \leq m, n \leq 4$.

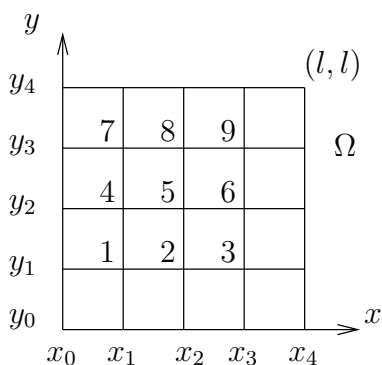
Notation : $U_{nj} \approx u$ at $x = nh$, $y = jh$.

Substituting the partial derivatives in (1.1) by central differences, we obtain the system of linear equations $Au = B$ given by

$$(1.2) \quad \left\{ \begin{aligned} \frac{U_{m+1,n} - 2U_{mn} + U_{m-1,n}}{h^2} + \frac{U_{m,n+1} - 2U_{mn} + U_{m,n-1}}{h^2} &= f_{mn}, && m \in \{1, 2, 3\}, n \in \{1, 2, 3\} \\ U_{mn} &= g_{mn}, && m \in \{0, 4\} \text{ or } n \in \{0, 4\} \end{aligned} \right.$$

where $f_{mn} = f(mh, nh)$. This is called the **finite difference method**, because derivatives are substituted by finite differences.

Example 1.1



We define

$$U_1 := U_{11}, U_2 := U_{21}, \dots$$

$$\text{If } g \equiv 0 \Rightarrow B_j \equiv -h^2 f_j.$$

From (1.2) then follows

$$\left(\begin{array}{ccc|ccc} 4 & -1 & 0 & -1 & & \cdots & & 0 \\ -1 & 4 & -1 & 0 & -1 & & & \\ 0 & -1 & 4 & 0 & 0 & -1 & & \\ \hline -1 & 0 & 0 & 4 & -1 & 0 & -1 & \\ & -1 & 0 & -1 & 4 & -1 & 0 & -1 \\ & & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ \hline \vdots & & & -1 & 0 & 0 & 4 & -1 & 0 \\ & & & & -1 & 0 & -1 & 4 & -1 \\ 0 & \cdots & & & & -1 & 0 & -1 & 4 \end{array} \right) \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \\ U_5 \\ U_6 \\ U_7 \\ U_8 \\ U_9 \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \\ B_6 \\ B_7 \\ B_8 \\ B_9 \end{pmatrix}$$

Lemma 1.2

Provided the boundary value problem (BVP) (1.1) has a unique solution and is sufficiently smooth, then the system

$$(1.3) \quad Au = B$$

has exactly one solution.

1.2 Convergence of point iteration methods

Let A be the matrix from (1.3). We decompose

$$A = -L + D - U$$

where L is the **lower triangular**, D the **diagonal** and U the **upper triangular** matrix. Suppose $\det D \neq 0$.

We could use the following **iterative solvers** for $Au = B$:

Jacobi :

$$u^{k+1} = T_J u^k + C_J \quad , \quad T_J := D^{-1}(L+U) \quad , \quad C_J := D^{-1}B$$

Gauss-Seidel :

$$u^{k+1} = T_G u^k + C_G \quad , \quad T_G := (D-L)^{-1}U \quad , \quad C_G := (D-L)^{-1}B$$

SOR :

$$u^{k+1} = T_\omega u^k + C_\omega \quad , \quad T_\omega := (D - \omega L)^{-1}[(1 - \omega)D + U] \quad , \quad C_\omega := (D - \omega L)^{-1}B$$

Theorem 1.3

The series $\{u^k\}$ defined by $u^{k+1} = Tu^k + C$ with arbitrary u^0 converges to a unique u^* with $Au^* = B \Leftrightarrow \rho(T) < 1$ where $\rho = \max |\lambda_j|$, λ_j eigenvalue of T . (ρ is called the **spectral radius** of T).

Theorem 1.4

If the matrix A in $Au = B$ fulfills

- (i) $a_{ij} \leq 0$ for $i \neq j$ and $a_{ii} > 0$ and
- (ii) $a_{ii} \geq \sum_{i \neq j} |a_{ij}|$ with strict inequality for some i ,

then both Jacobi and Gauss-Seidel iterations converge.

We want to study the eigenvalues of A . We can write

$$A = \begin{pmatrix} H & -I & 0 \\ -I & H & -I \\ 0 & -I & H \end{pmatrix} \quad \text{with} \quad H = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 4 \end{pmatrix}.$$

As we will show in the following, the **eigenvalues of A** are

$$\gamma_{ik} = \lambda_k - 2 \cos\left(\frac{j\pi}{4}\right) \quad (j = 1, 2, 3),$$

where λ_k is the k^{th} eigenvalue of H :

$$\lambda_k = 4 - 2 \cos\left(\frac{k\pi}{4}\right) \quad (k = 1, 2, 3).$$

$$\Rightarrow \quad \gamma_{ik} = 4 - 2 \left(\cos\left(\frac{j\pi}{4}\right) + \cos\left(\frac{k\pi}{4}\right) \right) \quad (j = 1, 2, 3, \quad k = 1, 2, 3).$$

Definition 1.5

A positive definite \Leftrightarrow A symmetric and $\langle Au, u \rangle > 0 \quad \forall u \neq 0$.

Remark 1.6

- (a) A positive definite \Rightarrow Solve (1.3) with SOR ($0 < \omega < 2$).
- (b) Symmetric matrix is positive definite \Leftrightarrow All eigenvalues are positive.

1.2.1 Eigenvalues of tridiagonal matrices

Let A be a $(N \times N)$ -matrix:

$$A =: \begin{pmatrix} b & c & \cdots & 0 \\ a & b & c & \vdots \\ & \ddots & \ddots & \ddots \\ \vdots & & a & b & c \\ 0 & \cdots & & a & b \end{pmatrix}$$

$$U_0 = U_{N+1} = 0 : \quad AU = \tilde{\lambda}U \quad \Leftrightarrow$$

$$(1.4) \quad aU_{n-1} + (b - \tilde{\lambda})U_n + cU_{n+1} = 0 \quad (1 \leq n \leq N)$$

This equation can be solved with the ansatz $U_n \sim r^n$:

Let r_1, r_2 be the solutions of

$$(1.5) \quad cr^2 + (b - \tilde{\lambda})r + a = 0.$$

Then set

$$U_n = \alpha r_1^n + \beta r_2^n \quad , \quad \alpha, \beta \in \mathbb{C}.$$

$$U_0 = U_{N+1} = 0 :$$

$$(1.6) \quad \begin{aligned} \alpha + \beta &= 0, \\ \alpha r_1^{N+1} + \beta r_2^{N+1} &= 0. \end{aligned}$$

It follows $\left(\frac{r_1}{r_2}\right)^{N+1} = 1$, and if $r_1 \neq r_2$ we have

$$(1.7) \quad \frac{r_1}{r_2} = e^{i2k\pi/(N+1)} \quad , \quad (1 \leq k \leq N) .$$

From (1.5) we get

$$(1.8) \quad \begin{aligned} r_{1,2} &= -\frac{(b-\tilde{\lambda})}{2c} \pm \sqrt{\left(\frac{b-\tilde{\lambda}}{2c}\right)^2 - \frac{a}{c}} \\ \Rightarrow \quad &\begin{cases} r_1 \cdot r_2 = \frac{a}{c} \neq 0 & \text{(by assumption)} \\ r_1 + r_2 = -\frac{b-\tilde{\lambda}}{c} \end{cases} \end{aligned}$$

and so it follows from (1.7) and (1.8):

$$r_1 = \sqrt{\frac{a}{c}} e^{ik\pi/(N+1)} \quad , \quad r_2 = \sqrt{\frac{a}{c}} e^{-ik\pi/(N+1)} .$$

Finally

$$(1.9) \quad \tilde{\lambda}_k = c(r_1 + r_2) + b = b + 2c\sqrt{\frac{a}{c}} \cos\left(\frac{k\pi}{(N+1)}\right) \quad (1 \leq k \leq N) .$$

1.2.2 Eigenvalues of block tridiagonal matrices

Let $H \in \mathbb{R}^{M \times M}$ with M distinct eigenvalues $\lambda_1, \dots, \lambda_M$.

Consider

$$(1.10) \quad AV = \gamma V$$

with

$$A = \begin{pmatrix} H & -I & \cdots & 0 \\ -I & H & -I & \vdots \\ & \ddots & \ddots & \ddots \\ \vdots & & -I & H & -I \\ 0 & \cdots & & -I & H \end{pmatrix}$$

where $I \in \mathbb{R}^{M \times M}$ and

$$V = \left[\alpha_1(U^k)^T, \alpha_2(U^k)^T, \dots, \alpha_N(U^k)^T \right]^T ,$$

where U^k is the eigenvector of H corresponding to λ_k and $\alpha_i \in \mathbb{R}$, $\alpha_i \neq 0$ for at least one i . From (1.10) we get the system

$$\begin{aligned} \alpha_1 H U^k &- \alpha_2 I U^k && = \gamma \alpha_1 U^k \\ -\alpha_1 I U^k &+ \alpha_2 H U^k &- \alpha_3 I U^k & = \gamma \alpha_2 U^k \\ && \cdots & \vdots \\ && & -\alpha_{N-1} I U^k + \alpha_N H U^k = \gamma \alpha_N U^k \end{aligned}$$

and from that with the equation $HU^k = \lambda_k U^k$:

$$(1.11) \quad \begin{pmatrix} \lambda_k & -1 & \cdots & 0 \\ -1 & \lambda_k & -1 & \vdots \\ & \ddots & \ddots & \ddots \\ \vdots & & -1 & \lambda_k & -1 \\ 0 & \cdots & & -1 & \lambda_k \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} = \gamma \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix}.$$

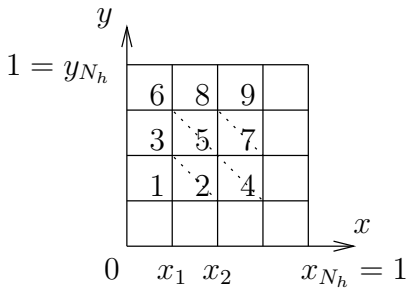
With (1.9) we get the eigenvalues of (1.11) which are the $M \cdot N$ eigenvalues of A :

$$\gamma_{jk} = \lambda_k - 2 \cos\left(\frac{j\pi}{N+1}\right) \quad (1 \leq k \leq M, 1 \leq j \leq N).$$

1.3 An example

We study the order of convergence of the difference method for the following problem:

$$(1.12) \quad \begin{aligned} -\Delta u &= -(u_{xx} + u_{yy}) = f(x, y) && \text{in } \Omega = (0, 1)^2, \quad f \in C^0(\bar{\Omega}) \\ u(x, y) &= 0 && \text{on } \partial\Omega. \end{aligned}$$



The grid points are defined as

$$(x_i, y_k) = (ih, kh) \in \Omega_h$$

with $i, k = 0, \dots, N_h$ and $N_h = \frac{1}{h}$.

Suppose $u \in C^4(\bar{\Omega})$ solves (1.12). Then there holds

$$(1.13) \quad u_{xx}(x_i, y_k) = \frac{1}{h^2} [u(x_{i+1}, y_k) - 2u(x_i, y_k) + u(x_{i-1}, y_k)] + \varepsilon_{i,k}(h),$$

$$(1.14) \quad u_{yy}(x_i, y_k) = \frac{1}{h^2} [u(x_i, y_{k+1}) - 2u(x_i, y_k) + u(x_i, y_{k-1})] + \eta_{i,k}(h)$$

with

$$\varepsilon_{i,k}(h) = \frac{h^2}{12} \left(\frac{\partial^4 u}{\partial x^4} \right) (x_i + \vartheta_1 h, y_k) \quad , \quad -1 \leq \vartheta_1 \leq 1,$$

$$\eta_{i,k}(h) = \frac{h^2}{12} \left(\frac{\partial^4 u}{\partial y^4} \right) (x_i, y_k + \vartheta_2 h) \quad , \quad -1 \leq \vartheta_2 \leq 1.$$

Inserting (1.13) and (1.14) in (1.1) gives us

$$(1.15) \quad \begin{aligned} & (-\Delta u)(x_i, y_k) - f(x, y) \\ &= \frac{1}{h^2} [4u(x_i, y_k) - u(x_{i-1}, y_k) - u(x_{i+1}, y_k) - u(x_i, y_{k-1}) - u(x_i, y_{k+1})] \\ &\quad - f(x, y) - \varepsilon_{i,k}(h) - \eta_{i,k}(h) \\ &= 0. \end{aligned}$$

By neglecting the remainder term

$$(1.16) \quad \varepsilon_{i,k}(h) + \eta_{i,k}(h) = O(h^2) \quad (i, k = 1, \dots, N_h - 1),$$

we compute the approximate values $u_{i,k}^h$ with the following system of linear equations:

$$(1.17) \quad \begin{aligned} -(L_h u^h)_{i,k} &= \frac{1}{h^2} \left(4u_{i,k}^h - u_{i-1,k}^h - u_{i+1,k}^h - u_{i,k-1}^h - u_{i,k+1}^h \right) \\ &= f(x_i, y_k) \quad (i, k = 1, \dots, N_h - 1). \end{aligned}$$

Set

$$\mathbf{u}(h) := [u(h, h), u(2h, h), u(h, 2h), \dots, u((N_h - 1)h, (N_h - 1)h)]^T$$

and the error vector $\varepsilon^h = \mathbf{u}^h - \mathbf{u}(h)$ in the same fashion.

We obtain the system

$$(1.18) \quad A(h)\mathbf{u}^h = \mathbf{b}(h) \quad \text{with} \quad \mathbf{b}(h) := h^2 \mathbf{f}.$$

Inserting (1.15) and (1.16) under consideration of the factor h^2 gives us

$$(1.19) \quad A(h)\mathbf{u}(h) = \mathbf{b}(h) + h^2 O(h^2).$$

Subtracting (1.19) from (1.18) gives $A(h)\varepsilon^h = h^2 O(h^2)$ and hence

$$(1.20) \quad \varepsilon^h = h^2 A(h)^{-1} O(h^2).$$

$O(h^2)$ is a $(N_h - 1)^2$ -vector, whose components can be estimated by Ch^2 .

Suppose

$$\|A(h)^{-1}\|_\infty \leq Kh^{-2} \quad \text{with} \quad K > 0,$$

where K is independent from h . Then we obtain from (1.20)

$$\|\varepsilon^h\|_\infty \leq Ch^2 K = Mh^2 \quad \text{mit} \quad M > 0,$$

where M is also independent from h . For $h \rightarrow 0$ there follows

$$|u_{i,k}^h - u(x_i, y_k)| = |u_{i,k}^h - u(ih, kh)| \leq Mh^2 \quad (i, k = 1, \dots, N_h - 1),$$

i.e. the order of convergence of the method is 2.

Theorem 1.7

Let $u \in C^4(\bar{\Omega})$ solve (1.12). Then the **finite difference method** (1.17) converges with 2nd order and there holds for $h \rightarrow 0$:

$$u_{i,k}^h - u(x_i, y_k) = O(h^2) \quad (i, k = 1, \dots, N_h - 1).$$

2 Difference Methods for Parabolic Equations

partial derivative = difference quotient + truncation error

Forward Difference:

$$u_t(x, t) = \frac{u(x, t+k) - u(x, t)}{k} - \frac{k}{2} u_{tt}(x, \bar{t}), \quad t < \bar{t} < t+k$$

Centered Difference

$$u_x(x, t) = \frac{u(x+h, t) - u(x-h, t)}{2h} - \frac{h^2}{6} u_{xxx}(\bar{x}, t), \quad x-h < \bar{x} < x+h$$

Centered Difference

$$u_{xx}(x, t) = \frac{u(x+h, t) - 2u(x, t) + u(x-h, t)}{h^2} - \frac{h^2}{12} u_{xxxx}(\bar{x}, t), \quad x-h < \bar{x} < x+h.$$

Differential Equation:

$$(2.21) \quad Lu = f, \quad (x, t) \in \Omega$$

Difference Equation

$$(2.22) \quad DU_{nj} = f_{nj}, \quad (x_n, t_j) \in \Omega$$

Local truncation error: $T_{nj} = Du_{nj} - f_{nj}$.

(2.22) is called **consistent** with (2.21) $:\Leftrightarrow \lim_{h,k \rightarrow 0} T_{nj} = 0$.

(2.22) is called **convergent** $:\Leftrightarrow \lim_{h,k \rightarrow 0} |U_{nj} - u_{nj}| = 0, (x_n, t_j) \in \Omega$.

Remark: A difference method can be consistent but not convergent.

Stability: Let U_{nj} satisfy (2.22) with initial values U_{n0} . Let V_{nj} be the solution to a perturbed difference system which differs only in the initial values, and write $V_{n0} \equiv U_{n0} + E_{n0}$. Then, assuming exact arithmetic, the initial perturbation, or "error", E_{n0} , can be shown to propagate, with increasing j , according to the homogeneous difference equation $DE_{nj} = 0$.

(2.22) is called **stable**, if E_{nj} is uniformly bounded in n as $j \rightarrow \infty$, i.e., if for some constant M and some positive integer J

$$(2.23) \quad |E_{nj}| < M, \quad (j < J)$$

If h and k must be functionally related for (2.23) to hold, the difference method is **conditionally stable**.

Remark: One of the concerns in applying a difference method is whether or not rounding errors in the calculation grow to such an extent that they dominate the numerical solution. When a stable method is used, rounding errors do not generally cause any difficulties.

Theorem 2.8 (Lax-Richtmyer Equivalence Theorem)

Given a well-posed initial-boundary value problem and a finite-difference problem consistent with it, stability is both necessary and sufficient for convergence.

von Neumann stability criterion: A difference method for an initial-boundary value problem with a bounded solution is *von Neumann stable* if every extended solution to $DU_{nj} = 0$ of the form

$$U_{nj} = \xi^j e^{i\beta n} \quad (\beta \text{ real}, \xi = \xi(\beta) \text{ complex})$$

has the property $|\xi| \leq 1$.

Stability in the von Neumann sense is a necessary condition for stability in the general sense (2.23).

2.1 Parabolic equations

Differential Equation:

$$u_t = a^2 u_{xx}$$

Explicit (Forward Difference) Method

$$(2.24) \quad \frac{U_{n,j+1} - U_{nj}}{k} = a^2 \frac{U_{n+1,j} - 2U_{nj} + U_{n-1,j}}{h^2}, \quad r := \frac{a^2 k}{h^2}.$$

Lemma 2.9

(2.24) has the local truncation error $\mathcal{O}(k + h^2)$. If $\frac{k}{h^2} = \frac{1}{6a^2}$ the local truncation error can be reduced to $\mathcal{O}(k^2 + h^4)$

Proof:

With $(x_n, t_j) = (nh, jk)$ there is

$$(u_t - a^2 u_{xx})_{nj} = \frac{u_{n,j+1} - u_{nj}}{k} - a^2 \frac{u_{n+1,j} - 2u_{nj} + u_{n-1,j}}{h^2} - \frac{k}{2} u_{tt}(x_n, \bar{t}_j) + a^2 \frac{h^2}{12} u_{xxxx}(\bar{x}_n, t_j)$$

where $t_j < \bar{t}_j < t_{j+1}$ and $x_{n-1} < \bar{x}_n < x_{n+1}$. The amount by which the solution of $u_t - a^2 u_{xx} = 0$ fails to satisfy the difference equation (2.24) is

$$T_{nj} = \frac{k}{2} u_{tt}(x_n, \bar{t}_j) - a^2 \frac{h^2}{12} u_{xxxx}(\bar{x}_n, t_j) = \mathcal{O}(k + h^2)$$

provided u_{tt} and u_{xxxx} are bounded.

Now, by Taylor's theorem and $(u_t - a^2 u_{xx})_{nj} = 0$,

$$(2.25) \quad \frac{u_{n,j+1} - u_{nj}}{k} - a^2 \frac{u_{n+1,j} - 2u_{nj} + u_{n-1,j}}{h^2} = \left[\frac{k}{2} u_{tt} - a^2 \frac{h^2}{12} u_{xxxx} \right]_{nj} + \mathcal{O}(k^2) + \mathcal{O}(h^4).$$

Since $u_t = a^2 u_{xx}$, $u_{tt} = a^2 u_{xxt} = a^2 (u_t)_{xx}$; whence

$$u_{tt} = a^2 (a^2 u_{xx})_{xx} = a^4 u_{xxxx}.$$

This shows that the bracketed terms in (2.25) can be written as

$$\left(\frac{k}{2} a^4 - a^2 \frac{h^2}{12} \right) u_{xxxx}(x_n, t_j)$$

which will be zero if we choose $k = h^2/6a^2$. ■

Lemma 2.10

If $r \leq 1/2$, then the explicit method (2.24) is convergent when applied to the problem

$$\begin{aligned} u_t - a^2 u_{xx} &= 0 & 0 < x < 1, t > 0 \\ u(x, 0) &= f(x) & 0 < x < 1 \\ u(0, t) = p(t), u(1, t) &= q(t) & t > 0 \end{aligned}$$

Proof:

Let Ω be the region $0 < x < 1$, $0 < t < T$; take $(x_n, t_j) = (nh, jk)$ for $n = 0, 1, 2, \dots, N$ and $j = 0, 1, \dots, J$, with $Nh = 1$ and $Jk = T$. Let U_{nj} satisfy the difference system

$$\begin{aligned} U_{n,j+1} &= U_{nj} + r(U_{n+1,j} - 2U_{nj} + U_{n-1,j}) & (r = a^2k/h^2) \\ U_{n0} &= f(x_n) & U_{0j} = p(t_j) & U_{Nj} = q(t_j) \end{aligned}$$

and set $w_{nj} = U_{nj} - u_{nj}$. Then w_{nj} satisfies

$$(2.26) \quad \begin{aligned} w_{n,j+1} &= rw_{n-1,j} + (1 - 2r)w_{nj} + rw_{n+1,j} + \frac{k^2}{2}u_{tt}(x_n, \bar{t}_j) - \frac{kh^2a^2}{12}u_{xxxx}(\bar{x}_n, t_j) \\ w_{n0} &= 0 & w_{0j} &= 0 & w_{Nj} &= 0 \end{aligned}$$

where $t_j < \bar{t}_j < t_{j+1}$ and $x_{n-1} < \bar{x}_n < x_{n+1}$.

If u_{tt} and u_{xxxx} are continuous and if we write

$$A = \max \left| \frac{1}{2}u_{tt}(x, t) \right| \quad B = \max \left| \frac{a^2}{12}u_{xxxx}(x, t) \right|$$

for (x, t) in $\bar{\Omega}$, then, since $r \leq 1/2$, it follows from (2.26) that

$$\begin{aligned} |w_{n,j+1}| &\leq r|w_{n-1,j}| + (1 - 2r)|w_{nj}| + r|w_{n+1,j}| + Ak^2 + Bkh^2 \\ &\leq \|w_j\| + Ak^2 + Bkh^2 \quad (\|w_j\| = \max_{0 < n < N} |w_{nj}|). \end{aligned}$$

From this we have

$$\|w_{j+1}\| \leq \|w_j\| + Ak^2 + Bkh^2.$$

Because $\|w_0\| = 0$ this implies

$$\|w_j\| \leq j(Ak^2 + Bkh^2) \leq T(Ak + Bh^2)$$

which shows that $|w_{nj}| \rightarrow 0$ uniformly in Ω as $h, k \rightarrow 0$. ■

Example 2.11

Use the von Neumann criterion to establish the condition $r \leq 1/2$ for the stability of the explicit method (2.24).

With (2.24) expressed in the form

$$(2.27) \quad U_{n,j+1} = rU_{n+1,j} + (1 - 2r)U_{nj} + rU_{n-1,j}$$

suppose that, at level j , an error is introduced at one or more of the x -nodes, perturbing the exact solution, U_{nj} , by an amount E_{nj} . If $U_{nj} + E_{nj}$ is used to advance the numerical solution to level $j + 1$, the result is the exact solution, $U_{n,j+1}$, plus an error, $E_{n,j+1}$. Putting $U_{nj} + E_{nj}$

and $U_{n+1,j} + E_{n+1,j}$ into (2.27), we see that E_{nj} satisfies the equation. Using separation of variables, we identify complex solutions of (2.27) of the form

$$(2.28) \quad a_j b_n = \xi^j e^{in\beta}$$

where ξ is some (possibly complex-valued) function of the real parameter β . From this, by superposition, we are led to the following expression for the error E_{nj} :

$$(2.29) \quad E_{nj} = \int_{-\infty}^{\infty} \xi(\beta)^j e^{in\beta} d\beta.$$

(Strictly, the real part of the integral should be taken.)

Substitution of (2.28) in (2.27) gives, after division by $\xi^j e^{in\beta}$,

$$\xi = re^{i\beta} + (1 - 2r) + re^{-i\beta} = 1 - 2r(1 - \cos \beta) = 1 - 4r \sin^2 \frac{\beta}{2}$$

and so $-1 \leq \xi \leq 1$ for all β - in particular, for $\beta = \pi$ - if and only if $0 \leq r \leq 1/2$.

Example 2.12

implicit (backwards)

$$\frac{U_{n,j+1} - U_{nj}}{k} = a^2 \frac{U_{n+1,j+1} - 2U_{n,j+1} + U_{n-1,j+1}}{h^2} \quad (r = a^2 k/h^2)$$

$$\Rightarrow -rU_{n-1,j+1} + (1 + 2r)U_{n,j+1} - rU_{n+1,j+1} = U_{nj}$$

Inserting $\xi^j e^{i\beta n}$ gives

$$\xi[-re^{i\beta} + (1 + 2r) - re^{-i\beta}] = 1 \quad \text{or } \xi = (1 + 4r \sin^2 \frac{\beta}{2})^{-1}$$

$$\Rightarrow |\xi| \leq 1 \forall \beta \text{ independent of } r$$

Chapter C

The CG algorithm revisited

1 Conjugate Gradient Method

For the algorithm and the notation look into the NuMa I script.

Theorem 1.1

If the conjugate gradient algorithm stops after the m^{th} step, we have $x^m = A^{-1}b$.

Proof:

If the algorithm has stopped after m steps we have $p_i \neq 0 \quad \forall i = 0, \dots, m-1$. Defining $r_j := Ax_j - b_j \quad (j = 0, \dots, m)$ we get by the construction of p_{j+1}

$$(1.1) \quad p_{j+1} = r_{j+1} - \beta_j p_j,$$

and

$$(1.2) \quad r_{j+1} = Ax_{j+1} - b = Ax_j - b - \alpha_j Ap_j = r_j - \alpha_j Ap_j.$$

Multiplication of this equation by p_j yields

$$(1.3) \quad \langle r_{j+1}, p_j \rangle = \langle r_j, p_j \rangle - \alpha_j \langle Ap_j, p_j \rangle = 0.$$

First, we show for all $k = 1, \dots, m-1$, that

$$\langle r_k, r_j \rangle = 0 \text{ and } \langle Ap_k, p_j \rangle = 0, \quad \forall j = 0, \dots, k-1,$$

by induction:

$k = 1$: We have $\langle Ap_1, p_0 \rangle = 0$ by the construction of β_0 and $\langle r_1, r_0 \rangle = 0$ by (1.3) (note $p_0 = r_0$).

$k \rightarrow k+1$:

$$\begin{aligned} \langle r_{k+1}, r_k \rangle &\stackrel{(1.1)}{=} \underbrace{\langle r_{k+1}, p_k \rangle}_{=0 \text{ (1.3)}} + \beta_{k-1} \langle r_{k+1}, p_{k-1} \rangle \\ &\stackrel{(1.2)}{=} \beta_{k-1} \underbrace{\langle r_k, p_{k-1} \rangle}_{=0 \text{ (1.3)}} - \beta_{k-1} \alpha_k \langle Ap_k, p_{k-1} \rangle = 0, \\ \langle r_{k+1}, r_0 \rangle &\stackrel{(1.2)}{=} \langle r_k, r_0 \rangle - \alpha_k \langle Ap_k, p_0 \rangle = 0. \end{aligned}$$

For all $0 < j < k$

$$\begin{aligned} \langle r_{k+1}, r_j \rangle &\stackrel{(1.2)}{=} \underbrace{\langle r_k, r_j \rangle}_{=0} - \alpha_k \langle Ap_k, r_j \rangle \\ &\stackrel{(1.1)}{=} -\alpha_k \langle Ap_k, p_j + \beta_{j-1} p_{j-1} \rangle = 0. \end{aligned}$$

We have $\langle Ap_{k+1}, p_k \rangle = 0$ by the construction of β_k and for all $j < k$

$$\begin{aligned} \langle Ap_{k+1}, p_j \rangle &\stackrel{(1.1)}{=} \langle r_{k+1}, Ap_j \rangle - \beta_k \underbrace{\langle p_k, Ap_j \rangle}_{=0} \\ &\stackrel{(1.2)}{=} \langle r_{k+1}, \frac{r_j - r_{j+1}}{\alpha_j} \rangle = 0. \end{aligned}$$

Since this has been shown we can discuss the conjugate gradient algorithm. If $m = n$ we have $x_n = A^{-1}b$.

If $m = 0$ we have $Ax_0 - b = r_0 = p_0 = 0$ and $r_0 = 0$ implies that x_0 is the solution of $Ax = b$.

If $0 < m < n$ we get ($p_m = 0$)

$$0 = \langle p_m, r_m \rangle \stackrel{(1.1)}{=} \langle r_m, r_m \rangle - \beta_{m-1} \underbrace{\langle p_{m-1}, r_m \rangle}_{=0 \text{ (1.3)}} = \langle r_m, r_m \rangle$$

and this implies $r_m = 0$ and thus $x_m = A^{-1}b$. ■

To show a further property of the conjugate gradient algorithm we define for all $k = 1, \dots, m$

$$V_k := \text{span}\{r_0, Ar_0, \dots, A^{k-1}r_0\} \quad , \quad r_0 := Ax_0 - b.$$

Then we have for all $k = 1, \dots, m$:

Lemma 1.2

$x_k^* := x_k - x_0$ satisfies

$$\langle Ax_k^*, \xi \rangle = -\langle r_0, \xi \rangle \quad , \quad \forall \xi \in V_k$$

and $x_k^* \in V_k$.

Proof:

First we show for all $i = 0, \dots, k-1$

$$\langle r_k, A^i r_j \rangle = 0 \quad , \quad \forall j = 0, \dots, k-i-1 \quad (r_i := Ax_i - b)$$

by induction.

$i = 0$: $\langle r_k, r_j \rangle = 0$, $\forall j = 0, \dots, k-1$ has been shown in the proof of Theorem 1.1.

$i \rightarrow i+1$: For $j = 0$ we have (note $i+1 \leq k-1$)

$$\langle r_k, A^{i+1} r_0 \rangle = \langle r_k, A^i Ap_0 \rangle \stackrel{(1.2)}{=} \langle r_k, A^i \frac{r_0 - r_1}{\alpha_0} \rangle = 0$$

and for $j = 1, \dots, k-(i+1)-1$

$$\begin{aligned} \langle r_k, A^{i+1} r_j \rangle &\stackrel{(1.1)}{=} \langle r_k, A^i (Ap_j + \beta_{j-1} Ap_{j-1}) \rangle \\ &\stackrel{(1.2)}{=} \langle r_k, A^i (\frac{r_j - r_{j+1}}{\alpha_j} + \beta_{j-1} \frac{r_{j-1} - r_j}{\alpha_{j-1}}) \rangle = 0, \end{aligned}$$

since $j+1 \leq k-i-1$ and $j-1 \geq 0$.

Thus:

$$\begin{aligned} & \langle r_k, A^i r_0 \rangle = 0 \quad , \quad \forall i = 0, \dots, k-1 \quad , \\ \iff & \langle Ax_k - b, \xi \rangle = 0 \quad , \quad \forall \xi \in V_k \quad , \\ \implies & \langle Ax_k^*, \xi \rangle = \langle Ax_k - b - \underbrace{(Ax_0 - b)}_{=r_0}, \xi \rangle = -\langle r_0, \xi \rangle \quad , \quad \forall \xi \in V_k \quad . \end{aligned}$$

Now we show $x_k^*, p_{k-1} \in V_k$, for all $k = 1, \dots, m$ by induction.

$$k = 1: x_1^* = x_1 - x_0 = x_0 - \alpha_0 p_0 - x_0 = -\alpha_0 r_0 \in V_1 .$$

$k \rightarrow k+1$: We assume $x_k^* = x_k - x_0 \in V_k$ and $p_{k-1} \in V_k$. Note:

$$(1.4) \quad x \in V_k \implies x \in V_{k+1} \text{ and } Ax \in V_{k+1}$$

$$\begin{aligned} x_{k+1}^* &= x_{k+1} - x_0 = x_k - x_0 - \alpha_k p_k \\ &\stackrel{(1.1)}{=} x_k^* - \alpha_k (r_k - \beta_k p_{k-1}) = x_k^* - \alpha_k (Ax_k - b - \beta_k p_{k-1}) \\ &= x_k^* - \alpha_k (Ax_k^* + \underbrace{Ax_0 - b}_{=r_0} - \beta_k p_{k-1}) \in V_{k+1} \quad (\text{see (1.4)}) \end{aligned}$$

■

2 The Preconditioned Conjugate Gradient Method

The preconditioner B (symmetric and positive definite) transforms the problem $Ax = b$ to

$$(2.5) \quad \hat{A}x = Bb =: \hat{b} \quad , \quad \hat{A} := BA .$$

\hat{A} is symmetric and positive definite with respect to the inner product $[\cdot, \cdot] := \langle B^{-1} \cdot, \cdot \rangle$. Applying the conjugate gradient algorithm using this inner product to solve the problem (2.5), we get the following algorithm:

Preconditioned Conjugate Gradient Algorithm

$x_0 \in \mathbb{R}^n$ starting vector (arbitrary)
 $p_0 := BAx_0 - Bb$
 $x_{k+1} := x_k - \hat{\alpha}_k p_k \quad , \quad k = 0, \dots, n-1$
 $p_{k+1} := BAx_{k+1} - Bb - \hat{\beta}_k p_k$

For $\hat{\alpha}_k$ we obtain

$$(2.6) \quad \hat{\alpha}_k = \frac{[\hat{A}x - \hat{b}, p_k]}{[\hat{A}p_k, p_k]} = \frac{\langle B^{-1}(BAx - Bb), p_k \rangle}{\langle B^{-1}BAp_k, p_k \rangle} = \alpha_k$$

and for $\hat{\beta}_k$

$$(2.7) \quad \hat{\beta}_k = \frac{[\hat{A}x_{k+1} - \hat{b}, \hat{A}p_k]}{[\hat{A}p_k, p_k]} = \frac{\langle Ax - b, BA p_k \rangle}{\langle Ap_k, p_k \rangle} .$$

Let \hat{x} be the solution of (2.5). Then we have

$$(2.8) \quad r_0 := p_0 = \hat{A}x_0 - \hat{b} = \hat{A}(x_0 - \hat{x})$$

and by Lemma 1.2:

$$(2.9) \quad \hat{x}_k^* = x_k - x_0 \in V_k = \{r_0, Ar_0, \dots, A^{k-1}r_0\},$$

$$(2.10) \quad [\hat{A}\hat{x}_k^* + r_0, \xi] = 0, \quad \forall \xi \in V_k .$$

The error $e_k := x_k - \hat{x}$ satisfies for all $\xi \in V_k$:

$$(2.11) \quad \begin{aligned} \hat{A}e_k &= \hat{A}(x_k - \hat{x}) \stackrel{(2.9)}{=} \hat{A}(\hat{x}_k^* + x_0 - \hat{x}) \stackrel{(2.8)}{=} \hat{A}\hat{x}_k^* + r_0 . \\ \implies [\hat{A}e_k, e_k]^2 &= [\hat{A}\hat{x}_k^* + r_0, \hat{x}_k^* + x_0 - \hat{x}]^2 \\ &\stackrel{(2.10)}{=} [\hat{A}\hat{x}_k^* + r_0, \xi + x_0 - \hat{x}]^2 = [\hat{A}e_k, \xi + x_0 - \hat{x}]^2 \\ &\leq [\hat{A}e_k, e_k] \cdot [\hat{A}(\xi + x_0 - \hat{x}), \xi + x_0 - \hat{x}] \end{aligned}$$

$$(2.12) \quad \implies [\hat{A}e_k, e_k] \leq [\hat{A}(\xi + x_0 - \hat{x}), \xi + x_0 - \hat{x}].$$

Let \tilde{Q}_k be a polynomial of degree k degree satisfying $\tilde{Q}_k(1) = 1$. By the smallest eigenvalue λ_{\min} and the largest eigenvalue λ_{\max} of \hat{A} we define a polynomial Q_k of degree k as follows:

$$(2.13) \quad Q_k(t) := \tilde{Q}_k\left(1 - \frac{2}{\lambda_{\min} + \lambda_{\max}}t\right) .$$

Since $Q_k(0) = \tilde{Q}_k(1) = 1$ we can find a polynomial P_{k-1} of degree $k-1$ such that

$$(2.14) \quad Q_k(t) = 1 + P_{k-1}(t) \cdot t .$$

We estimate the error e_k by Q_k :

$$(2.15) \quad \begin{aligned} Q_k(\hat{A})(x_0 - \hat{x}) &\stackrel{(2.14)}{=} (I + P_{k-1}(\hat{A})\hat{A})(x_0 - \hat{x}) \\ &\stackrel{(2.8)}{=} x_0 - \hat{x} + \underbrace{P_{k-1}(\hat{A})r_0}_{\in V_k} . \\ \stackrel{(2.12)}{\implies} [\hat{A}e_k, e_k] &\leq [\hat{A}Q_k(\hat{A})(x_0 - \hat{x}), Q_k(\hat{A})(x_0 - \hat{x})] \\ &\leq |||Q_k(\hat{A})|||^2 \cdot [\hat{A}(x_0 - \hat{x}), x_0 - \hat{x}] \end{aligned}$$

Defining

$$p(t) := 1 - \frac{2}{\lambda_{\max} + \lambda_{\min}}t \quad , \quad M := p(\hat{A}) \quad ,$$

we get by Lemma 2.3

$$(2.16) \quad \begin{aligned} \frac{\lambda_{\min} - \lambda_{\max}}{\lambda_{\max} + \lambda_{\min}} &= \min_{\lambda \in \sigma(\hat{A})} p(\lambda) \leq |||M||| \leq \max_{\lambda \in \sigma(\hat{A})} p(\lambda) = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} . \\ \implies \sigma(M) &\subseteq [-\rho, \rho] \quad , \quad \rho := \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} < 1 . \end{aligned}$$

Now we can estimate $|||Q_k(\hat{A})|||$

$$(2.17) \quad |||Q_k(\hat{A})||| = |||\tilde{Q}_k(M)||| \stackrel{2.3}{\leq} \max_{\lambda \in \sigma(M)} |\tilde{Q}_k(\lambda)| \stackrel{(2.16)}{\leq} \max_{\lambda \in [-\rho, \rho]} |\tilde{Q}_k(\lambda)| .$$

Now we introduce the Chebychev polynomials C_k :

$$C_k(t) := \begin{cases} \cos(k \cdot \arccos(t)), & |t| \leq 1 \\ \cosh(k \cdot \operatorname{arccosh}(t)), & |t| > 1 \end{cases}.$$

Setting $\tilde{Q}_k(t) := \frac{C_k(\frac{t}{\rho})}{C_k(\frac{1}{\rho})}$, we ensure that $\tilde{Q}_k(1) = 1$ and get by (2.17)

$$(2.18) \quad |||Q_k(\hat{A})||| \leq \max_{\lambda \in [-\rho, \rho]} |\tilde{Q}_k(\lambda)| = \frac{1}{C_k(\frac{1}{\rho})} \underbrace{\cos(k \arccos(\frac{t}{\rho}))}_{\leq 1} \leq \frac{1}{C_k(\frac{1}{\rho})}.$$

Let $\tau := \operatorname{arccosh}(\frac{1}{\rho})$. It is easy to show that τ satisfies $e^\tau = \frac{1 + \sqrt{1 - \rho^2}}{\rho}$. Using the definition of ρ we get

$$(2.19) \quad \begin{aligned} e^\tau &= \frac{1 + \sqrt{1 - \rho^2}}{\rho} = \frac{1}{\rho} \left(1 + \sqrt{(1 - \rho)(1 + \rho)} \right) \\ &= \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \left(1 + \sqrt{\frac{4 \cdot \lambda_{\min} \lambda_{\max}}{(\lambda_{\max} + \lambda_{\min})^2}} \right) \\ &= \frac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \left(1 + 2 \frac{\sqrt{\lambda_{\min} \lambda_{\max}}}{\lambda_{\max} + \lambda_{\min}} \right) \\ &= \frac{1}{\lambda_{\max} - \lambda_{\min}} (\lambda_{\max} + \lambda_{\min} + 2\sqrt{\lambda_{\min} \lambda_{\max}}) \\ &= \frac{(\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}})^2}{(\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}) \cdot (\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}})} \\ &= \frac{\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{\frac{1}{2}} + 1}{\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right)^{\frac{1}{2}} - 1} = \frac{\kappa(\hat{A})^{\frac{1}{2}} + 1}{\kappa(\hat{A})^{\frac{1}{2}} - 1} \end{aligned}$$

where $\kappa(\hat{A}) := \frac{\lambda_{\max}}{\lambda_{\min}}$ is called the condition number of \hat{A} . $\rho < 1$ implies

$$\begin{aligned} C_k\left(\frac{1}{\rho}\right) &= \cosh(k\tau) = \frac{1}{2}e^{k\tau}(1 + e^{-2k\tau}) \geq \frac{1}{2}e^{k\tau} \\ &\stackrel{(2.19)}{=} \frac{1}{2} \left(\frac{\kappa(\hat{A})^{\frac{1}{2}} + 1}{\kappa(\hat{A})^{\frac{1}{2}} - 1} \right)^k \end{aligned}$$

and so we get by (2.15) and (2.18)

$$(2.20) \quad [\hat{A}e_k, e_k]^{\frac{1}{2}} \leq 2 \cdot \left(\frac{\kappa(BA)^{\frac{1}{2}} - 1}{\kappa(BA)^{\frac{1}{2}} + 1} \right)^k [\hat{A}(x_0 - \hat{x}), x_0 - \hat{x}]^{\frac{1}{2}}.$$

Lemma 2.3

Let $C \in \mathbb{R}^{N \times N}$ be a symmetric matrix and $\sigma(C)$ denote the set of eigenvalues (spectrum) of C . Then with C and every polynomial $p(x)$ there holds

$$\min_{\lambda \in \sigma(C)} |p(\lambda)| \leq |||p(C)||| \leq \max_{\lambda \in \sigma(C)} |p(\lambda)|,$$

where $|||\cdot|||$ denotes the matrix norm

$$|||p(C)||| := \sup_{u \in \mathbb{R}^N} \frac{\langle p(C)u, u \rangle}{\langle u, u \rangle}.$$

Chapter D

Iterative methods for finding eigenvalues of symmetric matrices

1 The von-Mises method (power method, 1929)

Aim: Determination of certain eigenvalues; first that of largest absolute value (the **dominant eigenvalue**).

Remark 1.1

Let $A = A^T \in \mathbb{R}^{n \times n}$ (A symmetric). Then it is known that

(a) all eigenvalues are real. We ennumerate λ_i such that:

$$|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$$

(with respect to multiplicity);

(b) there are n linear independent eigenvectors $x^{(1)}, \dots, x^{(n)}$ with $Ax^{(i)} = \lambda_i x^{(i)}$ (A is **diagonalizable**). The eigenvectors can be chosen as to form an orthonormal basis, i.e.

$$(1.1) \quad x^{(i)T} x^{(k)} = \begin{cases} 1 & , \quad i = k \\ 0 & , \quad i \neq k \end{cases} .$$

Choose arbitrary $z^{(0)} \in \mathbb{R}^n$. Since the $x^{(i)}$ form a basis of \mathbb{R}^n , we can write:

$$z^{(0)} = c_1 x^{(1)} + \dots + c_n x^{(n)} .$$

Let $c_1 \neq 0$ (this can often not be verified, but it is “usually” satisfied. To make sure, one can always choose several different $z^{(0)}$).

v.Misesmethod :

$$(1.2) \quad z^{(v)} = A z^{(v-1)} \quad (v = 1, 2, \dots)$$

So we have

$$\begin{aligned}
 z^{(1)} &= A z^{(0)} = c_1 A x^{(1)} + \dots + c_n A x^{(n)} = c_1 \lambda_1 x^{(1)} + \dots + c_n \lambda_n x^{(n)} \\
 z^{(2)} &= A z^{(1)} = A^2 z^{(0)} = \dots = c_1 \lambda_1^2 x^{(1)} + \dots + c_n \lambda_n^2 x^{(n)} \\
 &\vdots \\
 (1.3) \quad z^{(v)} &= A z^{(v-1)} = A^v z^{(0)} = \dots = \underbrace{c_1 \lambda_1^v x^{(1)}}_{\star} + \dots + c_n \lambda_n^v x^{(n)}.
 \end{aligned}$$

★ outweighs the other terms if $c_1 \neq 0$.

For $v \rightarrow \infty$ we obtain the asymptotic behavior

$$(1.4) \quad z^{(v)} \sim c_1 \lambda_1^v x^{(1)}, \quad z^{(v+1)} \sim c_1 \lambda_1^{v+1} z^{(v)}.$$

Therefore if $z_i^{(v)} \neq 0$ there follows

$$q_i^{(v+1)} = \frac{z_i^{(v+1)}}{z_i^{(v)}} \rightarrow \lambda_1 \quad (i = 1, \dots, n).$$

When computing, the results are often normed so the numbers do not become too large or too small:

$$\hat{z}^{(1)} = A z^{(0)}, \quad z^{(1)} = \frac{\hat{z}^{(1)}}{\|\hat{z}^{(1)}\|_\infty} \quad (\text{etc.})$$

Example 1.2

$A (= A^T)$	$z^{(0)}$	$\hat{z}^{(1)}$	$z^{(1)}$	$\hat{z}^{(2)}$	$z^{(2)}$	
5 -2 -4	1	5	1	9	1	$z^{(v)} \rightarrow \begin{pmatrix} 1 \\ -0.5 \\ -1 \end{pmatrix} \quad \text{EV von } A$
-2 2 2	0	-2	-0.4	-4.4	-0.489	
-4 2 5	0	-4	-0.8	-8.8	-0.978	

Due to the norming, we now have convergence instead of the asymptotic behavior (1.4).

$$q_1^{(4)} = 9.9888, \quad q_2^{(4)} = 10.0086, \quad q_3^{(4)} = 10.0085.$$

There holds $q_j^{(v+1)} \rightarrow 10 = \lambda_1$. We have gut convergence, since $\lambda_2 = \lambda_3 = 1 \ll 10$.

From (1.3) there follows

$$(1.5) \quad \left| \lambda_1 - q_j^{(v+1)} \right| = \left| \frac{\lambda_2}{\lambda_1} \right|^v \cdot O(1),$$

i.e. the convergence for $|\lambda_2| \approx |\lambda_1|$ is only moderate.

Remark 1.3

The von-Mises method can also be applied to nonsymmetric ($A \neq A^T$) matrices, but it is vital that A possesses n linear independent eigenvectors.

1.0.1 Improvement for $A = A^T$ with the Rayleigh-quotient

Let $x \in \mathbb{R}^n$, $x \neq 0$.

Die real value

$$(1.6) \quad R[x] = \frac{x^T Ax}{x^T x}$$

is called the **Rayleigh-quotient** (of A for the vector x).

Theorem 1.4

Let $A = A^T$. The Rayleigh-quotient reaches its maximum resp. minimum at the eigenvector belonging to the largest resp. smallest eigenvalue.

Proof:

Let $0 \neq x \in \mathbb{R}^n$ with the representation

$$x = c_1 x^{(1)} + \dots + c_n x^{(n)},$$

where the $x^{(i)}$ are the eigenvectors of A . Then there holds

$$\begin{aligned} x^T Ax &= \left(c_1 x^{(1)T} + \dots + c_n x^{(n)T} \right) \left(\lambda_1 c_1 x^{(1)} + \dots + \lambda_n c_n x^{(n)} \right) \\ &\stackrel{(1.1)}{=} \lambda_1 c_1^2 + \dots + \lambda_n c_n^2 \end{aligned}$$

and $x^T x = c_1^2 + \dots + c_n^2$.

We conclude

$$\lambda_{min} \leq R[x] = \frac{\lambda_1 c_1^2 + \dots + \lambda_n c_n^2}{c_1^2 + \dots + c_n^2} \leq \lambda_{max}.$$

■

We can compute the Rayleigh-quotient $R[z^{(v)}]$ in the von-Mises method with little additional effort:

$$R[z^{(v)}] = \frac{z^{(v)T} A z^{(v)}}{z^{(v)T} z^{(v)}} = \frac{z^{(v)T} \hat{z}^{(v+1)}}{z^{(v)T} z^{(v)}}.$$

From (1.3) there follows

$$(1.7) \quad \left| \lambda_1 - R[z^{(v)}] \right| = \left(\frac{\lambda_2}{\lambda_1} \right)^{2v} \cdot O(1).$$

This is a considerable improvement compared to (1.5).

Example 1.2 yields:

$$R[z^{(3)}] = \frac{z^{(3)T} \hat{z}^{(4)}}{z^{(3)T} z^{(3)}} = 9.9997 < 10 = \lambda_1 = \lambda_{max}.$$

Here some further remarks to the von-Mises method.

Remark 1.5

(a) Let $\lambda_1 = \dots = \lambda_p$ and $|\lambda_p| > |\lambda_{p+1}|$. Then there follows

$$z^{(0)} = \underbrace{c_1 x^{(1)} + \dots + c_p x^{(p)}}_{=: y} + \sum_{i=p+1}^n c_i x^{(i)},$$

where y is an eigenvector for the eigenvalue λ_1 . This gives

$$z^{(v)} = \lambda_1^v y + \sum_{i=p+1}^n c_i \lambda_i^v x^{(i)},$$

and so there holds

$$q_i^{(v)} \rightarrow \lambda_1 \quad , \quad z^{(v)} \sim \lambda_1^v y,$$

i.e. there is practically no change in the behavior of q_i .

(b) Let $\lambda_1 = -\lambda_2$. Then $q_i^{(v)}$ is useless due to the “oscillation” of the iteration. Therefore take

$$\tilde{q}_i^{(v)} := \frac{z_i^{(v)}}{z_i^{(v-2)}}.$$

Then there holds $\tilde{q}_i^{(v)} \rightarrow \lambda_1^2$.

(c) Inverse Iteration

In many cases (oscillations, buckling loads, ...) the eigenvalue of lowest absolute value is of interest (i.g. $\neq 0$).

$$Ax = \lambda_n x \quad \Rightarrow \quad \frac{1}{\lambda_n} x = A^{-1} x$$

i.e. we must determine the dominant eigenvalue $K_n (= \frac{1}{\lambda_n})$ of A^{-1} .

von-Mises:

$$z^{(v+1)} = A^{-1} z^{(v)} \quad \Leftrightarrow \quad Az^{(v+1)} = z^{(v)}.$$

So we must solve a linear equation system at each iteration step, where the matrix is the same and only the right-hand side changes.

(d) Wielandt correction for eigenvalues

Let l be an approximation for λ_j ($1 \leq j \leq n$). Then $A - lE$ has the eigenvalues $\lambda_i - l$ ($i = 1, \dots, n$):

$$Ax = \lambda_i x \quad \Rightarrow \quad (A - lE)x = (\lambda_i - l)x.$$

If l is a “good” approximation of λ_j , then $\lambda_j - l$ is the eigenvalue of $A - lE$ of smallest absolute value.

(\Rightarrow inverse iteration: $(A - lE)z^{(v+1)} = z^{(v)}$.)

Critical: Note that $A - lE$ is “almost singular” and becomes more so, the closer l is to λ . But there is a “stable connection” to the QR algorithm.

(e) Determination of higher eigenvalues through matrix deflation

(e.g. λ_2 in oscillation problems)

Determine $x^{(1)}$, λ_1 approximately with von-Mises and $z^{(0)}$ such that $z^{(0)T} x^{(1)} = 0$.

$$z^{(0)} = \underbrace{c_1}_{=0} x^{(1)} + c_2 x^{(2)} + \dots + c_n x^{(n)} \quad , \quad x^{(1)T} x^{(1)} = 1,$$

then repeated application of the von-Mises-method.

But: Unfortunately $c_1 \neq 0$ due to approximation error (or rounding error), and so the method actually converges to λ_1 if executed long enough.

Better : The influence of λ_1 can be reduced by modification of the matrix A :

$$\begin{aligned} B &= A - \lambda_1 x^{(1)} x^{(1)T} \\ Bx^{(1)} &= Ax^{(1)} - \lambda_1 x^{(1)} \left(x^{(1)T} x^{(1)} \right) = 0 \\ Bx^{(i)} &= Ax^{(i)} - \lambda_1 x^{(1)} \underbrace{\left(x^{(1)T} x^{(i)} \right)}_{= 0} = \lambda_i x^{(i)} \end{aligned}$$

i.e. B has the eigenvalues $0, \lambda_2, \dots, \lambda_n$ with eigenvectors $x^{(1)}, x^{(2)}, \dots, x^{(n)}$.

in practice: Determine λ_1 , $x^{(1)}$ approximately.

\Rightarrow B may not have the exact eigenvalue 0, but λ_2 is certainly dominant.

2 The Jacobi method (1846)

Let $A = A^T$. Then the following matrices are called **Jacobi-Givens matrices** :

$$(2.8) \quad \Omega_{jk} = \begin{pmatrix} 1 & & & & & \cdots & 0 \\ & \ddots & & & & & \vdots \\ & & 1 & & & & \\ & & & c & & & -s \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & +s & & & c \\ & & & & & & & 1 \\ \vdots & & & & & & & \ddots \\ 0 & \cdots & & & & & & & 1 \end{pmatrix},$$

in which $c = \cos \rho$ und $s = \sin \rho$ for a given angle ρ with $-\frac{\pi}{4} \leq \rho \leq \frac{\pi}{4}$.

Ω_{jk} is a **orthogonal matrix**, i.e.

$$\Omega_{jk}^T \Omega_{jk} = I$$

Aim: $A \rightarrow A^{(1)} (= \Omega_{jk}^T A \Omega_{jk}) \rightarrow A^{(2)} (= \Omega_{j'k'}^T A^{(1)} \Omega_{j'k'}) \rightarrow \dots$

with

$$A^{(v)} \rightarrow D = \begin{pmatrix} \lambda_1 & & \cdots & 0 \\ & & & \vdots \\ & & \ddots & \\ \vdots & & & \\ 0 & \cdots & & \lambda_n \end{pmatrix},$$

where the λ_i are the eigenvalues of A (not necessarily arranged according to von-Mises).

Example 2.6 (Jacobi)

Denote the dominant element in A outside of the diagonal by $a_{jk} = a_{kj}$ ($j \neq k$). Determine the “angle of rotation” ρ of Ω_{jk} such that

$$a_{jk}^{(1)} = a_{kj}^{(1)} = 0$$

and then repeat the method. Note that a generated 0 can be replaced by a non-zero number in the following step, but the method converges nevertheless.

To generate a zero at the matrix coordinate (j, k) , use the following numerically stable formulas:

Formulas for $A = (a_{jk}) \rightarrow A^{(1)} = (a'_{jk})$:

$$(2.9) \quad \left\{ \begin{array}{l} \vartheta := \frac{a_{jj} - a_{kk}}{2a_{jk}} \quad (= \cot 2\rho), \\ t := \frac{s(\vartheta)}{|\vartheta| + \sqrt{1 + \vartheta^2}} \quad (= \tan \rho) \quad \text{with } s(\vartheta) = \begin{cases} 1 & , \vartheta \geq 0 \\ -1 & , \vartheta < 0 \end{cases}, \\ c := \frac{1}{\sqrt{1 + t^2}} \quad (= \cos \rho), \\ s := tc \quad (= \sin \rho), \\ \tau := \frac{s}{1 + c} \quad (= \tan \frac{\rho}{2}), \\ \Rightarrow \begin{array}{l} a'_{rj} = a'_{jr} = a_{rj} + s(a_{rk} - \tau a_{rj}) \quad , \quad r \neq j, k, \\ a'_{rk} = a'_{kr} = a_{rk} - s(a_{rj} + \tau a_{rk}) \quad , \quad r \neq j, k, \\ a'_{jj} = a_{jj} + ta_{jk}, \\ a'_{jk} = a'_{kj} = 0, \\ a'_{kk} = a_{kk} - ta_{jk}. \end{array} \end{array} \right.$$

Remark 2.7

- (a) In case ϑ is so large that ϑ^2 causes an overflow (which means that we have almost attained diagonal form), set $t = \frac{1}{2\vartheta}$.
- (b) Since A is similar to D , the product of the transformation matrices Ω_{jk} yields an orthonormal basis of eigenvectors of A .
- (c) The Jacobi method is computationally expensive but very stable. According to some authors, it has been replaced by the QR method.

5 With $R_k = Q_k^T(A_k - \sigma_k E)$ there follows for A_{k+1} :

3 The QR method

3.1 The QR decomposition

Definition 3.8

The QR decomposition of an $n \times n$ matrix is the (multiplicative) decomposition $A = QR$, where Q is an orthogonal matrix ($Q^{-1} = Q^T$) and R is an upper triangular matrix.

Theorem 3.9

Let A be an $n \times n$ matrix and Q an orthogonal $n \times n$ matrix. Then there holds

$$\text{cond}_2(QA) = \text{cond}_2(A).$$

Proof:

We have to show that

$$\text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2 \stackrel{!}{=} \text{cond}_2(QA) = \|QA\|_2 \|(QA)^{-1}\|_2,$$

where $\|A\|_2 = \sqrt{\sigma(A^T A)}$.

It is

$$A^T A = A^T Q^{-1} Q A = A^T Q^T Q A = (QA)^T (QA)$$

therefore $A^T A$ and $(QA)^T (QA)$ have the same eigenvalues and $\|A\|_2 = \|QA\|_2$.

$$((QA)^{-1})^T (QA)^{-1} = (A^{-1} Q^{-1})^T A^{-1} Q^{-1} = Q(A^{-1})^T A^{-1} Q^T.$$

On the right side there is a similarity transformation of $(A^{-1})^T A^{-1}$. Therefore $((QA)^{-1})^T (QA)^{-1}$ and $(A^{-1})^T A^{-1}$ have the same eigenvalues and it is $\|A^{-1}\|_2 = \|(QA)^{-1}\|_2$. ■

The QR decomposition (Algorithm)

First step: Calculate $A^{(1)}$:

If $a_{21} = \dots = a_{n1} = 0$ go to next step

1. $s := \sqrt{\sum_{i=1}^n a_{i1}^2}$
2. Calculate the vector $\omega^{(1)}$ by

$$\begin{aligned} \omega_1 &= \sqrt{\frac{1}{2} \left(1 + \frac{|a_{11}|}{s} \right)} \\ \omega_k &= \frac{1}{2\omega_1 s} \cdot \sigma(a_{11}) \text{ with } \sigma(t) = \begin{cases} 1 & t \geq 0 \\ -1 & t < 0 \end{cases} \end{aligned}$$

3. The transformation matrix

$$P^{(1)} = E - 2\omega^{(1)} \cdot \omega^{(1)T}$$

and calculate (Remark: $(P^{(1)})^{-1} = P^{(1)}$):

$$A^{(1)} = P^{(1)} A$$

Now in $A^{(1)}$ all elements in the the first column under a_{11} are 0.

Second step: Repeat the method with

$$P^{(2)} = E - 2\omega^{(2)} \cdot \omega^{(2)T}, \quad \text{where } \omega^{(2)} = \begin{pmatrix} 0 \\ * \\ \vdots \\ * \end{pmatrix}$$

etc.

finally (generated zeros are preserved)

$$R = P^{(n-1)} \dots P^{(1)} A =: Q^T A$$

So we get $A = QR$.

Theorem 3.10

A real $n \times n$ matrix can always be decomposed into a product $A = QR$, where Q is an orthogonal and R an upper triangular matrix.

Theorem 3.11

The QR decomposition of a real regular $n \times n$ matrix is unique, if the signs of the diagonal elements of R are given.

Proof:

Assume that

$$A = Q_1 R_1 = Q_2 R_2,$$

where the diagonal elements of R_1 and R_2 have the same signs.

The regular upper triangular matrices form a group with respect to the multiplication.

Because A is regular, R_1 and R_2 are regular.

$$D := Q_2^{-1} Q_1 = Q_2^T Q_1 = R_2 R_1^{-1}$$

Q_2^{-1} is orthogonal and therefore also $Q_2^T Q_1$. $R_2 R_1^{-1}$ is an upper triangular matrix. Therefore D is orthogonal and an upper triangular matrix. There holds

$$D^{-1} = D^T,$$

where D^{-1} is an upper and D^T a lower triangular matrix. Therefore D is a diagonal matrix.

Because $R_2 = DR_1$ and the diagonal elements of R_1 and R_2 have the same signs, D has only positive diagonal elements. Because D is orthogonal there holds $D^T D = E$ and therefore $D = E$. It follows that $R_1 = R_2$ and $Q_1 = Q_2$. ■

3.2 QR decomposition and linear equations systems

$$\begin{aligned} Ax = b &\Rightarrow Ax = QRx = b \\ &\Rightarrow Rx = y \quad \text{and} \quad Qy = b \end{aligned}$$

If one has calculated the QR decomposition it is easy to compute $y = Q^{-1}b = Q^T b$.

3.3 Shift**Theorem 3.12**

Let A be an $n \times n$ matrix and $a \in \mathbb{R}$. If A has got the eigenvalue λ then $A - aE$ has got the eigenvalue $\lambda - a$.

Proof:

$$\det((A - aE) - (\lambda - a)E) = \det(A - aE - \lambda E + aE) = \det(A - \lambda E) = 0.$$

■

3.4 The QR method (1961 Francis & Kublanowskaj)

First, take arbitrary $A \in \mathbb{R}^{n \times n}$.

If A is invertible ($\det A \neq 0$), there exists a LR decomposition $PA = LR$ with permutation matrices P . In contrast, the **QR-method** seeks an orthogonal matrix Q ($Q^{-1} = Q^T$) and a **QR decomposition of A** :

$$A = QR \quad , \quad R = \begin{pmatrix} * & \cdots & * \\ & \ddots & \vdots \\ \vdots & & \\ 0 & \cdots & * \end{pmatrix}.$$

The QR decomposition can be applied to any matrix; Q can be defined as product of Householder matrices or Jacobi-Givens matrices.

Iteration method:

$$(3.10) \quad \left\| \begin{aligned} A_1 &= A \quad \text{where } Q_i, R_i \text{ as above,} \\ A_i &= Q_i R_i, \\ A_{i+1} &= R_i Q_i \\ &= Q_i^T A_i Q_i \quad \text{where } Q_i^{-1} = Q_i^T \\ &= \dots \\ &= \underbrace{Q_i^T \cdots Q_1^T}_{=: U_i^T} A_1 \underbrace{Q_1 \cdots Q_i}_{=: U_i} \\ &= U_i^T A U_i \quad \text{where } U_i U_i^T = E. \end{aligned} \right.$$

All iteration matrices A_i ($i = 1, \dots, n$) are “orthogonally”-similar to one another.

The execution of the QR-decomposition using Jacobi-Givens matrices (2.8) is very costly for general matrices, but it is much less expensive when restricted to Hessenberg matrices resp. tridiagonal matrices (if $A = A^T$). This suggests that one should first use the methods of Wilkinson or Householder transform a general matrix A into Hessenberg form before applying the QR method.

Theorem 3.13

If the matrix A is regular and all the eigenvalues have pairwise different absolut values, then the matrices A_i converge to an upper triangular matrix.

Often the QR method is not used in the above mentioned form. Instead, one introduces a **spectral shift**:

To this aim, let $\sigma_k \in \mathbb{R}$:

$$(3.11) \quad \left\| \begin{aligned} A_1 &= A \\ A_k - \sigma_k E &= Q_k R_k \quad (\text{QR-decomposition of } A_k - \sigma_k E) \\ A_{k+1} &= R_k Q_k + \sigma_k E \end{aligned} \right.$$

Remark 3.14

With $R_k = Q_k^T(A_k - \sigma_k E)$ there follows for A_{k+1} :

$$A_{k+1} = Q_k^T(A_k - \sigma_k E)Q_k + \sigma_k E = Q_k^T A_k Q_k$$

Two methods for determining a “good” shift σ_k :

(a) **Rayleigh shift** : $\sigma_k = a_{nn}^{(k)}$.

(b) **Wilkinson- hift** :

$$C_k = \begin{pmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n-1,n}^{(k)} & a_{n,n}^{(k)} \end{pmatrix}$$

where $A = A^T \Rightarrow a_{n,n-1}^{(k)} = a_{n-1,n}^{(k)}$.

Determine the eigenvalues of C_k and set σ_k as the eigenvalue with smallest distance from $a_{n,n}^{(k)}$.

The QR algorithm

(A) Bring A to Hessenberg form

(B) Set $A^{(0)} := A$.

(I) Determination of the shift

(α) $\sigma = a_{nn}$ 'Rayleigh shift'

(β) σ as the eigenvalue of $\begin{pmatrix} a_{n-1,n-1} & a_{n-1,n} \\ a_{n,n-1} & a_{nn} \end{pmatrix}$ with smallest distance from a_{nn} , 'Wilkinson shift'.

(II) Shifting: $\tilde{A}^{(0)} = A^{(0)} - \sigma E$.

(III) QR decomposition: $\tilde{A}^{(0)} = QR$

(IV) Calculate $\tilde{A}^{(1)} = RQ = Q^{-1}\tilde{A}^{(0)}Q$

(V) Backwards shift: $A^{(1)} = \tilde{A}^{(1)} + \sigma E$

Repeat the steps (I) to (V) until $a_{n,n-1} \approx 0$. So the Hessenberg matrix A has been transformed via similarity transformations into a matrix of the form

$$\begin{pmatrix} a_{11} & \cdots & \cdots & \cdots & a_{1,n-1} & a_{1n} \\ a_{21} & \ddots & & & \vdots & \vdots \\ 0 & \ddots & \ddots & & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & a_{n-1,n-2} & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \cdots & 0 & 0 & 0 & a_{n,n} \end{pmatrix}$$

So, $a_{n,n}$ is an eigenvalue of A .

(C) Repeat the steps (I) to (V) with the $(n-1) \times (n-1)$ Hessenberg matrix until $a_{n-1,n-2} \approx 0$.

(D) Repeat the previous steps until one gets a triangular matrix with the eigenvalues of A in the diagonal.

Chapter E

Approximation of periodic functions with trigonometric polynomials

1 Representation theorem

Let f be continuous on $[0, 2\pi]$ with $f(0) = f(2\pi)$ and let f be extendable periodically, i.e. $f(x) = f(x \pm 2\pi)$ (e.g. periodic currents). Let l be the period length and $t = \frac{x}{l}2\pi$.

Trigonometric polynomial:

$$f \sim T_n(x) = \frac{a_0}{2} + \sum_{k=1}^n a_k \cos kx + b_k \sin kx \quad , \quad n \in \mathbb{N}.$$

Remark 1.1

One sometimes writes a_0 instead of $\frac{a_0}{2}$, though the fraction representation is more practical and more common.

We would like to determine the coefficients a_k, b_k , such that

$$S = \int_0^{2\pi} (f(x) - T_n(x))^2 dx \stackrel{!}{=} \min$$

This is the continuous form of the least squares method, confer (0.2) further on in the script for the discrete version. There we have a finite number of pairs x_i, y_i and

$$\sum_{i=1}^n \left(\underbrace{y_i}_{\hat{=} f} - \underbrace{f(x_i)}_{\hat{=} T_n} \right)^2 \stackrel{!}{=} \min \quad \Bigg) .$$

For S minimal there holds

$$\frac{\partial S}{\partial a_k} \stackrel{!}{=} 0 \quad , \quad \frac{\partial S}{\partial b_k} \stackrel{!}{=} 0$$

and this yields

$$(1.1) \quad \begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(x) \cos kx \, dx \quad (k = 0, 1, \dots, n), \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(x) \sin kx \, dx \quad (k = 1, \dots, n). \end{aligned}$$

The function $[0, 2\pi]$ -continuous function f kann be extended to a piecewise continuous periodic function f . Note that the left and right function limits exist in all points.

Now consider $T(x)$ for $n \rightarrow \infty$:

$$T(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kx + b_k \sin kx .$$

Representation problem: Does there hold $\forall x \in [0, 2\pi]$ (d.h. $\forall x \in \mathbb{R}$) $f(x) = T(x)$?

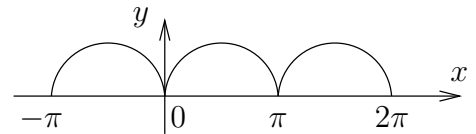
Theorem 1.2 (Representation theorem)

Let f, f' be piecewise continuous and all left and right function limits of f and of f' are supposed to exist. then $T(x)$ converges on $[0, 2\pi]$ and there holds $T(x_0) = f(x_0)$ for all continuous points x_0 of f , otherwise

$$T(x_0) = \frac{1}{2} \left(\lim_{h \rightarrow +0} (f(x_0 + h) + f(x_0 - h)) \right) .$$

Example 1.3

$$f(x) = x(\pi - x) \quad , \quad 0 \leq x \leq \pi$$



\Rightarrow extend π -periodically.

Computation of the integrals yields

$$T(x) = \frac{\pi^2}{6} - \left(\frac{\cos 2x}{1^2} + \frac{\cos 4x}{2^2} + \frac{\cos 6x}{3^2} + \dots \right) .$$

Theorem 1.2 gives us: $\forall x \quad f(x) = T(x)$.

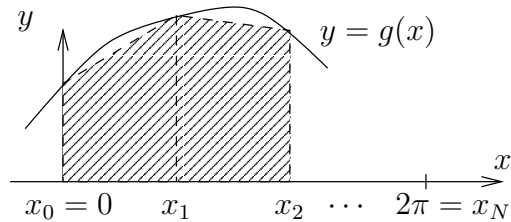
Set $x = 0$: $0 = \frac{\pi^2}{6} - \frac{1}{1^2} - \frac{1}{2^2} - \dots$

i.e. :

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6} .$$

2 Fast Fourier Transformation

The integrals in formulas (1.1) can often only be computed approximately.



Use trapezoid rule with the first $N + 1$ nodes

$$x_j = \frac{2\pi}{N} \cdot j \quad (j = 0, 1, \dots, N)$$

where $h = \frac{2\pi}{N}$.

$$\begin{aligned} \int_0^{2\pi} g(x) dx &\approx h \left[\frac{g(x_0) + g(x_1)}{2} + \frac{g(x_1) + g(x_2)}{2} + \dots + \frac{g(x_{N-1}) + g(x_N)}{2} \right] \\ &= h \sum_{j=1}^N g(x_j) \quad \text{da } g(x_0) = g(x_N). \end{aligned}$$

Remark 2.4

Due to periodicity and requested translation indifference other quadrature rules do not come into consideration.

Computation of the integrals in (1.1) gives us the approximate values a_k^* , b_k^* for a_k , b_k :

$$(2.2) \quad \begin{aligned} a_k^* &= \frac{2}{N} \sum_{j=1}^N f(x_j) \cos kx_j, \quad k \in \mathbb{N} \cup \{0\} \\ b_k^* &= \frac{2}{N} \sum_{j=1}^N f(x_j) \sin kx_j, \quad k \in \mathbb{N}. \end{aligned}$$

The functions $\cos kx$, $\sin kx$ define an orthogonal system:

$$\begin{aligned} \int_{-\pi}^{\pi} \cos(jx) \cos(kx) dx &= \begin{cases} 0 & \text{for } j \neq k \\ 2\pi & \text{for } j = k = 0 \\ \pi & \text{for } j = k > 0 \end{cases} \\ \int_{-\pi}^{\pi} \sin(jx) \sin(kx) dx &= \begin{cases} 0 & \text{for } j \neq k, j, k > 0 \\ \pi & \text{for } j = k > 0 \end{cases} \\ \int_{-\pi}^{\pi} \cos(jx) \sin(kx) dx &= 0 \quad \text{for } j \geq 0, k > 0. \end{aligned}$$

With the nodes x_j there hold the so-called **discrete orthogonal relations**:

$$\sum_{j=1}^N \cos(kx_j) \cos(lx_j) = \begin{cases} 0, & \text{if } \frac{k+l}{N} \notin \mathbb{Z} \text{ and } \frac{k-l}{N} \notin \mathbb{Z} \\ \frac{N}{2}, & \text{if either } \frac{k+l}{N} \in \mathbb{Z} \text{ or } \frac{k-l}{N} \in \mathbb{Z} \\ N, & \text{if } \frac{k+l}{N} \in \mathbb{Z} \text{ and } \frac{k-l}{N} \in \mathbb{Z} \end{cases}$$

$$\sum_{j=1}^N \sin(kx_j) \sin(lx_j) = \begin{cases} 0, & \text{if } \frac{k+l}{N} \notin \mathbb{Z} \text{ and } \frac{k-l}{N} \notin \mathbb{Z} \\ & \text{or } \frac{k+l}{N} \in \mathbb{Z} \text{ and } \frac{k-l}{N} \in \mathbb{Z} \\ -\frac{N}{2}, & \text{if } \frac{k+l}{N} \in \mathbb{Z} \text{ and } \frac{k-l}{N} \notin \mathbb{Z} \\ \frac{N}{2}, & \text{if } \frac{k+l}{N} \notin \mathbb{Z} \text{ and } \frac{k-l}{N} \in \mathbb{Z} \end{cases}$$

$$\sum_{j=1}^N \cos(kx_j) \sin(lx_j) = 0 \quad \text{for all } k, l \in \mathbb{N}_0.$$

In the following let N be even, i.e. $N = 2n$.

$$(2.3) \quad T_n^*(x) = \frac{a_0^*}{2} + \sum_{k=1}^{n-1} (a_k^* \cos kx + b_k^* \sin kx) + \frac{a_n^*}{2} \cos nx.$$

There holds:

- (i) The a_k^* , b_k^* ($k = 0, \dots, n$ bzw. $k = 1, \dots, n-1$) can be determined using the quadrature rule

$$T_n(x_j) = f(x_j) \quad (j = 1, \dots, N = 2n)$$

$$\left(\text{for } T_n(x) = \frac{\tilde{a}_0}{2} + \sum_{k=1}^{n-1} (\tilde{a}_k \cos kx + \tilde{b}_k \sin kx) + \frac{\tilde{a}_n}{2} \right).$$

Proof:

The interpolation conditions lead to a linear equation system whose coefficient determinant is unequal to zero (cf. Vandermonde).

\Rightarrow The interpolation problem has a unique solution which is determined by (2.2). ■

- (ii) If $m < n (= \frac{N}{2})$ then of all trigonometric polynomials of the form

$$T_m(x) = \frac{\tilde{a}_0}{2} + \sum_{k=1}^m \tilde{a}_k \cos kx + \tilde{b}_k \sin kx,$$

the polynomial $T_m^*(x)$ with $\tilde{a}_k = a_k^*$, $\tilde{b}_k = a_k^*$ approximates the function f at the nodes x_1, \dots, x_N best in the discrete least squares norm.

This means that

$$S = \sum_{j=1}^N (T_m(x_j) - f(x_j))^2$$

is minimized for $\tilde{a}_k = a_k^*$, $\tilde{b}_k = a_k^*$.

Proof:

$$\frac{\partial S}{\partial \tilde{a}_k} \stackrel{!}{=} 0, \quad \frac{\partial S}{\partial \tilde{b}_k} \stackrel{!}{=} 0$$

leads to the affirmation with help of the orthogonal relations. ■

It follows from (i) and (ii) that (2.2) and (2.3) are very well suited for approximating f .

2.0.1 Computation of the a_k^* , b_k^*

The formulas (2.2) require N^2 multiplications and N^2 trigonometric function evaluations. This gives us reason to seek a more suitable method.

There are two possibilities:

- (a) **The Runge scheme:** (antiquated)

Let $N = 4m$ ($m \in \mathbb{N}$)

$$x_{N-j} = (N-j)\frac{2\pi}{N} = 2\pi - x_j \quad , \quad x_{\frac{N}{2}-j} = \pi - x_j .$$

It follows

$$\begin{aligned} \cos kx_{4m-j} &= \cos kx_j \\ \sin kx_{4m-j} &= -\sin kx_j \\ \cos kx_{2m-j} &= (-1)^k \cos kx_j \\ \sin kx_{2m-j} &= (-1)^{k+1} \sin kx_j \end{aligned}$$

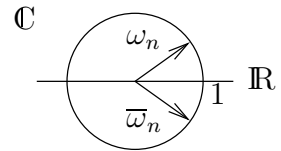
and similar indices in the sum of (2.2).

Altogether we obtain for the number of multiplications and trigonometric function evaluations: $\frac{1}{4}N^2$.

- (b) **The FFT algorithm:** (fast Fourier transformation)

Let $n = \frac{1}{2}N = 2^\gamma$ ($\gamma \in \mathbb{N}$). Set $b_0^* = 0 = b_n^*$ and execute all computations in \mathbb{C} where ω_n

$$\omega_n := e^{-i\frac{2\pi}{n}} ,$$



is the n -th complex root of 1 ($\omega_n^n = 1$). Further set

$$\begin{aligned} y_j &:= f(x_{2j}) + if(x_{2j+1}) \quad (j = 0, \dots, n-1) , \\ (2.4) \quad c_k &:= \sum_{j=0}^{n-1} y_j \omega_n^{jk} \quad (k = 0, \dots, n-1) . \end{aligned}$$

The a_k^* , b_k^* can be written using the c_k :

$$\begin{aligned} a_k^* - ib_k^* &= \frac{1}{N} [c_k + \bar{c}_{n-k}] - \frac{i}{N} [c_k - \bar{c}_{n-k}] e^{-\frac{ik\pi}{n}} , \\ a_{n-k}^* - ib_{n-k}^* &= \frac{1}{N} [\bar{c}_k + c_{n-k}] - \frac{i}{N} [\bar{c}_k - c_{n-k}] e^{\frac{ik\pi}{n}} \end{aligned}$$

for $k = 1, \dots, n$ (set $c_n = c_0$). This can be proven using Euler's formula.

Aim: We wish to combine the n sums c_k ($k = 0, \dots, n-1$) in (2.4) of length n to two times $\frac{n}{2}$ sums of length $\frac{n}{2}$, then combine these to sums of length $\frac{n}{4}$ and so on ...

Let k be even, i.e. $k = 2l$ ($l = 0, \dots, m - 1$) with $m = \frac{n}{2}$.
Then in (2.4) we have

$$\begin{aligned} y_j \omega_n^{2lj} + y_{m+j} \omega_n^{2l(m+j)} &= (y_j + y_{m+j}) \omega_n^{2lj} && \text{mit } \omega_n^{2lm} = 1 \\ &= (y_j + y_{m+j}) \omega_m^{lj} && \text{mit } \omega_m = e^{-i \frac{2\pi}{m}} \quad \text{d.h. } \omega_n^2 = \omega_m, \end{aligned}$$

and thus we obtain

$$(2.5) \quad \begin{aligned} z_j &= y_j + y_{m+j} \quad (j = 0, \dots, m - 1), \\ c_{2l} &= \sum_{j=0}^{m-1} z_j \omega_m^{jl} \quad (l = 0, \dots, m - 1) \end{aligned}$$

and analogously for k odd

$$(2.6) \quad \begin{aligned} z_{m+j} &= (y_j - y_{m+j}) \omega_n^j \quad (j = 0, \dots, m - 1), \\ c_{2l+1} &= \sum_{j=0}^{m-1} z_{m+j} \omega_m^{jl} \quad (l = 0, \dots, m - 1). \end{aligned}$$

Note that multiplications only arise in (2.6)!!

Each of these sums c_{2l} , c_{2l+1} of length m can be reduced by the same scheme ($n = 2^\gamma$), until we are left with sums of length 1.

The number of complex multiplications (only at (2.6)):

$$\frac{n}{2} + 2 \frac{n}{4} + 4 \frac{n}{8} + \dots = \frac{n}{2} \log_2 n.$$

This corresponds to $2n \cdot \log_2 n$ real multiplications instead of $O(n^2)$.

Chapter F

Discrete approximation problems

Serie of measurements:

(time)	t_i	1	2	3	4	5
(voltage)	y_i	0.14	0.38	0.52	0.76	0.63

Theory lets us expect

$$y_i = \alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2 \quad (i = 1, \dots, 5).$$

This cannot be exactly fulfilled by our experimental data. We therefore determine $\alpha_0, \alpha_1, \alpha_2$, such that the “total error” becomes as small as possible.

In general: We “fit” a polynomial

$$\alpha_0 + \alpha_1 t + \dots + \alpha_{n-1} t^{n-1} \quad (n < N)$$

through the N data points.

Let

$$r_i = \alpha_0 + \alpha_1 t_i + \alpha_2 t_i^2 + \dots + \alpha_{n-1} t_i^{n-1} - y_i$$

be the components of the **residual vector**, $\mathbf{r} = (r_1, \dots, r_N)^T$, also called **error vector**. This vector should become as small as possible, i.e. there should hold

$$\|\mathbf{r}\| \stackrel{!}{=} \min .$$

There are two main methods:

Discrete Chebyshev Approximation:

$$(0.1) \quad \|\mathbf{r}\|_\infty = \max_{i=1, \dots, N} |r_i| \stackrel{!}{=} \min_{\alpha_0, \dots, \alpha_{n-1}} .$$

Discrete Gaussian Approximation:

$$(0.2) \quad \|\mathbf{r}\|_2 = \sqrt{\sum_{i=1}^N r_i^2} \stackrel{!}{=} \min_{\alpha_0, \dots, \alpha_{n-1}} .$$

The Gaussian method is also called the **least squares approximation method**. In the following we will consider only (0.2), as (0.1) is very complicated, though also important. A first simplification arises from the fact that (0.2) is equivalent to

$$(0.3) \quad S = \|\mathbf{r}\|_2^2 = \sum_{i=1}^N r_i^2 \stackrel{!}{=} \min_{\alpha_0, \dots, \alpha_{n-1}} .$$

Because $S = S(\alpha_0, \dots, \alpha_{n-1})$ is a scalar function from \mathbb{R}^N to \mathbb{R} , the following condition is necessary for minimality (and also sufficient in this case):

$$\begin{aligned} \frac{\partial S}{\partial \alpha_j} &= \frac{\partial}{\partial \alpha_j} \sum_{i=1}^N (\alpha_0 + \dots + \alpha_{n-1} t_i^{n-1} - y_i)^2 \quad (j = 0, \dots, n-1) \\ &= \sum_{i=1}^N 2 (\alpha_0 + \dots + \alpha_{n-1} t_i^{n-1} - y_i) t_i^j = 0 . \end{aligned}$$

These are the so-called **normal equations**.

$$(0.4) \quad \begin{array}{ccccccc} \alpha_0 N & + & \alpha_1 \sum_{i=1}^N t_i & + & \dots & + & \alpha_{n-1} \sum_{i=1}^N t_i^{n-1} & = & \sum_{i=1}^N y_i \\ \alpha_0 \sum_{i=1}^N t_i & + & \alpha_1 \sum_{i=1}^N t_i^2 & + & \dots & + & \alpha_{n-1} \sum_{i=1}^N t_i^n & = & \sum_{i=1}^N t_i y_i \\ \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\ \alpha_0 \sum_{i=1}^N t_i^{n-1} & + & \alpha_1 \sum_{i=1}^N t_i^n & + & \dots & + & \alpha_{n-1} \sum_{i=1}^N t_i^{2n-2} & = & \sum_{i=1}^N t_i^{n-1} y_i \end{array}$$

Another somewhat antiquated form of writing is $\sum t_i^\rho = [t^\rho]$ etc.

(0.4) is a linear equation system with the symmetric coefficient matrix

$$A = \begin{pmatrix} N & \sum_{i=1}^N t_i & \dots & \sum_{i=1}^N t_i^{n-1} \\ \sum_{i=1}^N t_i & \sum_{i=1}^N t_i^2 & \dots & \sum_{i=1}^N t_i^n \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^N t_i^{n-1} & \sum_{i=1}^N t_i^n & \dots & \sum_{i=1}^N t_i^{2n-2} \end{pmatrix} .$$

It is easy to see that A is positive definit:

$$A = \begin{pmatrix} 1 & \dots & 1 \\ t_1 & \dots & t_N \\ \vdots & & \vdots \\ t_1^{n-1} & \dots & t_N^{n-1} \end{pmatrix} \cdot \begin{pmatrix} 1 & t_1 & \dots & t_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & t_N & \dots & t_N^{n-1} \end{pmatrix} =: C^T C ,$$

where $\text{rang } C = n$ (cf. Vandermonde).

Therefore a possible solution procedure for (0.4) is the Cholesky method. But the matrices are usually of high condition numbers ($\kappa_2(A) = 7377$ in the above example), so other methods are more preferable (as also for other linear ansatz, e.g. $y = \alpha_0 e^{-t} + \alpha_1 \sin t \dots$).

In general: Overdetermined system:

$$(0.5) \quad C\mathbf{x} + \mathbf{d} = \mathbf{r}$$

where $C \in \mathbb{R}^{N \times n}$ ($n < N$), $\mathbf{x} \in \mathbb{R}^n$ (analogously to $(\alpha_0, \dots, \alpha_{n-1})^T$), $\mathbf{d} \in \mathbb{R}^N$ (analogously to $-(y_1, \dots, y_N)^T$) and $\mathbf{r} = (r_1, \dots, r_N)^T \in \mathbb{R}^N$ is the residual vector.

Approximation according to Gauss, i.e. $\|\mathbf{r}\|_2^2 = \min_{\mathbf{x} \in \mathbb{R}^n}$ leads to a linear equation system as above:

$$A\mathbf{x} = -\mathbf{b} \quad \text{with } A = C^T C \in \mathbb{R}^{n \times n}, \quad \mathbf{b} = C^T \mathbf{d}.$$

We do not solve this system directly; instead we apply an orthogonal transformation to (0.5). Let $Q \in \mathbb{R}^{N \times N}$ be orthogonal.

$$(0.6) \quad \begin{aligned} Q^T(C\mathbf{x} + \mathbf{d}) &= Q^T \mathbf{r} = \mathbf{s}, \\ \|\mathbf{s}\|_2^2 &= \mathbf{s}^T \mathbf{s} = \mathbf{r}^T \underbrace{Q Q^T}_{= E} \mathbf{r} = \|\mathbf{r}\|_2^2 \end{aligned}$$

i.e. the norm of the residual vector is indifferent to Q -transformations, which means we can choose arbitrary Q .

Aim: Choose Q such that

$$C = Q\hat{R} \quad \text{mit } \hat{R} = \begin{pmatrix} R \\ 0 \end{pmatrix} \in \mathbb{R}^{N \times n}$$

with a regular triangular matrix

$$R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ \vdots & & \\ 0 & \cdots & r_{nn} \end{pmatrix}.$$

Inserting into (0.6) leads to

$$\begin{aligned} \underbrace{Q^T Q}_{= E} \hat{R} \mathbf{x} + \underbrace{Q^T \mathbf{d}} &= \underbrace{\mathbf{s}} \\ &= \boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_N \end{pmatrix} = \begin{pmatrix} s_1 \\ \vdots \\ s_N \end{pmatrix}. \end{aligned}$$

So we obtain the linear system

$$\begin{aligned} r_{11}x_1 + \dots + r_{1n}x_n + \delta_1 &= s_1 \\ &\vdots \\ r_{nn}x_n + \delta_n &= s_n \\ \delta_{n+1} &= s_{n+1} \\ &\vdots \\ \delta_N &= s_N \end{aligned}$$

$\Rightarrow \|\mathbf{r}\|_2^2 = \|\mathbf{s}\|_2^2$ is minimal for $s_1 = \dots = s_n = 0$, because $s_{n+1} = \delta_{n+1}, \dots, s_N = \delta_N$ are fixed anyway.

$\Rightarrow x_1, \dots, x_n$ are solutions of the system $R\mathbf{x} = -\boldsymbol{\delta}$.

Remark 0.1

We perform the orthogonal transformations with Householder matrices. Since we do not require any similarity transformations, we need only multiply one side of the system with Householder matrices. So we do not transform the system to a Hessenberg, but to a triangular matrix. Note that C is generally not quadratic.

Transformation matrices:

$$(P^T =) \quad P = E + \frac{1}{c} \omega \omega^T$$

with $c = \frac{1}{2} \omega^T \omega$. There holds $P^2 = E$.

For the first step we have

$$\omega^{(1)} = \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_N \end{pmatrix} \quad \text{mit } \omega^T \omega = 1,$$

i.e. $P = E - 2\omega\omega^T$.

$$C^{(1)} = P_1 C \stackrel{!}{=} \begin{pmatrix} * & & & \\ 0 & & & \\ \vdots & * & & \\ 0 & & & \end{pmatrix} \quad (\text{only one-sided!})$$

leads to

$$\omega_1 = \sqrt{\frac{1}{2} \left(1 + \frac{|c_{11}|}{\sqrt{\sum_{i=1}^N c_{i1}^2}} \right)}, \quad \omega_k = \text{sgn}(c_{11}) \frac{1}{2} \frac{c_{k1}}{\omega_1 \sqrt{\sum_{i=1}^N c_{i1}^2}} \quad (k = 2, \dots, N)$$

(cf. eigenvalue problems, stability). Repeat the method with $P_2 = E - 2\omega^{(2)}\omega^{(2)T}$ with

$$\omega^{(2)} = \begin{pmatrix} 0 \\ * \\ \vdots \\ * \end{pmatrix}$$

etc., finally (generated zeros are preserved)

$$\hat{R} = \underbrace{P_n P_{n-1} \cdots P_1}_{= Q^T} C.$$

Chapter G

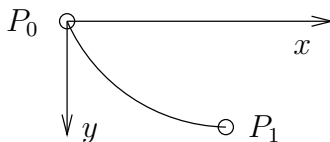
Calculus of variations - the Euler differential equation

1 Introduction

Example 1.1

(i) **The brachistochron** (*Johann Bernoulli 1696*)

Take a homogenous gravity field and points P_0, P_1 . We search the curve connecting P_0 to P_1 for which a point mass gliding without friction requires the least possible time to travel.



$$P_0 = (0, 0), \quad P_1 = (x_1, y_1), \quad y_1 > 0$$

Using elementary mechanics we can compute the falling time

$$T = \int_{P_0}^{P_1} \frac{1}{v} ds, \quad ds = \sqrt{1 + (y')^2} dx, \quad v = \sqrt{2gy},$$

i.e. we search (out of a class K of functions with certain differential qualities) the function y satisfying

$$I[y] = \int_0^{x_1} \frac{\sqrt{1 + (y')^2}}{\sqrt{y}} dx \stackrel{!}{=} \min_K$$

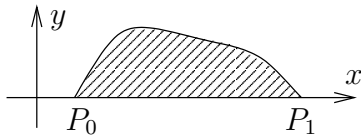
and with boundary values $y(0) = 0$, $y(x_1) = y_1$.

(ii) **Dido's problem** (*Carthagian queen*)

Given a rope of length l spanned between two points P_0, P_1 , we search the curve which cuts the largest area out of the plane:

with boundary values $y(x_0) = 0$, $y(x_1) = 0$ and the constraint

$$\int_{x_0}^{x_1} \sqrt{1 + (y')^2} dx = l.$$



$$I[y] = \int_{x_0}^{x_1} y \, dx \stackrel{!}{=} \max_K$$

The so-called **simplest variational problem** :

$$(1.1) \quad I[y] = \int_a^b f(x, y, y') \, dx \stackrel{!}{=} \min_K \quad , \quad y(a) = y_a \quad , \quad y(b) = y_b,$$

in which $K = \{y \in C^1[a, b] \mid y_1(x) \leq y(x) \leq y_2(x)\}$ and y_1, y_2 are given functions and

$$f : [a, b] \times [c, d] \times \mathbb{R} \rightarrow \mathbb{R} \quad \text{mit} \quad c \leq y_1(x) \leq y_2(x) \leq d .$$

f possesses continuous partial derivatives up to 2nd order.

Remark 1.2

Note that f in Example 1 (i) does not satisfy the continuity assumption at $x = 0$. We will disregard this fact.

2 Euler's differential equation

2.1 Derivation of a necessary condition for the solution of (1.1)

Let $u \in K$ be the solution of (1.1). Embed u in a class of "rival" functions:

$\eta \in C^1[a, b]$ arbitrary (i.e. η, η' continuous on $[a, b]$) with $\eta(a) = \eta(b) = 0$ and for $\varepsilon \in \mathbb{R}$ arbitrary set

$$y(x) = u(x) + \varepsilon\eta(x) .$$

For each η determine $\varepsilon_0 > 0$ such that for all ε with $|\varepsilon| \leq \varepsilon_0$ there holds $y \in K$. (The requirement $y(a) = u(a)$, $y(b) = u(b)$ is automatically fulfilled.) Because u is a solution, this means

$$I[u] = \int_a^b f(x, u(x), u'(x)) \, dx \leq \int_a^b f(x, \underbrace{y(x)}_{u + \varepsilon\eta}, \underbrace{y'(x)}_{u' + \varepsilon\eta'}) \, dx = I[u + \varepsilon\eta] .$$

For fixed η the real valued function $\rho(\varepsilon) = I[u + \varepsilon\eta]$ has a relative minimum at $\varepsilon = 0$, so there necessarily holds

$$\left. \frac{d\rho}{d\varepsilon} \right|_{\varepsilon=0} \stackrel{!}{=} 0 .$$

Remember: ($F : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ continuous, D is open)

(i) There holds

$$\frac{d}{dt} \int_a^b F(x, t) dx = \int_a^b \frac{\partial F(x, t)}{\partial t} dx$$

if $\frac{\partial F}{\partial t}$ is also continuous.

Generalization for variable limits $a(t)$, $b(t)$:

$$\frac{d}{dt} \int_{a(t)}^{b(t)} F(x, t) dx = \int_{a(t)}^{b(t)} \frac{\partial F(x, t)}{\partial t} dx + F(b(t), t) \cdot b'(t) - F(a(t), t) \cdot a'(t)$$

if a, b, a', b' are continuous.

(ii) **Chain rule for functions of several Variables**

$$\frac{d}{dt} F(x_1(t), x_2(t), \dots, x_n(t)) = \frac{\partial F}{\partial x_1} \cdot \frac{dx_1}{dt} + \dots + \frac{\partial F}{\partial x_n} \cdot \frac{dx_n}{dt} = \text{grad } F \cdot \dot{x}(t),$$

where $x = (x_1(t), \dots, x_n(t))^T$ and all functions are continuous.

(iii) **Partial integration**

$$\int_a^b u(x)v'(x) dx = u(x)v(x) \Big|_a^b - \int_a^b u'(x)v(x) dx$$

(follows from $(uv)' = u'v + v'u$).

So

$$\begin{aligned} \left. \frac{d\rho}{d\varepsilon} \right|_{\varepsilon=0} &= \frac{d}{d\varepsilon} \int_a^b f(x, u(x) + \varepsilon\eta(x), u'(x) + \varepsilon\eta'(x)) dx \\ &\stackrel{(i),(ii)}{=} \int_a^b \left[\frac{\partial f}{\partial y}(x, u + \varepsilon\eta, u' + \varepsilon\eta') \cdot \eta + \frac{\partial f}{\partial y'}(x, u + \varepsilon\eta, u' + \varepsilon\eta') \cdot \eta' \right] dx \Big|_{\varepsilon=0} \\ &= \int_a^b \left[\frac{\partial f}{\partial y}(x, u, u') \cdot \eta + \frac{\partial f}{\partial y'}(x, u, u') \cdot \eta' \right] dx \\ &\stackrel{!}{=} 0, \end{aligned}$$

where in the first sum we took the derivative with respect to the second variable y of f and in the second sum the derivative with respect to the third variable y' of f .

Partial integration of the second sum (assumed continuous):

$$\int_a^b \frac{\partial f}{\partial y'} \cdot \eta' dx = \frac{\partial f}{\partial y'} \cdot \eta \Big|_a^b - \int_a^b \frac{d}{dx} \left(\frac{\partial f}{\partial y'} \right) \cdot \eta dx$$

Since we assumed $\eta(a) = \eta(b) = 0$, the boundary terms vanish. Therefore the following condition is necessary for (relative) extrema (minimum or maximum):

$$(2.2) \quad \int_a^b \left[\frac{\partial f}{\partial y}(x, u(x), u'(x)) - \frac{d}{dx} \frac{\partial f}{\partial y'}(x, u(x), u'(x)) \right] \eta(x) dx = 0$$

for arbitrary $\eta \in C^1[a, b]$ with $\eta(a) = \eta(b) = 0$.
 The following lemma holds:

Lemma 2.3 (Fundamental lemma of variational calculus)

Let G be continuous on $[a, b]$. If

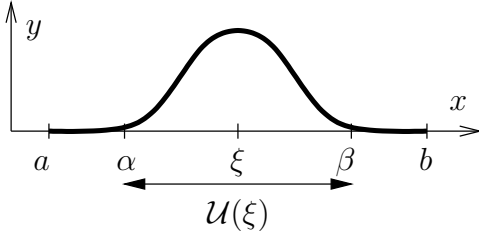
$$\forall \eta \in C^1[a, b], \eta(a) = \eta(b) = 0 \quad : \quad \int_a^b G(x)\eta(x) dx = 0,$$

then $G(x) \equiv 0$ on $[a, b]$.

Proof:

Suppose there exists a $\xi \in [a, b]$ such that $G(\xi) \neq 0$ (e.g. $G(\xi) > 0$, $\xi \in (a, b)$ is possible due to the continuity of G).

Since G is continuous, there exists an environment $\mathcal{U}(\xi)$ with $G(x) > 0$ for all $x \in \mathcal{U}(\xi) = (\alpha, \beta) \subset (a, b)$.



Now construct a suitable η , e.g.

$$\eta = \begin{cases} (x - \alpha)^2(\beta - x)^2 & , \quad x \in \mathcal{U}(\xi) \\ 0 & , \quad \text{otherwise} \end{cases}$$

$$\Rightarrow \eta \in C^1[a, b], \eta(a) = \eta(b) = 0.$$

Therefore we have

$$\int_a^b G(x)\eta(x) dx = \int_\alpha^\beta G(x)\eta(x) dx > 0,$$

since $G > 0$, $\eta > 0$ in (α, β) . Contradiction! ■

According to (2.2) und Lemma 2.3, a necessary condition for a relative extremum is

$$(2.3) \quad \frac{\partial f}{\partial y}(x, u(x), u'(x)) - \frac{d}{dx} \frac{\partial f}{\partial y'}(x, u(x), u'(x)) = 0.$$

(2.3) is called **Euler's differential equation** (of variational calculus). The solution curves of Euler's DE are called **extremals**. For (1.1) a solution u must satisfy the boundary conditions

$$u(a) = y_a, \quad u(b) = y_b$$

as well. (2.3) with these boundary conditions is called **Euler's boundary value problem**.

Using the chain rule (ii), it follows from (2.3) (assuming that $\frac{d}{dx} \frac{\partial f}{\partial y'}$ is differentiable)

$$(2.4) \quad \frac{\partial f}{\partial y} - \frac{\partial^2 f}{\partial y' \partial x} - \underbrace{\frac{\partial^2 f}{\partial y' \partial y} \cdot u'}_{= \frac{\partial}{\partial y} \left(\frac{\partial f}{\partial y'} \right) \cdot \frac{du}{dx}} - \frac{\partial^2 f}{\partial y' \partial y'} \cdot u'' = 0,$$

denoted by the **explicit 2nd order differential equation** for $\frac{\partial^2 f}{\partial y' \partial y'} \neq 0$.

Remark 2.4

- (i) In (1.1) we considered rival functions from K , i.e. $y \in C^1$. If the partial derivatives of f are continuous, then the solution $u(x)$ of (2.4) has a continuous partial derivative for $\frac{\partial^2 f}{\partial y' \partial y'} \neq 0$.
- (ii) (1.1) is also defined for rival functions which are continuous on $[a, b]$, piecewise continuously differentiable (differentiable except for a finite number of exceptions, but with existing left and right continuous function limits of y' everywhere). Then the solution $u(x)$ must also fulfill (2.3). (Even more general rival functions: $\mathcal{L}^2[a, b]$.)
- (iii) **Legendre's necessary condition:**
If u is a solution of the variational problem (1.1), then there necessarily holds (assuming continuity of the partial derivatives of f up to 2nd order)

$$\frac{\partial^2 f}{\partial y' \partial y'}(x, u(x), u'(x)) \geq 0 \quad \forall x \in [a, b]$$

(cf. $f''(x_{ex}) \geq 0$ for minimum).

The condition (2.3), Legendre with " > 0 " and a further condition (Jacobi) are also sufficient for relative minima of (1.1).

2.2 Special cases of Euler's differential equation

- (a) f depends only on x and y' : $f = f(x, y')$

Then

$$\frac{\partial f}{\partial y} = 0.$$

From (2.3) we obtain via integration

$$\frac{d}{dx} \frac{\partial f}{\partial y'} = 0 \quad \Rightarrow \quad \frac{\partial f}{\partial y'}(x, u'(x)) = \text{const}.$$

1st order differential equation

- (b) f depends only on y and y' (i.e. depends only implicitly on x): $f = f(y, y')$

Multiplication of (2.3) by u' leads to

$$\frac{\partial f}{\partial y} = \frac{d}{dx} \frac{\partial f}{\partial y'} \quad | \cdot u'.$$

There holds

$$u' \left(\frac{d}{dx} \frac{\partial f}{\partial y'} - \frac{\partial f}{\partial y} \right) = \frac{d}{dx} \left(u' \frac{\partial f}{\partial y'} - f \right),$$

since the right side can be transformed into

$$\frac{d}{dx} \left(u' \frac{\partial f}{\partial y'} - f \right) = u'' \frac{\partial f}{\partial y'} + u' \frac{d}{dx} \frac{\partial f}{\partial y'} - \frac{d}{dx} f = u'' \frac{\partial f}{\partial y'} + u' \frac{d}{dx} \frac{\partial f}{\partial y'} - \left[\frac{\partial f}{\partial y} u' + \frac{\partial f}{\partial y'} u'' \right].$$

Integration as in (a):

$$u'(x) \frac{\partial f}{\partial y'}(u(x), u'(x)) - f(u(x), u'(x)) = \text{const}.$$

Example 2.5

(i) Brachistochron

$$f(x, y, y') = \frac{\sqrt{1 + y'^2}}{\sqrt{y}}$$

Here we have case (b). (Not really a regular case because of $u(0) = 0$, but (1.1) exists, passing to the limit.)

$$u' \frac{\partial f}{\partial y'} - f = u' \cdot \frac{1}{\sqrt{u}} \cdot \frac{u'}{\sqrt{1 + u'^2}} - \frac{\sqrt{1 + u'^2}}{\sqrt{u}} = \frac{1}{c} = \text{const},$$

where

$$\frac{u'}{\sqrt{1 + u'^2}} = \frac{\partial}{\partial y'} \left(\sqrt{1 + y'^2} \right) \Big|_{y = u \quad (y' = u')},$$

$$\Rightarrow c^2 = u(1 + u'^2) \quad \text{resp.} \quad \frac{du}{dx} = \sqrt{\frac{c^2 - u}{u}}$$

which is a 1st order differential equation which can be solved by separation of variables. Parameter representation of u with parameter τ (cycloids):

$$u = \frac{c^2}{2}(1 + \cos \tau), \quad x = \frac{c^2}{2}(\tau + \sin \tau) + k$$

(where k is an integration constant generated by solving the 1st order DE). k and c are obtained from the boundary conditions

$$u(0) = 0, \quad u(x_1) = y_1.$$

(Tracks and roads are often described using similar curves, although civil engineers must often take other facts into consideration as well.)

$$\begin{aligned} u(0) = 0 &= \frac{c^2}{2}(1 + \cos \tau_0) && \stackrel{c \neq 0}{\Rightarrow} \tau_0 = \pi \\ x = 0 &= \frac{c^2}{2}(\tau_0 + \sin \tau_0) + k && \Rightarrow k = -\pi \frac{c^2}{2}, \end{aligned}$$

so

$$u = \frac{c^2}{2}(1 + \cos \tau), \quad x = \frac{c^2}{2}(\tau + \sin \tau - \pi).$$

Derive $\frac{c^2}{2}$ from $y_1 = \frac{c^2}{2}(1 + \cos \tau_1)$, $x_1 = \frac{c^2}{2}(\tau_1 + \sin \tau_1 - \pi)$.

The Legendre condition (iii) applied to brachistochrons:

$$\frac{\partial^2 f}{\partial y' \partial y'} = \frac{1}{\sqrt{u}(1 + u'^2)^{\frac{3}{2}}} > 0.$$

3 Extensions and generalizations

3.1 Natural boundary conditions

Given the problem (1.1) without boundary conditions $y(a) = y_a$, $y(b) = y_b$. Derivation of Euler's DE as above with the embedding approach

$$y(x) = u(x) + \varepsilon \eta(x)$$

(without the necessary condition $\eta(a) = \eta(b) = 0$). Using partial integration, one obtains

$$(3.5) \quad 0 = \frac{\partial f}{\partial y'} \cdot \eta \Big|_a^b + \int_a^b \left[\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial y'} \right] \eta dx .$$

First choose η such that

$$\eta(a) = \eta(b) = 0 ,$$

so we are left with considering

$$\int_a^b \left[\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial y'} \right] \eta dx = 0 .$$

The Fundamental Lemma 2.3 gives us Euler's DE (cf. (2.3))

$$\left[\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial y'} \right] = 0 .$$

Now consider the situation of (3.5) and choose arbitrary η :

$$0 = \eta(b) \frac{\partial f}{\partial y'}(b, u(b), u'(b)) - \eta(a) \frac{\partial f}{\partial y'}(a, u(a), u'(a)) .$$

Now choose η such that $\eta(b) = 0$, $\eta(a) \neq 0$ (analogously $\eta(a) = 0$, $\eta(b) \neq 0$). Thus for an extremum of (1.1) without boundary conditions we have the necessary **natural boundary conditions**:

$$(3.6) \quad \begin{aligned} \frac{\partial f}{\partial y'}(a, u(a), u'(a)) &= 0 \\ \frac{\partial f}{\partial y'}(b, u(b), u'(b)) &= 0 \end{aligned}$$

(proceed analogously if, e.g., $y(a) = y_a$, $y(b) = y_b$ are not given).

3.2 Variational problems in parametric representation

We seek a curve

$$\left\{ \mathbf{x}(t) = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} \mid t \in [\alpha, \beta] \right\}$$

going through the points $(a, y_a)^T$ and $(b, y_b)^T$ and which satisfies

$$(3.7) \quad I(x, y) = \int_{\alpha}^{\beta} f \left(x(t), y(t), \frac{dx(t)}{dt}, \frac{dy(t)}{dt} \right) dt \stackrel{!}{=} \min$$

with $a = x(\alpha)$, $b = x(\beta)$. Note that we could also pose the analog problem without boundary conditions and then find natural boundary conditions.

We assume that the derivatives $\frac{dx(t)}{dt}$ and $\frac{dy(t)}{dt}$ exist, are continuous and are not both equal

to zero in any point. This formulation of the problem only makes sense if $I(x, y)$ is independent of the choice of parametric representation. A necessary and sufficient condition for this is:

$$f \text{ is positive homogenous of order 1 in } \frac{dx}{dt}, \frac{dy}{dt}.$$

For arbitrary $k > 0$ there holds

$$f\left(x, y, k \frac{dx}{dt}, k \frac{dy}{dt}\right) = k \cdot f\left(x, y, \frac{dx}{dt}, \frac{dy}{dt}\right).$$

The embedding theorem leads to the Euler DE system

$$(3.8) \quad \begin{aligned} \frac{d}{dt} \frac{\partial f}{\partial x'}(x, y, x', y') &= \frac{\partial f}{\partial x}(x, y, x', y') \\ \frac{d}{dt} \frac{\partial f}{\partial y'}(x, y, x', y') &= \frac{\partial f}{\partial y}(x, y, x', y') \end{aligned},$$

where x' stands for $\frac{dx}{dt}$ and y' for $\frac{dy}{dt}$.

3.3 Isoperimetric problems

$$(3.9) \quad I[y] = \int_a^b f(x, y, y') dx \stackrel{!}{=} \min, \quad y(a) = y_a, \quad y(b) = y_b$$

with the constraint (cf. Dido's problem)

$$I[y] = \int_a^b g(x, y, y') dx = \gamma$$

(cf. max, min with constraint, Lagrange function resp. Lagrange multiplier). We obtain the Euler equation by starting with

$$\tilde{I}[y] = \int_a^b [f(x, y, y') + \lambda g(x, y, y')] dx \stackrel{!}{=} \min$$

with $y(a) = y_a, y(b) = y_b$ (problem of the form (1.1)).

Analogously for two constraints, e.g.

$$\begin{aligned} \int_a^b g_1(x, y, y') dx &= \gamma_1, \\ \int_a^b g_2(x, y, y') dx &= \gamma_2. \end{aligned}$$

λ is determined using $I[y] = \gamma$.

Example 3.6**Dido's problem**

($I[y] \rightarrow -I[y]$ because of minimum)

$$I[y] = - \int_a^b y \, dx \stackrel{!}{=} \min \quad , \quad y(a) = y(b) = 0$$

with the constraint

$$I[y] = \int_a^b \sqrt{1 + y'^2} \, dx = l .$$

We have $F = f + \lambda g = -y + \lambda \sqrt{1 + y'^2}$ and therefore Euler's DE becomes

$$\frac{\partial F}{\partial y} = -1 = \lambda \frac{d}{dx} \frac{y'}{\sqrt{1 + y'^2}} = \frac{d}{dx} \frac{\partial F}{\partial y'}$$

By integrating we obtain

$$-\lambda \frac{y'}{\sqrt{1 + y'^2}} = x - c_1$$

where c_1 is an integration constant.

Set $y' = \tan \psi$ ($-\frac{\pi}{2} < \psi < \frac{\pi}{2}$) with the supplementary variable ψ .

$$\Rightarrow \quad x - c_1 = -\lambda \sin \psi .$$

Then

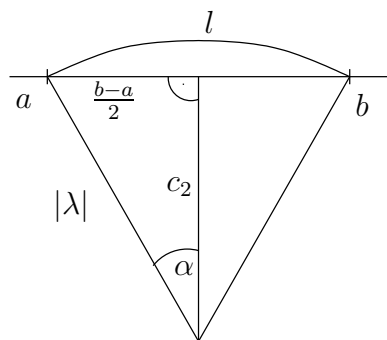
$$\begin{aligned} dy &= y' dx = \tan \psi \, dx = \tan \psi (-\lambda \cos \psi \, d\psi) = -\lambda \sin \psi \, d\psi \\ \Rightarrow \quad \frac{dy}{d\psi} &= -\lambda \sin \psi \\ \Rightarrow \quad y - c_2 &= \lambda \cos \psi . \end{aligned}$$

With $\sin^2 \psi + \cos^2 \psi = 1$ we now obtain $(x - c_1)^2 + (y - c_2)^2 = \lambda^2$, and the constraint $y(a) = y(b) = 0$ finally yields

$$(a - c_1)^2 + c_2^2 = \lambda^2 \quad , \quad (b - c_1)^2 + c_2^2 = \lambda^2 ,$$

therefore

$$c_1 = \frac{a + b}{2} \quad , \quad \left(\frac{b - a}{2} \right)^2 + c_2^2 = \lambda^2 .$$



$$l = 2|\lambda|\alpha \quad , \quad |\lambda| \sin \alpha = \frac{b - a}{2} .$$

Eliminate $|\lambda|$:

$$\frac{\sin \alpha}{\alpha} = \frac{b - a}{l} = \rho .$$

The acute angle α exists for $\frac{2}{\pi} < \rho < 1$ because of $b - a < l$.

With $\rho(\varepsilon) = I[u + \varepsilon\eta]$ we obtain

$$0 \stackrel{!}{=} \left. \frac{d\rho}{d\varepsilon} \right|_{\varepsilon=0} = \int_a^b \left[\frac{\partial f}{\partial y} \eta + \frac{\partial f}{\partial y'} \eta' + \dots + \frac{\partial f}{\partial y^{(n)}} \eta^{(n)} \right] dx .$$

Integrate the second addend once partially, the third addend twice partially and so on:

$$0 = \int_a^b \eta \left[\frac{\partial f}{\partial y} - \frac{d}{dx} \frac{\partial f}{\partial y'} + \frac{d^2}{dx^2} \frac{\partial f}{\partial y''} - \dots + (-1)^n \frac{d^n}{dx^n} \frac{\partial f}{\partial y^{(n)}} \right] dx + R(a, b) .$$

The integrated boundary term $R(a, b)$ disappears because of (3.13).

Variate the Fundamental Lemma:

$$\eta = \begin{cases} (x - \alpha)^{2n}(\beta - x)^{2n} & , \quad x \in \mathcal{U}(\xi) \\ 0 & , \quad \text{sonst} \end{cases} .$$

We then obtain Euler's DE of $2n - th$ order for the solution $u(x)$:

$$(3.14) \quad \begin{aligned} 0 &= \frac{\partial f}{\partial y}(x, u, u', \dots, u^{(n)}) - \frac{d}{dx} \frac{\partial f}{\partial y'}(x, u, u', \dots, u^{(n)}) + \dots \\ &\dots + (-1)^n \frac{d^n}{dx^n} \frac{\partial f}{\partial y^{(n)}}(x, u, u', \dots, u^{(n)}) . \end{aligned}$$

Example 3.7

Now consider

$$\int_0^1 [y^{(n)}]^2 dx \stackrel{!}{=} \min$$

with boundary conditions

$$y^{(r)}(0) = \alpha_r , \quad y^{(r)}(1) = \beta_r \quad (r = 0, \dots, n - 1) .$$

Equation (3.14) yields

$$(-1)^n \frac{d^n}{dx^n} 2u^{(n)} = 0 \quad \text{d.h.} \quad u^{(2n)} = 0 .$$

The coefficients of the solution $u(x) = c_0 + c_1x + \dots + c_{2n-1}x^{2n-1}$ can be uniquely determined from the boundary conditions.

3.7 Weighted variational problems

In this generalized case of (3.12), we now consider the problem

$$(3.15) \quad \begin{aligned} I[y] &= \int_a^b f(x, y, \dots, y^{(n)}) dx + G(y(a), \dots, y^{(n-1)}(a)) \\ &\quad - H(y(b), \dots, y^{(n-1)}(b)) \stackrel{!}{=} \min \end{aligned}$$

besides possible boundary conditions at $x = a$ and $x = b$ with boundary terms up to and including $(n - 1)$ -th order. Compare the following considerations with the case of natural boundary conditions.

We take an embedding approach and first choose a specific η . Once again, condition (3.14) is necessary, and for arbitrary η

$$\begin{aligned}
 R(a, b) &= \frac{\partial f}{\partial y'} \eta + \left[\frac{\partial f}{\partial y''} \eta' - \frac{d}{dx} \left(\frac{\partial f}{\partial y''} \right) \eta \right] \\
 &+ \dots \\
 &+ \left[\frac{\partial f}{\partial y^{(n)}} \eta^{(n-1)} - \dots + (-1)^{n-1} \frac{d^{n-1}}{dx^{n-1}} \left(\frac{\partial f}{\partial y^{(n)}} \right) \eta \right] \Big|_a^b \\
 &+ \left[\frac{\partial G}{\partial y(a)} \eta(a) + \frac{\partial G}{\partial y'(a)} \eta'(a) + \dots + \frac{\partial G}{\partial y^{(n-1)}(a)} \eta^{(n-1)}(a) \right] \\
 &- \left[\frac{\partial H}{\partial y(b)} \eta(b) + \frac{\partial H}{\partial y'(b)} \eta'(b) + \dots + \frac{\partial H}{\partial y^{(n-1)}(b)} \eta^{(n-1)}(b) \right] \\
 &\stackrel{!}{=} 0.
 \end{aligned}$$

Weighted variational problems are of interest in connection with Ritz methods for solving boundary value problems.

Example 3.8

2nd order differential equation with boundary conditions

$$(3.16) \quad -u'' + h(x)u = r(x) \quad , \quad u(0) = 0 \quad , \quad u'(1) = 1 .$$

We want to pose a variational problem whose Euler boundary value problem is (3.16):

$$I[y] = \frac{1}{2} \int_0^1 (y'^2 + h(x)y^2 - 2yr(x)) dx - y(1) \stackrel{!}{=} \min \quad , \quad y(0) = 0,$$

where $y(1)$ is a weight term.

Euler:

$$h(x)u - r(x) = \frac{d}{dx} u' = u''$$

\Rightarrow differential equation.

Embedding: $\eta(0) = 0$.

$$R(a, b) = \underbrace{\frac{\partial f}{\partial y'}}_{y'} \eta \Big|_0^1 - \underbrace{1}_{\frac{\partial H}{\partial y(1)}} \eta(1) \stackrel{\eta(0)=0}{=} (y'(1) - 1)\eta(1) \stackrel{!}{=} 0$$

η can be chosen such that $\eta(1) \neq 0$.

$\Rightarrow y'(1) = 1$ is the natural boundary condition.

The above problem without the weight term would have $y'(1) = 0$ as natural boundary condition.

3.8 The Ritz method revisited

The Ritz method supplies approximate solutions of boundary value problems transformed into variational problems (see above example). Here we consider only the simplest variational problem (1.1) without resp. with partly stipulated boundary conditions

$$y(a) = y_a \quad , \quad y(b) = y_b$$

and with possible weight terms.

Let ρ_0 be continuously differentiable and satisfy the given boundary conditions, and let $\rho_1, \dots, \rho_n \in C^1$ satisfy the homogenous boundary conditions, i.e. if $y(a) = y_a$ is given, then there holds $\rho_1(a) = \dots = \rho_n(a) = 0$ and so on.

Insert the ansatz

$$y_n(x) = \rho_0(x) + \sum_{r=1}^n a_r \rho_r(x)$$

into (1.1) and und define a_1, \dots, a_n such that

$$I[y_n] \stackrel{!}{=} \min_{a_1, \dots, a_n} .$$

This is an extremum problem without constraints, which means there must hold

$$\frac{\partial I}{\partial a_r} \stackrel{!}{=} 0 .$$

Example 3.9

Consider the boundary value problem

$$I[y] = \frac{1}{2} \int_0^1 (y'^2 + xy^2 - 2y) dx \stackrel{!}{=} \min \quad , \quad y(0) = 0 \quad , \quad y(1) = 1 .$$

Interpolating with polynomials

$$\begin{aligned} \rho_0(x) &= x \\ \rho_r(x) &= x^r(1-x) \quad (r = 1, \dots, n) \end{aligned}$$

results in the n -dimensional Ritz ansatz

$$y_n(x) = x + a_1 x(1-x) + \dots + a_n x^n(1-x) .$$

For $n = 2$:

$$\frac{\partial I}{\partial a_1} = \frac{1}{2} \int_0^1 \left[2y_2' \underbrace{(1-2x)}_{\frac{\partial y_2}{\partial a_1}} + 2xy_2x(1-x) - 2x(1-x) \right] dx \stackrel{!}{=} 0 ,$$

and analogously for $\frac{\partial I}{\partial a_2}$. This yields a 2×2 linear equation system for a_1, a_2 .

$$\begin{aligned} 147a_1 + 74a_2 &= 49 & \Rightarrow & a_1 = 0.420202 \\ 148a_1 + 117a_2 &= 42 & & a_2 = -0.172563 \end{aligned}$$

This is a pretty good approximation. The Ritz method forms the basis of finite element methods!

3.9 Variational problems for two independent variables

Let $G \subset \mathbb{R}^2$ be a domain in the (x, y) -plane and Γ its boundary. We consider

$$I[u] = \iint_G f\left(x, y, z, \frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right) dG = 0$$

with $z|_{\Gamma} = \rho(s)$ with respect to the arc length.

A corresponding embedding theorem leads to the partial differential equation

$$\frac{\partial f}{\partial z} - \frac{\partial}{\partial x} \underbrace{\frac{\partial f}{\partial \left(\frac{\partial z}{\partial x}\right)}}_{\text{nach 4.Var.}} - \frac{\partial}{\partial y} \underbrace{\frac{\partial f}{\partial \left(\frac{\partial z}{\partial y}\right)}}_{\text{nach 5.Var.}} = 0.$$

We get analog results if we omit the boundary conditions or add weight terms. With arc length s :

$$\int_{\Gamma} = \psi\left(s, z|_{\Gamma}, \frac{dz}{ds}|_{\Gamma}\right) ds.$$

Remark 3.10 (Final remark)

Calculus of variations is in fact an extremum problem, only not in \mathbb{R}^n but in function spaces in which "derivatives" (Fréchet derivatives) are defined. Euler's DE corresponds to finding the zeros of the Fréchet derivative (resp. $f'(x) = 0$ or $\text{grad } f \stackrel{!}{=} 0$). Several Legendre conditions correspond to $f''(x) > 0$, as also the Lagrange multipliers in (iii) and (iv).