# Numerical Mathematics II
# SS 2019

Lecture by Konstantin Fackeldey

July 8, 2019

## Contents

# I Basic Facts on Ordinary Differential Equations

**Definition I.1.** An ODE of first order in some interval $I \subset \mathbb{R}$ is an equation of the form

$$y'(t) = f(t, y(t)),\ t \in I$$

where $y : I \to \mathbb{C}^n,\ y \in C^1(I)$ and $f : I \times \mathbb{C}^n \to \mathbb{C}^n$. The order is the highest derivative in the ODE. We call an ODE **explicit** if we can solve it for $y'$ and **implicit** otherwise.

**Definition I.2.** An ordinary differential equation of order $n$ is given as

$$y^{(n)}(t) = f(t, y(t), y'(t), \dots, y^{(n-1)}(t))$$

for $t \in I \subset \mathbb{R}$ where $y$ is a $n$-times differentiable function on $I$ and $f : I \times (\mathbb{C}^n)^n \to \mathbb{C}^n$ is a function.

A solution $y$ of an ODE on some $J \subset I$ is a (multiple) continuously differentiable function $y : I \to \mathbb{C}^n$ which solves the ODE

**Remark.** An ODE of order $n$ can be transferred to an ODE of first order by transformation.

**Definition I.3.** We call an ODE

$$y^{(n)}(t) = f(t, y(t), y'(t), \dots, y^{(n-1)}(t)), t \in I$$

an **initial value problem (IVP)** for $y$ if we have additionally the constraints $y(t_0) = y_0, \dots, y^{(n-1)}(t_0) = y_{n-1}$ for $t_0 \in I$.

**Remark.** An ODE has a swarm of solutions, IVP has specific solutions. The swarm of solutions with all constraints is called general solution.

**Theorem I.4** (Picard-Lindelöf)**.** For $t_0 \in \mathbb{R},\ y_0 \in \mathbb{R}^n,\ a, b > 0$ we set

$$I = [t_0 - a, t_0 + a] \text{ and } Q = \{z \in \mathbb{C}^n \mid \|z - y_0\|_\infty \le b\}.$$

Let furthermore $F : I \times Q \to \mathbb{C}^n$ be continuous, with bounded components by some constant $R$ and Lipschitz-continuous in the second argument, i.e.

$$\left|F_j(t, u) - F(t, v)\right| \le L \sum_{k=1}^{n} |u_k - v_k|, \quad j = 1, \dots, n\ , \quad t \in I, \quad u, v \in Q.$$

Then the IVP $y'(t) = F(t, y(t)),\ y(b) = y_0$ has on $J = [t_0 - \alpha, t_0 + \alpha] \subset I$ with $\alpha = \min\{a, \frac{b}{R}\}$ exactly one differentiable solution.

*Proof.* No Proof. □

**Remark.** The existence is local around $t_0$.

**Definition I.5.** The system $y'(t) = A(t)y(t) + f(t)$ for some interval $I \subset \mathbb{R}$ with $A(t) = (a_{ij}(t))_{ij} \in \mathbb{C}^{n,n},\ a_{ij} : I \to \mathbb{C}$ for $i, j \in \{1, \dots, n\},\ n \in \mathbb{N},\ y : I \to \mathbb{C}^n$ and $f : I \to \mathbb{C}^n$ is a linear system of ODEs.

The function $f$ is called inhomogeneity.

The system is called homogeneous if $f = 0$ and inhomogeneous otherwise.

**Theorem I.6.** Let $y_1, y_n$ be two solutions of the homogeneous ODE

$$y'(t) = A(t)y(t),\ t \in I \subset \mathbb{R}.$$

Then each linear combination

$$\alpha y_1 + \beta y_2,\ \alpha, \beta \in \mathbb{C}$$

of $y_1$ and $y_2$ is also a solution

*Proof.*

$$(\alpha y_1(t)+\beta y_2(t))' = \alpha y_1'(t)+\beta y_2'(t) = \alpha A(t)y_1(t)+\beta A(t)y_2(t) = A(t)(\alpha y_1(t)+\beta y_2(t))$$

□

**Remark.** The set of all solutions of homogeneous linear ODEs $y'(t) = A(t)y(t)$ form an $n$-dimensional subspace of the vector space $C^1(I, \mathbb{C}^n)$.

Terms of linear dependence / independence in the context of vector valued functions.

**Definition I.7.** Let $y_1, \dots, y_n$ be $\mathbb{C}^n$ valued functions. We call these functions linear independent on $I \subset \mathbb{R}$ if

$$c_1 y_1(t) + \cdots + c_n y_n(t) = 0 \forall t \in I$$

has the only solution $c_1 = \cdots = c_n = 0$. Otherwise we call them linear independent.

## I.3 Qualitative Behaviour of ODEs

**Example.** Let us consider the $n$-dimensional non contonomous system of first order

$$y'(t) = f(t, y(t))$$
$$y(t_0) = y_0$$

where $f : D \to \mathbb{R}^n,\ D \subset I \times \mathbb{R}^n,\ t_0 \in I,\ I \subset \mathbb{R}$. The questions we are dealing with are:

1. Why only first order?
2. What is the relation between a non-autonomous and an autonomous system?

The reason behind 1. is that any ODE of $n$-th order can be transformed into a $n$-dimensional ODE of first order. Consider the ODE

$$x^{(n)} = F(t, x(t), x'(t), \dots, x^{(n-1)}(t))$$

and define a vector $y$ with its components $y_i,\ i = 1, \dots, n$ by

$$y_i(t) = x^{(i-1)}(t)$$

and a vector field $f(t, y)$ by

$$f(t, y) = \Big(t, y_1, y_2, \dots, y_n, F(t, y_1, y_2, \dots, y_n)\Big)^T$$

Then the ODE of $n$-th order is equivalent to $y'(t) = f(t, y)$.

A system of the form $y'(t) = f(t, y)$ is called an non-autonomous system, a system of the form $y = f(y)$ is called autonomous. We can transform a non-autonomous system to an autonomous system.

Consider the ODE

$$y'(t) = f(t, y) \text{ and } y(t_0) = y_0.$$

We set

$$z = \begin{bmatrix} y \\ s \end{bmatrix} \text{ and } \hat{f} = \begin{bmatrix} f(s, y) \\ 1 \end{bmatrix}, \quad s \in \mathbb{R}$$

Then

$$z'(t) = \hat{f}(z(t)),\ z(t_0) = z_0 = \begin{bmatrix} y_0 \\ t_0 \end{bmatrix}$$

is an autonomous system.

In short, each ODE in $\mathbb{R}^n$ can be transformed to an autonomous ODE in $\mathbb{R}^{n+1}$

**Remark.** In the theorem of Picard-Lindelöf the ODE is of the form $f(t, y)$, where $y$ has to be Lipschitz-continuous.

In the autonomous system the right hand side looks like $f(y(t))$, where $t$ and $y$ have to be Lipschitz-continuous.

**Analytic Continuation**

"Local solutions can be spread onto a maximum time interval."

**Definition I.14** (Local Lipschitz)**.** A function $f : X \to Y$ is local Lipschitz in $x \in X$ if the exists a neighbourhood $U_x \subseteq X$ around $x$ such that $f|_{U_x}$ is Lipschitz-continuous.

For $G := I \times Q$ with $I = [t_0 - a, t_0 + a]$, $Q = \{z \in \mathbb{C} \mid \|z - y_0\| \leq b\}$ with $a, b > 0$ the theorem of Picard-Lindelöf gives for local Lipschitz $f$ the existence of a solution $y_0(t)$ of the IVP

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(t_0) &= y_0 \end{aligned} \tag{1.4}$$

on some (small) inverval $I_0 = [t_0 - a_0, t_0 + a_0]$ with $a_0 = a > 0$.

We will have a look at what happens if we apply the theorem of Picard-Lindelöf on one side of the interval $I_0$. Let now be $t_1 := t_0 + a_0$ and $y_1 = y_0(t_1)$. We then have that $(y_1, t_1) \in G$ and according to Picard-Lindelöf we know that the IVP with $y(t_1) = y_1$ has a unique solution $y_1(t)$ on $I_1 := [t_1 - a_0, t_1 + a_1]$ where $a_1 > 0$.

Due to the uniqueness of the solution if hold $y_0(t) = y_1(t)$ on $I_0 \cap I_1$ we are defining a continuoation of our solution on the greater interval.

It holds

$$y_+(t) = y_0(t) \text{ for } t \in [t_0, t_1]$$

and

$$y_+(t) = y_1(t) \text{ for } t \in (t_1, t_1 + a_1]$$

analogue for $y_-(t)$. Thus there exists a unique solution on the interval $[t_0, t_0 + a_0 + a_1 + \dots]$ if $\sum_{k=0}^{\infty} a_k < \infty$. If $\sum_{k=0}^{\infty} a_k$ diverges, the solution exists globally in forward time.

**Remark.** It can happen that $a_n$ can arbitrary small when $(t_k, y_+(t_k))$ approaches the boundary of $G$. Then either $\|f((t_k), y_+(t_k))\|$ or the Lipschitz-constant $L$ might get arbitrary large.

**Definition I.15.** Let $f : G \to \mathbb{R}^n$ be continuous and local Lipschitz with respect to $y$ and let $(t_0, y_0) \in G$. Let furthermore $t_\pm := t_\pm(t_0, y_0) \in \mathbb{R}$ be defined as

$$\begin{aligned} t_+ &= \sup\{\tau > t_0 \mid \text{there exists a continuation } y_+ \text{ of (1.4) on } [t_0, \tau]\} \\ t_- &= \inf\{\tau > t_0 \mid \text{there exists a continuation } y_- \text{ of (1.4) on } [t_0, \tau]\}. \end{aligned}$$

The interval $(t_-, t_+)$ is the largest interval of existence of the IVP with some initial point $y(t_0) = y_0$.

The maximum solution $y(t)$ is

$$y(t) = \begin{cases} y_+(t) \text{ for } t \in [t_0, t_+) \\ y_-(t) \text{ for } t \in (t_-, t_0]. \end{cases}$$

**Example.** Consider

$$y' = y^2,\ y(t_0) = y(0) = 1,\ y(t) = \frac{1}{1 - t}.$$

Then we have $(t_-, t_+) = (-\infty, 1)$ or $(1, \infty)$.

**Remark.** In case of $t_+ < \infty$ the maximum solution approches for $t \to t_+$, it can then happen that $\|y(t)\|$ is unbounded. This is also called "blow up".

**Solutions and Initial Data**

"What is the influence of a perturbation in $f,\ y_0$ or $t_0$ on the solution?"

To consider this, we need the following Lemma.

**Lemma I.16** (Grönwall-Lemma)**.** Let $I = [a, b] \subseteq \mathbb{R}$ and $g : I \to \mathbb{R}$ be a continuous function. If

$$0 \leq g(t) \leq \delta + \gamma \int_a^t g(x)\ dx$$

holds for all $t \in I,\ \delta, \gamma > 0$, then it holds

$$g(t) \leq \delta e^{\gamma(t-a)}.$$

*Proof.* We set

$$\varphi(t) = \delta + \gamma \int_a^t g(x)\ dx.$$

Then we have

$$\varphi'(t) = \gamma \cdot g(t) \leq \gamma\varphi(t).$$

Since

$$\Big(\varphi \cdot e^{-\gamma t}\Big)' = \varphi' \cdot e^{-\gamma t} + \varphi \cdot (-\gamma)e^{-\gamma t} = e^{-\gamma t}\Big(\varphi'(t) - \gamma\varphi(t)\Big) \leq 0$$

we have that $\varphi e^{-\gamma t}$ is monotone falling. It thus follows

$$g(t) \cdot e^{-\gamma t} \leq \varphi(t) \cdot e^{\gamma t} \leq \varphi(a) \cdot e^{-\gamma a} = \delta \cdot e^{-\gamma a}$$

for all $t \geq a$. □

The Grönwall-Lemma allows us to prove the following theorem.

**Theorem I.17** (Dependence on initial data)**.** Let $D \subset I \times \mathbb{R}^n$ be open, $f : D \to \mathbb{R}^n$ continuous and local Lipschitz with respect to $y$ and $(t_0, y_0) \in D$. If the solution of

$$y'(t) = f(t, y(t))$$
$$y(t_0) = y_0,\ y_0 \in \mathbb{R}^n$$

exists for all $t \in I = [a, b]$ then for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

(i) If $\|y_0 - z_0\| < \delta$ there also exists a solution of

$$z'(t) = f(t, z(t))$$
$$z(t_0) = z_0,\ z_0 \in \mathbb{R}^n$$

for $t \in I$.

(ii) It holds

$$\max_{t \in I} \|y(t) - z(t)\| < \varepsilon.$$

*Proof.* Since $D$ is open, there exists a $\bar{\delta} > 0$ and a compact set

$$K := \{(t, z(t)) \mid t \in I,\ \|y(t) - z(t)\| \leq \bar{\delta}\} \subset D.$$

On $K$ the function $f$ is Lipschitz (with respect to $y$) with a Lipschitz-constant $L$. Let now $\delta < \bar{\delta}$ and $\|y_0 - z_0\| < \delta$. Then for all $t_0, t \in [a, b]$ it holds

$$\|z(t) - y(t)\| \leq \delta + L \int_{t_0}^t \|y(x) - z(x)\|\ dx.$$

This can be seen by the integral representation of $y(t)$. Applying Grönwall's Lemma with $\gamma = L$ yields

$$\|y(t) - z(t)\| \leq \delta \cdot e^{L(t-t_0)} \tag{I.5}$$

and by choosing $\delta \leq \bar{\delta} \cdot e^{L(a-b)}$ it holds $\|y(t) - z(t)\| \leq \bar{\delta}$ for all $t \in I$. Thus it holds $(t, z(t)) \in K$ for $t \in [a, b]$ and hence we have shown *(i)*.

By choosing $\delta < \varepsilon \cdot e^{L(a-b)}$ it follows *(ii)*. □

**Remark.** We have thus shown, that the solution $y(t)$ of the IVP with initial value $y(t_0) = y_0$ depends continuously on the initial data. The solution is often written as $y(t; t_0, y_0, f)$.

**Example.** Let us consider the ODE

$$\begin{aligned} y' &= \lambda y, \ \lambda \in \mathbb{R} \\ y(0) &= y_0 \end{aligned}$$

Here we have $L = |\lambda|$. The equation (I.5) gives

$$|y(t) - z(t)| \leq e^{|\lambda| \cdot t} |y_0 - z_0|.$$

For $\lambda < 0$ we know that $|y(t) - z(t)|$ decreases exponentially.

## I.4 Stability and Flow

### Vector field

A solution of an ODE is a function $y : I \to \mathbb{R}^n$ which is differentiable on $I$. Its graph $\{(t, y(t)) \mid t \in I\}$ is a differentiable curve in $\mathbb{R}^{n+1}$ also known as *solution curve* or *integral curve*. In each point $(t, y(t))$ the direction of the tangent is given by the $(1, f(t, y(t)))$. In other words, $f$ is assigning a direction to each point.

### Stability and small perturbations

Consider

$$y'(t) = f(t, y(t)), \ y(t_0) = y_0.$$

We are now interested in a comparison of different solutions for $t \in [t_0, \infty)$ with respect to the initial condition. We denote the solution by $y(t) = y(t, t_0)$.

Stability means that $y(t_0) = \tilde{y}$ with $\tilde{y}$ near by $y_0$. The question we are dealing with is "How does $y(t, \tilde{y})$ behave in comparison with $y(t, y_0)$?".

Let us consider an autonomous ODE, i.e. an ODE of the form $y'(t) = f(y(t))$.

**Definition I.18** (Equilibrium Point)**.** A point $\bar{y} \in D \subset \mathbb{R}^n$ is called an equilibrium point of a mapping $f : D \to \mathbb{R}^n$ if $f(\bar{y}) = 0$. The constant solution $y(t) = \bar{y}$ is the only solution with $y(t_0) = \bar{y}$.

**Remark.** Other names for equilibrium points are fixed points, equilibria and stationary points.

**Definition I.19** (Stability and asymptotic stability)**.** An equilibrium point is **stable** (in the sense of Ljapunov) if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that for $t \geq t_0$ and for all trajectories $y(t)$ with $\|y(t_0) - \bar{y}\| \leq \delta$ it holds that

$$\|y(t) - \bar{y}\| \leq \varepsilon.$$

An equilibrium point is **instable** if it is not stable.

An equilibrium point $\bar{y}$ is **asymptotic stable** if there exists a neighbourhood $U_{\bar{y}}$ of $\bar{y}$ such that

$$y(t_0) \in U_{\bar{y}} \Rightarrow \lim_{t \to \infty} y(t) = \bar{y}.$$

In this case $\bar{y}$ is called a sink.

An equilibrium point $\bar{y}$ is a spring if for each solution $y(t)$ with $y(t_0) \in U_{\bar{y}}$ and $y(t_0) \neq \bar{y}$ there exists a $t_1 > t_0$ such that $y(t) \notin U_{\bar{y}}$ for all $t \geq t_1$.

**Example.** Consider an ODE in $\mathbb{R}^1$ given by $y'(t) = f(t, y(t))$. The equilibrium point is asymptotic stable if in $U_{\bar{y}}$ it holds that

$$f(y) < 0 \text{ for } y < \bar{y} \quad \text{and} \quad f(y) > 0 \text{ for } y > \bar{y}.$$

**Definition I.20** (Stability of solutions)**.** Let $y(t;y_0)$ be a solution of $y'(t) = f(y(t)),\ y(t_0) = y_0\ \forall t \geq t_0$. Then the solution is **stable** if for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\|y_0 - \tilde{y}_0\| \leq \delta \Rightarrow \|y(t;y_0) - y(t,\tilde{y}_0)\| < \varepsilon$$

for all $t > t_0$. The solution is **attractive** if there exists a $\delta > 0$ such that

$$\|y_0 - \tilde{y}_0\| < \delta \Rightarrow \lim_{t\to\infty} \|y(t;y_0) - y(t,\tilde{y}_0)\| = 0.$$

The solution is **asymptotic stable** if its stable and attractive.

**Flow and Dynamical System**

A Dynamical System is a mathematical model to understand a time independent (autonomous) process. This process shall not depend on the initial time but only on the initial state. Formally, a dynamical system is triple $(T, S, \Phi)$ where $T$ is the time space, $S$ is the state space and $\Phi : T \times S \to S$ is the flow. The time space can either be discrete ($T = \mathbb{N}$) or continuous ($T = \mathbb{R},\ S = \mathbb{R}^n$). This dynamical system is described by an ODE: The entity of all solutions of an ODE is a dynamical system

$$y'(t) = f(y)$$

where $f$ is a differentiable vector field.

**Definition I.21** (Flow of an autonomous ODE)**.** The flow $\Phi(t, y_0)$ or $\Phi_t(y_0)$ of an autonomous ODE

$$y'(t) = f(y(t)),\ y(t_0) = y_0$$

is a mapping $\Phi : \mathbb{R}^{n+1} \to \mathbb{R}^n,\ \Phi(t, y_0) = y(t)$ and with the following properties:

(i) $\Phi(t_0, y_0) = y_0$ for all $y_0 \in \mathbb{R}^n$

(ii) $\Phi(t_1 + t_2, \cdot) = \Phi\Big(t_2, \Phi(t_1, \cdot)\Big)$ for $t_1, t_2 \in \mathbb{R}$.

**Remark.**

- $\Phi(t, y_0)$ is the solution of the ODE $y'(t) = f(y(t))$ which starts in $y_0$ at $t_0$.
- $\Phi : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is differentiable, i.e. $\Phi(t, y_0)$ is a $C^1$/function and it holds

$$\frac{\partial}{\partial t}\Phi(t, y_0) = f\Big(\Phi(t, y_0)\Big).$$

**Example.** For the ODE

$$\begin{aligned} y'(t) &= Ay(t) \\ y(t_0) &= y_0 \end{aligned}$$

with $A \in \mathbb{R}^{n,n}$ it holds

$$\Phi(t, y_0) = e^{At} y_0$$

for all $t \in \mathbb{R}$.

**Lemma I.22.** Under the assumptions of the theorem of Picard-Lindelöf on the ODE

$$y'(t) = f(y(t))$$

the solutions $y_1$ and $y_2$ of different initial conditions do not intersect.

*Proof.* Let us assume towards a contradiction that we have two solutions $\Phi(t_1, y_1)$ and $\Phi(t_2, y_2)$ with different initial conditions which intersect at $y^*$, i.e.

$$\Phi(t_1, y_1) = \Phi(t_2, y_2) = y^*.$$

We define

$$v(t) := \Phi(t + t_1, y_1) = \Phi\big(t, \Phi(t_1, y_1)\big) = \Phi(t, y^*)$$

and

$$w(t) := \Phi(t + t_2, y_2) = \Phi\big(t, \Phi(t_2, y_2)\big) = \Phi(t, y^*).$$

Then by the theorem of Picard-Lindelöf it follows that

$$v(t) = w(t)$$

what ends the proof. □

By

$$\mathcal{O}(y_0) := \{y \in \mathbb{R}^n \mid \exists t \in \mathbb{R} : \ y = \Phi(t, y_0)\}$$

we denote the image of the mapping $t \to \Phi(t, y_0)$. The set $\mathcal{O}(y_0)$ is called **trajectory** or **orbit**.

**Example** (Predator-Prey-Model, Räuber-Beute-Modell)**.** Let $x$ represent the number of prey (maybe a goat) and $y$ the number of the predators (maybe a wolf). We can model

$$\begin{aligned} x' &= x(a - by) \\ y' &= y(-c + dx) \end{aligned} \tag{I.6}$$

where $a, b, c, d \in \mathbb{R}_{>0}$. In the absence of preadotrsthe number of prey is growing exponentially. An increase in the number of predators means a decrease in the number of preys. Note that the decrease of the preys is propotional to $x \cdot y$. In the absence of preys, the predators die. An increase in the number of preys means an increase in the number of predators.

Also note that we assume that the wolf only eats goats and that no further enemies of the goat exist.

These equations belong to the Lotha-Volterra equations.

The origin $(0, 0)$ is the only equilibrium point on the boundary of the state space $\mathbb{R}^2_{\geq 0}$. In the interior of $\mathbb{R}^2_{\geq 0}$ there exists also only one equilibrium point which is given by $(\bar{x}, \bar{y}) = (\frac{c}{d}, \frac{a}{b})$.

The curves of the solutions are closed. To see this, reconsider (I.6). Using simple calculations we get

$$x'\Big(\frac{c}{x} - d\Big) = (a - by)(c - dx)$$

and

$$y'\Big(\frac{a}{y} - b\Big) = (-c + dx)(a - by).$$

By adding up, we obtain

$$\Big(\frac{c}{x} - d\Big)x' + \Big(\frac{a}{y} - b\Big)y' = 0$$

or (using the method of *scharf hinsehen*)

$$\frac{\partial}{\partial t}\Big(c\ln(x) - dx + a\ln(y) - by\Big) = 0.$$

Setting

$$B(x) := \bar{x} \cdot \ln(x) - x \quad \text{and} \quad R(y) := \bar{y} \cdot \ln(y) - y$$

it holds for $V(x, y) := dB(x) + bR(y)$ that

$$\frac{\partial}{\partial t} V\big(x(t), y(t)\big) = 0$$

or $V(x, y)$ is constant along the trajectories of the solutions. We see that $V(x, y)$ is a conserved quantity (*Erhaltungsgröße*) taking its maximum in the equilibrium point $(\bar{x}, \bar{y})$. This point is stable, too (Homework).

Let us now consider $V : D \to \mathbb{R},\ D \subseteq \mathbb{R}^n$ such that in $D$ there exists a equilibrium point $\bar{y}$ of the system $y' = f(y)$. Taking the derivative of $V$ along the solution $y(t)$ we obtain

$$V'(y(t)) = \frac{\partial}{\partial t} V(y(t)) = \nabla\Big(V \cdot y'(t)\Big) = \nabla V\big(f(y(t))\big).$$

If $V' \leq 0$, then $V$ is a monotone falling function along all solutions $y(t) \in D$.

**Theorem I.24** (Ljapunov-Stability)**.** Let $\bar{y} \in D \subseteq \mathbb{R}^n$ be an equilibrium point of $y' = f(y)$. Let further $V : D \to \mathbb{R}$ be a differentiable function on an open set $D$ and let $V(\bar{y}) = 0$ and $V(y) > 0$ for $y \neq \bar{y}$ and

$$V' = \frac{\partial}{\partial t} V \leq 0 \quad \text{on} \quad D \setminus \{\bar{y}\}.$$

Then the equilibrium point $\bar{y}$ is stable. If we have $V' < 0$ then $\bar{y}$ is asymptotic stable.

*Proof.* No proof. □

**Remark.** The function $V$ from theorem I.24 is called Ljapunov-function.

# II Numerics of ODEs

**Motivation II.1.** In the following we only consider first order ODEs for a bounded interval $[a, b] \subseteq R$ and a given function $f : [a, b] \times \mathbb{R} \to \mathbb{R}$. We seek for a function $y : [a, b] \to \mathbb{R}$ such that [1]

$$y'(t) = f(t, y(t)) \quad \forall t \in [a, b] \tag{II.1}$$

with initial condition

$$y'(a) = \hat{y}. \tag{II.2}$$

We devide the interval $[a, b]$ by

$$a = t_0 < t_1 < \cdots < t_n = b, \quad \Delta t_i = t_{i+1} - t_i.$$

At the beginning we only consider an equidistant mesh, i.e. $\Delta t_i$ is constant. Later we also consider variable meshsizes, since there might exist solutions where variable meshsizes can be helpful. We write

$$\Delta t = \frac{b-a}{n} \text{ and } t_i = t_0 + i \cdot \Delta t.$$

Given a starting value $y_0$ we compute our approximations $y_i$ of the exact solution $y(t_i)$ evaluated at $t_i$.

## II.1 Two different schemes

### Difference method

Replace the tangent of $y$ at $t_i$ by a secant with respect to $t_i$ and $t_{i+1}$, i.e.

$$y'(t_i) = \frac{y(t_{i+1}) - y(t_i)}{\Delta t}.$$

Inserting this into the ODE gives

$$\frac{y(t_{i+1}) - y(t_i)}{\Delta t} \approx f(t, y(t)).$$

This leads to the *explicit Euler-Method*

$$y_{i+1} = y_i + \Delta t \cdot f(t_i, y_i), \quad i = 0, \ldots, n-1.$$

### Integration method

We are using the equation

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(\tau)\, d\tau = \int_{t_i}^{t_{i+1}} f(\tau, y(\tau))\, d\tau.$$

Applying the quadrature rule leads to

$$\int_{t_i}^{t_{i+1}} f(\tau, y(\tau))\, d\tau \approx (t_{i+1} - t_i) \cdot f(t_{i+1}, y(t_{i+1})).$$

The *implicit Euler-Method* follows by that as

$$y_{i+1} = y_i + \Delta t \cdot f(t_{i+1}, y(t_{i+1})), \quad i = 0, \ldots, n-1.$$

[1] We assume in (II.1) that $f$ is sufficiently small, such that all necessery (Taylor-)expansions can be built and we also have uniqueness and existence of a solution for the IVP.

## II.2 One-Step Methods

"For computing $y_{i+1}$ of $y$ we only use the information at $t_i$."

**Definition II.2** (One-Step Method)**.** A method for approximating the IVP (II.1) and (II.2) of the form

$$y_{i+1} = y_i + \Delta t\ \Phi(t_1, y_i, y_{i+1}, \Delta t)$$

with some given starting value $y_0$ at $t_0$ and an incremental function (*Verfahrensfunktion*)

$$\Phi : [a, b] \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$$

is called a **one-step method**. We call it **explicit** if $\Phi$ depends not on $y_{i+1}$ and **implicit** otherwise.

**Example.** For the explicit Euler-Method the incremental function $\Phi$ is

$$\Phi(t_i, y_i, y_{i+1}, \Delta t) = f(t_1, y_i).$$

For the implicit Euler-Method the incremental function $\Phi$ is

$$\Phi(t_i, y_i, y_{i+1}, \Delta t) = f(t_{i+1}, y_{i+1}).$$

Note that in the following we use an abuse of notation: In the explicit case we write $\Phi(t_1, y_1, \Delta t)$.

But how do we measure the quality of our approximation?

**Definition II.3** (local discretization error (consistency))**.** A one-step method is **consistent of order** $p \in \mathbb{N}$ if for an ODE (II.1) with some solution $y$ and the local discretization error

$$\eta(t, \Delta t) = y(t) + \Delta t \cdot \Phi\Big(t, y(t), y(t + \Delta t), \Delta t\Big) - y(t + \Delta t)$$

for $t \in [a, b]$ and $0 \le \Delta t \le b - t$ it holds

$$\eta(t, \Delta t) = O(\Delta t^{p+1}) \ \text{ as } \ \Delta t \to 0.$$

In case of $p = 1$ we say that the method is **consistent**.

**Revision.** The Landau-Notation for functions $f$ and $g$ is defined as follows:
It holds "$f(x) = O(g(x))$ for $x \to a$" if $\left|\frac{f(x)}{g(x)}\right|$ is bounded when $x \to a$. Furthermore it holds "$f(x) = o(g(x))$ for $x \to a$" if $\lim_{x\to a} \frac{f(x)}{g(x)} = 0$. We make use of an abuse of notation by writing the equality sign, since formally $O(g(x))$ and $o(g(x))$ are sets.

**Remark.** For a consistent method it holds

$$\begin{aligned}\lim_{\Delta t\to 0} \Phi\Big(t, y(t), y(t + \Delta t), \Delta t\Big) &= \underbrace{\lim_{\Delta t\to 0} \frac{\eta(t, \Delta t)}{\Delta t}}_{=0} + \lim_{\Delta t\to 0} \frac{y(t + \Delta t) - y(t)}{\Delta t} \\ &= y'(t) = f(t, y(t)).\end{aligned}$$

**Theorem II.3** (Consistence of the explicit Euler-Method)**.** The explicit Euler-Method is consistent of order $p = 1$.

*Proof.* Expansion of $y$ in $t$ gives

$$\begin{aligned}y(t + \Delta t) &= y(t) + y(t)' \cdot \Delta t + \frac{y''(\varrho)}{2}\Delta t^2, \quad \varrho \in [t, t + \Delta t] \\ &= y(t) + f(t, y(t)) \cdot \Delta t + \frac{y''(\varrho)}{2}\Delta t^2.\end{aligned}$$

It thus follows

$$\begin{aligned}\eta(t,\Delta t) &= y(t) - \Delta t \cdot f(t,y(t)) - y(t+\Delta t)\\ &= -\frac{\Delta t^2}{2} y''(\varrho) = O(\Delta t^2)\end{aligned}$$

for $\Delta t \to 0$, since $y''$ is bounded in $[t, t+\Delta t]$. □

**Definition II.4** (Convergence of one-step methods)**.** A one-step method with starting value $y_0 = y(0) + O(\Delta t^p),\ \Delta t \to 0$ is **convergent of order** $p \in \mathbb{N}$ with respect to the IVP (II.1) and (II.2) if for the approximation $y_i$ of the solution $y(t_i)$ the **global approximation error**

$$e(t_i, \Delta t) = y(t_i) - y_i$$

for all $t_i,\ i = 1, \ldots, n$ meets

$$e(t_i, \Delta t) = O(\Delta t^p), \quad \Delta t \to 0$$

In case of $e(t, \Delta t) = O(1)$ we call the method **consistent**.

**Remark.** Note, that in $e(t_i, \Delta t)$ all $\eta(t, \Delta t)$ are summed up.

**Lemma II.5** (technical Lemma)**.** Let $\eta_i, \varrho_i, z_i \in \mathbb{R}_{\geq 0}$ for $i = 0, \ldots, m-1$ and $z_m \in \mathbb{R}$ and it holds

$$z_{i+1} \leq (1 + \varrho_i) z_i + \eta_i$$

for $i = 0, \ldots, m-1$. Then it holds

$$z_{i+1} \leq \left( z_0 + \sum_{k=0}^{i} \eta_k \right) e^{\sum_{k=0}^{i} \varrho_k}$$

for $i = 0, \ldots, m-1$.

*Proof.* We prove the statement by induction on $i$. For $i = 0$ the claim is true. Hence let the statement be valid for a $i-1$. Then we have

$$\begin{aligned} z_{i+1} &\leq (1 + \varrho_i) z_i + \eta_i\\ &\leq \underbrace{(1 + \varrho_i)}_{\leq e^{\delta_i}} \cdot \left( z_0 + \sum_{k=0}^{i-1} \eta_k \right) e^{\sum_{k=0}^{i-1} \varrho_k} + \eta_i\\ &\leq \left( z_0 + \sum_{k=0}^{i-1} \eta_k \right) e^{\sum_{k=0}^{i-1} \varrho_k} + \eta_i\\ &\leq \left( z_0 + \sum_{k=0}^{i} \eta_k \right) \cdot e^{\sum_{k=0}^{i} \varrho_k}, \end{aligned}$$

what ends the proof. □

**Theorem II.6** (Convergence of one-step methods)**.** Let $\Phi$ be an incremental function of a one-step method for the IVP (II.1) and (II.2) with

$$|\Phi(t, u, w, \Delta t) - \Phi(t, v, w, \Delta t)| \leq L|u - v| \tag{II.3}$$

$$|\Phi(t, w, u, \Delta t) - \Phi(t, w, v, \Delta t)| \leq L|u - v| \tag{II.4}$$

with $L \in \mathbb{R}$. Then it holds for $\Delta t < \frac{1}{L}$ that

$$|e(t_{i+1}, \Delta t)| \leq \left( |e(t_0, \Delta t)| + \frac{(t_{i+1} - t_0)}{1 - \Delta t \cdot L} \cdot \frac{\eta(\Delta t)}{\Delta t} \right) e^{2 \cdot \frac{t_{i+1} - t_0}{1 - \Delta t} \cdot L} \tag{II.5}$$

for $i = 0, \ldots, n-1$, where

$$\eta(\Delta t) := \max_{j=0,\ldots,n-1} |\eta(t_j, \Delta t)|.$$

*Proof.* Reconsider that

$$\eta(t_i, \Delta t) = y(t_i) + \Delta t \Phi\left(t_i, y(t_i), y(t_i + \Delta t), \Delta t\right) - y(t_{i+1}).$$

Rearranging gives

$$y(t_{i+1}) = y(t_i) + \Delta t \Phi\left(t_i, y(t_i), y(t_i + \Delta t), \Delta t\right) - \eta(t_i, \Delta t).$$

Consider now

$$\begin{aligned} e(t_{i+1}, \Delta t) &= y(t_{i+1}) - y_{i+1} \\ &= y(t_i) + \Delta t \Phi\left(t_i, y(t_i), y(t_{i+1}), \Delta t\right) - \eta(t_i, \Delta t) \\ &\quad - y_i - \Delta t \Phi\left(t_i, y_i, y_{i+1}, \Delta t\right) \pm \Delta t \Phi\left(t_i, y(t_i), y_{i+1}, \Delta t\right). \end{aligned}$$

Using (II.3) and (II.4) we obtain

$$\begin{aligned} |e(t_{i+1}, \Delta t)| &\leq |e(t_i, \Delta t)| + \Delta t L |y(t_{i+1}) - y_{i+1}| + \Delta t L |y(t_i) - y_i| - \eta(t_i, \Delta t) \\ &= |e(t_i, \Delta t)| + \Delta t L |e(t_{i+1}, \Delta t)| + \Delta t L |e(t_i, \Delta t)| + |\eta(t_i, \Delta t)|. \end{aligned}$$

This gives

$$(1 - \Delta t L)|e(t_{i+1}, \Delta t)| \leq (1 + \Delta t L)|e(t_i, \Delta t)| + |\eta(t_i, \Delta t)|,$$

so

$$|e(t_{i+1}, \Delta t)| \leq \frac{(1 + \Delta t L)}{(1 - \Delta t L)} |e(t_i, \Delta t)| + \frac{1}{(1 - \Delta t L)} |\eta(t_i, \Delta t)| \leq \frac{(1 + \Delta t L)}{(1 - \Delta t L)} |e(t_i, \Delta t)| + \frac{1}{(1 - \Delta t L)} \eta(\Delta t).$$

By setting

$$\begin{aligned} \varrho_i &:= \frac{(1 + \Delta t L)}{(1 - \Delta t L)} - 1 = \frac{2 \Delta t L}{1 - \Delta t L} \geq 0 \\ z_i &:= |e(t_i, \Delta t)| \geq 0 \\ \eta_i &:= \frac{1}{(1 - \Delta t L)} \eta(\Delta t) \geq 0 \end{aligned}$$

and applying Lemma II.5 we obtain

$$\begin{aligned} |e(t_{i+1}, \Delta t)| &= z_{i+1} \\ &\leq \left( z_0 + \sum_{k=0}^{i} \eta_k \right) e^{\sum_{k=0}^{i} \varrho_k} \\ &= \left( |e(t_0, \Delta t)| + \sum_{k=0}^{i} \frac{1}{1 - \Delta t L} \eta(\Delta t) \right) e^{\sum_{k=0}^{i} \frac{2 \Delta t L}{1 - \Delta t L}}. \end{aligned} \tag{$*$}$$

Observe that the two sums can be rewritten as

$$\sum_{k=0}^{i} \frac{1}{1 - \Delta t L} \eta(\Delta t) = \frac{i + 1}{1 + \Delta t L} \eta(\Delta t) = \frac{t_{i+1} - t_0}{1 + \Delta t L} \cdot \frac{\eta(\Delta t)}{\Delta t}$$

and

$$\sum_{k=0}^{i} \frac{2 \Delta t L}{1 - \Delta t L} = (t_{i+1} - t_0) \frac{2L}{1 - \Delta t L}.$$

Inserting this into $(*)$ gives the result. □

**Theorem II.7.** If a one-step method with Lipschitz conditions (II.3) and (II.4) is consistent of order $p \in \mathbb{N}$ for an ODE (II.1) and if the initial value $y_0$ meets

$$y_0 = \hat{y}_0 + O(\Delta t^p),$$

then the method is convergent of order $p$ with respect to (II.1) and (II.2).

**Remark.**

- The error grows exponentially in time.
- If in the underlying ODE the Lipschitz-constant $\hat{L}$ given by

  $$|f(t, y_1(t)) - f(t, y_2(t))| \leq \hat{L}|y_1 - y_2|$$

  is large, then $L$ from (II.4) and (II.5) will also be large.
- If the initial condition of the explicit Euler-method meets

  $$y_0 = \hat{y} + O(\Delta t)$$

  then it is convergent of first order with respect to the ODE

  $$y'(t) = f(t, y(t)), y(t_0) = \hat{y}_0.$$

**Runge-Kutta Methods**

We already know that

$$y(t_{i+1}) - y(t_i) \approx \Delta t f(t, y(t)).$$

Asking whether a better approximation leads to better convergence leads to the Runge-Kutta methods. Trying with the midpoint rule gives

$$\int_{t_i}^{t_{i+1}} f(t, y(t))\ dt \approx f\left(t_i + \tfrac{\Delta t}{2},\ y\left(t_i + \tfrac{\Delta t}{2}\right)\right) \cdot \Delta t.$$

But we have not evaluated $y(t_i, \frac{\Delta t}{2})$, what is a problem. The idea is using the explicit Euler-method to approximate $y(t_i + \frac{\Delta t}{2})$. Define

$$y_{i+\frac{1}{2}} := y_i + \frac{\Delta t}{2} f(t_i, y(t_i)).$$

By plugging in we obtain

$$y_{i+1} = y_i + \Delta t \cdot f\left(t_i + \tfrac{\Delta t}{2}, y_i + \tfrac{\Delta t}{2} f(t_i, y(t_i))\right)$$

which is also called the explicit midpoint rule.

**Excursion to quadrature**

We know, that

$$I : C \to \mathbb{R},\ f \mapsto \int_a^b f(\tau)\ d\tau$$

is a linear functional from some function space $C$ into the real numbers.

**Excursion Definition 1.** A function $Q_{n+1} \in C([a,b])$ with

$$Q_{n+1}(f) = \sum_{i=0}^{n} a_i f(x_i)$$

with nodes $x_i \in [a,b]$ and weights $a_i \in \mathbb{R}$ is called a quadrature rule. Its quadrature error is the linear functional

$$R_{n+1}(f) = I(f) - Q_{n+1}(f).$$

The rule converges, if it holds

$$\lim_{n\to\infty} Q_{n+1}(f) = I(f).$$

There exist many quadrature rules. Here we consider **quadrature by interpolation**. Let us assume, that we know $f$ only at $(n+1)$ points $x_0, \ldots, x_n$ an we interpolate $f$ by a polynomial $p$ with degree $n$.

**Excursion Definition 2** (Quadrature by interpolation)**.** Let $p_n$ be a polynomial of degree $n$ on the interval $[a,b]$. We call $R_{n+1}$ a **quadrature rule by interpolation** if

$$R_{n+1}(p_n) = 0,$$

i.e. if a polynomial of degree $n$ can be integrated exactly.

Consider $\xi_j = t_i + c_j\Delta t,\ c_j \in [0,1]$ for $j = 1,\dots,s$. Then we have

$$y(t_{i+1}) - y(t_i) = \int_{t_i}^{t_{i+1}} f(\tau, y(\tau))\ d\tau \approx \Delta t \sum_{j=1}^{s} b_j f(\xi_j, y(\xi_j)).$$

From the quadrature by interpolation we know

$$\sum_{j=1}^{s} b_j = 1.$$

But since we do not know the values $y(\xi_j)$, we have to think about how to get these values.

Applying the fundamental theorem gives

$$\begin{aligned} y(\xi_j) - y(t_i) &= \int_{t_i}^{t_i + c_j\Delta t} f(t, y(t))\ dt \\ &\approx c_j\Delta t \cdot \sum_{\nu=1}^{s} \tilde{a}_{j\nu} f(\xi_\nu, y(\xi_\nu)). \end{aligned}$$

This seems strange, since we are using the same $\xi_i$. Setting $a_{j\nu} := c_j\tilde{a}_{j\nu}$ we obtain

$$k_i = y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} f(\xi_\nu, y(\xi_\nu))$$

as an approximation of $y(\xi_i)$, where $i = 1,\dots,s$. These methods are called the *Runge-Kutta methods*.

**Definition II.8** (Runge-Kutta-Method/RKM)**.** For $b_j, c_j, a_{j\nu} \in \mathbb{R},\ j = 1,\dots,s$ we denote

$$k_j = y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} f(\xi_\nu, k_\nu)$$

for $j = 1,\dots,s$ and

$$y_{i+1} = y_i + \Delta t \sum_{j=1}^{s} b_j f(\xi_j, k_j)$$

with $\xi_j = t_i + c_j\Delta t$ as an **$s$-step Runge-Kutta method**. We call $c_j$ and $b_j$ **weights**.

**Remark.**

- A Runge-Kutta methods is defined by the parameters $a_{j\nu}, b_j, c_j \in \mathbb{R}$.
- The *Butcher table* or *Array* of a Runge-Kutta method can be denoted as

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array}.$$

**Example.**

(1) For the explicit Euler method, the Butcher table is given by $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$, which is equivalent to $k_1 = y_i$ and $y_{i+1} = y_i + \Delta t f(t_i, k_1)$.

(2) For the implicit Euler method, the Butcher table is given by $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$, which is equivalent to $y_{i+1} = k_1 = y_i + \Delta t f(t_i, k_1)$.

(3) For the explicit midpoint rule, the Butcher tabel is given by

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & 0 & 1 \end{array}$$

which is equivalent to

$$\begin{aligned} k_1 &= y_i \\ k_2 &= y_i + \frac{\Delta t}{2} f(t_1, k_1) \\ y_{i+1} &= y_i + \Delta t f\left(t_i + \tfrac{\Delta t}{2}, k_2\right). \end{aligned}$$

The Runge-Kutta method can also be seen from a predictor-corrector method-point of view. For that, consider the trapezoid rule given by

$$y(t_{i+1}) - y(t_i) \approx \frac{\Delta t}{2}\Big(f(t_i, y(t_i)) + f(t_{i+1}, y(t_{i+1}))\Big).$$

We obtain

$$y_{i+1} = y_i + \frac{\Delta t}{2}\Big(f(t_i, y_i) + f(t_{i+1}, y_{i+1})\Big).$$

Since we do not know $y_{i+1}$, we approximate $f(t_{i+1}, y_{i+1})$ by $f(t_{i+1}, k_2)$, where

$$k_2 = y_i + \Delta t f(t_i, y_i)$$

is derived from the explicit Euler method. We thus have

$$\begin{aligned} k_1 &= y_i \\ k_2 &= y_i + \Delta t f(t_i, y_i) \\ y_{i+1} &= y_i + \frac{\Delta t}{2}\Big(f(t_1, k_1) + f(t_{i+1}, k_2)\Big). \end{aligned}$$

In this case, $k_2$ is called a predictor and since we have

$$y_{i+1} = k_2 + \frac{\Delta t}{2}\Big(f(t_i, k_2) - f(t_i, k_1)\Big),$$

the term

$$\frac{\Delta t}{2}\Big(f(t_i, k_2) - f(t_i, k_1)\Big)$$

is called the corrector.

Instead of computing $k_j$ at $y(\xi_j)$ we can use the slopes or gradients

$$r_j = f(t_i + c_j \Delta t, k_j).$$

Within the predictor-corrector method we have

$$\begin{aligned} r_1 &= f(t_i, y_i) \\ r_2 &= f\left(t_i + \Delta t, y_i + \Delta t r_1\right) \\ y_{i+1} &= y_i + \frac{\Delta t}{2}(r_1 + r_2), \end{aligned}$$

where $r_1 + r_2$ can be seen as an intermediate slope. By setting $r_j = f(t_i + c_j \Delta t, k_j)$ we obtain a Runge-Kutta method

$$\begin{aligned} r_j &= f(t_i + c_j \Delta t, k_j) \\ &= f\left(t_i + c_j \Delta t, y_i + \Delta t \cdot \sum_{\nu=1}^{s} a_{j\nu} f(\xi_\nu, k_\nu)\right) \\ &= f\left(t_i + c_j \Delta t, y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} r_\nu\right). \end{aligned}$$

By summing up, we can write

$$y_{i+1} = y_i + \Delta t \sum_{j=1}^{s} b_j r_j.$$

When computing for example $r_3$, it can happen that we end up with the form

$$r_3 = f\left(\ldots, \sum_{\nu=1}^{s} a_{j\nu} r_\nu\right),$$

where $r_3$ depends on $r_3$. Analogue to the one-step methods, in this case we call the Runge-Kutta method implicit. Let us assume, that $A = [a_{j\nu}] \in \mathbb{R}^{s,s}$ is a strict lower triangle matrix, i.e. $a_{j\nu} = 0$ for $\nu \geq j$. Then we obtain

$$r_j = f\left(t_i + c_\nu \Delta t, y_i + \Delta t \cdot \sum_{\nu=1}^{j-1} a_{j\nu} r_\nu\right)$$

for $j = 1, \ldots, s$ and hence we have an **explicit Runge-Kutta method**. If we don't have such a matrix $A$, we have an **implicit Runge-Kutta method**.

Note, that unlike to one-step methods, in Runge-Kutta methods explicit and implicit only refers to the intermediate steps between $t_i$ and $t_{i+1}$.

Let us assume, that we have a full matrix A and $f : [a, b] \times \mathbb{R}^m \to \mathbb{R}^m$. Then

$$\begin{aligned} r_1 &= f\left(t_i + c_1 \Delta t, y_i + \Delta t \cdot \sum_{\nu=1}^{s} a_{1\nu} r_\nu\right) \\ &\ldots \\ r_s &= f\left(t_i + c_s \Delta t, y_i + \Delta t \cdot \sum_{\nu=1}^{s} a_{s\nu} r_\nu\right) \end{aligned} \tag{II.$*$}$$

is a system of dimension $s - m$ for computing the gradients $r_j \in \mathbb{R}^m$. It might be linear or non-linear, depending on the underlying system.

**Example** (Classical Runge-Kutta method)**.** Let the Butcher table be given by

$$\begin{array}{c|cccc} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}.$$

This Butcher table gives an explicit Runge-Kutte method, since $A$ is a strict lower triangular matrix. We thus have

$$y_{i+1} = y_i + \Delta t\left(\tfrac{1}{6} r_1 + \tfrac{1}{3} r_2 + \tfrac{1}{3} r_3 + \tfrac{1}{6} r_4\right)$$

and further

$$\begin{aligned} r_1 &= f(t_i + 0 \cdot \Delta t, y_i + \Delta t \cdot 0) = f(t_i, y_i) \\ r_2 &= f\left(t_i + \frac{\Delta t}{2}, y_i + \frac{\Delta t}{2} r_1\right) \\ r_3 &= f\left(t_i + \frac{\Delta t}{2}, y_i + \frac{\Delta t}{2} r_2\right) \\ r_4 &= f(t_1 + \Delta t, y_1 + \Delta t r_3). \end{aligned}$$

The drawing is missing. A Runge-Kutta method hence allows us to approximate $f(t, y(t))$ on intermediate steps, i.e. estimate $y(t)$ on intermediate steps.

Reconsider the (maybe non-linear) system (II.$*$) for computing the gradients $r_j$. The following theorem shows that the system can be solved if $f$ satisfies certain conditions. In particular, we can, in some sense, buy the solvability of the system (II.$*$) by choosing smaller time steps.

**Theorem II.9.** Let the mapping $f : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ be continuous so that it holds

$$\|f(t,\tilde{y}) - f(t,y)\|_\infty \leq L \|\tilde{y} - y\|_\infty$$

for all $t \in [a,b]$, where $L > 0$ is a Lipschitz constant. Consider the Runge-Kutta method $(A,b,c)$ with $\Delta t < \frac{1}{L\|A\|_\infty}$. Then for any $j = 1,\dots,s$, the iteration given by

$$r_j^{(\ell+1)} = f\left(t_i + c_i\Delta t,\ y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} r_\nu^{(\ell)}\right)$$

converges for $\ell \to \infty$ to an arbitrary initialization $r_1^{(0)},\dots,r_s^{(0)}$ to the unique solution of the system

$$r_j = f\left(t_i + c_j\Delta t,\ y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} r_\nu\right).$$

*Proof.* We set

$$R := \begin{bmatrix} r_1 \\ \vdots \\ r_s \end{bmatrix} \text{ and } F := \begin{bmatrix} F_1 \\ \vdots \\ F_2 \end{bmatrix} : \mathbb{R}^{s\cdot m} \to \mathbb{R}^{s\cdot m}$$

with

$$F_j(R) = f\left(t_i + c_j\Delta t,\ y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} r_\nu\right).$$

We thus have

$$\left\|F(R) - F(\tilde{R})\right\|_\infty \leq L \cdot \left\| \begin{bmatrix} \Delta t \sum_{\nu=1}^s a_{1\nu}(r_\nu - \tilde{r}_\nu) \\ \vdots \\ \Delta t \sum_{\nu=1}^s a_{s\nu}(r_\nu - \tilde{r}_\nu) \end{bmatrix} \right\|_\infty$$

with

$$\left\| \begin{bmatrix} \Delta t \sum_{\nu=1}^s a_{1\nu}(r_\nu - \tilde{r}_\nu) \\ \vdots \\ \Delta t \sum_{\nu=1}^s a_{s\nu}(r_\nu - \tilde{r}_\nu) \end{bmatrix} \right\|_\infty \leq \left\| \begin{bmatrix} \Delta t \sum_{\nu=1}^s a_{1\nu} \\ \vdots \\ \Delta t \sum_{\nu=1}^s a_{s\nu} \end{bmatrix} \right\|_\infty \cdot \left\|R - \tilde{R}\right\|_\infty.$$

we obtain

$$\left\|F(R) - F(\tilde{R})\right\| \leq L \cdot \Delta t \cdot \underbrace{\left(\max_{i=1,\dots,s} \sum_{\nu=1}^{s} |a_{i\nu}| \left\|R - \tilde{R}\right\|_\infty\right)}_{=\|A\|_\infty}$$

which is a contraction in the Banachspace $(\mathbb{R}^{s\cdot m}, \|\cdot\|_\infty)$ if $L \cdot \Delta t \|A\|_\infty < 1$. The Banach-fixpoint theorem then implies that there exists an $R \in \mathbb{R}^{s\cdot m}$ as the fixed point of this iteration. Furthermore, $(R^{(\ell)})_\ell$ with $R^{(\ell+1)} = F(R^{(\ell)})$ converges towards R. □

**Remark.** After Definition II.2 (consistency) we saw, that the minimum requirement for a consistent method is

$$\lim_{\Delta t \to 0} \Phi(t, y(t), y(t+\Delta t), \Delta t) = f(t,y). \tag{II.6}$$

We now are interested in the consistency of arbitrary Runge-Kutta methods. Reconsider

$$\begin{aligned} r_j &= f\left(t_i + c_j\Delta t,\ y(t_i) + \Delta t \sum_{\nu=1}^{s} a_{j\nu} r_\nu\right) \\ &= f(t_i, y(t_i)) + O(\Delta t), \end{aligned}$$

since

$$\begin{aligned} y_{i+1} &= y_i + \Delta t \cdot \Phi(t_i, y(t_i), y(t_i + \Delta t), \Delta t) \\ &= y_i + \Delta t \cdot \sum_{j=1}^{s} b_j r_j. \end{aligned}$$

We further have that

$$\Phi(t_i, y(t_i), y(t_i + \Delta t), \Delta t) = \sum_{j=1}^{s} b_j f(t, y(t_i)) + O(\Delta t)$$

and hence

$$\lim_{\Delta t \to 0} \Phi(t_i, y(t_i), y(t_i + \Delta t), \Delta t) = f(t_i, y(t_i)) \Leftrightarrow \sum_{j=1}^{s} b_j = 1.$$

In summary, we have proven the following result.

**Lemma II.10.** A Runge-Kutta method given by $(A, b, c)$ is consistent in the sense of (II.6) if and only if it holds

$$\sum_{j=1}^{s} b_j = 1.$$

This makes hope for determining the order of consistency of a Runge-Kutta method by looking at its Butcher table. The next theorem gives us precise conditions to that.

**Theorem II.11.** For a Runge-Kutta method $(A, b, c)$ it holds

(a) The method has order of consistency $p \geq 1$ if

$$\sum_{j=1}^{s} b_j = 1 \text{ and } \sum_{\nu=1}^{s} a_{j\nu} = c_j \tag{II.7}$$

holds for all $j = 1, \ldots, s$.

(b) The method has order of consistency $p \geq 2$ if it holds (II.7) and

$$\sum_{j=1}^{s} b_j c_j = \frac{1}{2} \tag{II.8}$$

holds.

(c) The method has order of consistency $p \geq 3$ if (II.7), (II.8) and

$$\sum_{j=1}^{s} b_j c_j^2 = \frac{1}{3} \text{ and } \sum_{j=1}^{s} b_j \sum_{\nu=1}^{s} a_{j\nu} c_\nu = \frac{1}{6}$$

hold.

*Proof.* The proof can be found in *Deuflhard/Bornemann: Numerische Mathematik II, Chapter 4* and is not given in this lecture. □

**Remark.**

- In the case of an explicit Runge Kutta method the conditions in theorem II.11 are equivalent conditions to the consistency.
- The proof of the theorem relies on Taylor expansion.

- In principle, this technique is also working for higher orders. However, for higher orders we get more conditions on the coefficients $(A, b, c)$ of the Runge-Kutta method. The number of conditions $N_p$ increases rapidly when the order increases, since for $p = 15$ there are already 141083 conditions needed.

- It seems to be, that the number of stages (steps) of an explicit Runge-Kutta method gives the order. But it holds

| order $p$ | 1 | 2 | 3 | 4 | 5 | 6 | $\cdots$ | 12 | 14 |
|---|---|---|---|---|---|---|---|---|---|
| $s$ (stages) | 1 | 2 | 3 | 4 | 6 | 7 | $\cdots$ | 25 | 35 |

**Theorem II.12.** For an explicit Runge-Kutta method with $s$ stages and order of consistency $p$ it holds

$$p \leq s.$$

*Proof.* If we can find one IVP such that $p > s$ leads to a contradiction, the claim is proven. Consider

$$y'(t) = y(t),\ y(0) = 1.$$

□

In chapter I we saw that any non-autonomous system

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(t_0) &= y_0 \end{aligned}$$

can be transformed to an autonomous system

$$\begin{aligned} z'(t) &= \hat{f}(z(t)) \\ z'(t_0) &= z_0 = \begin{bmatrix} y_0 \\ t_0 \end{bmatrix} \end{aligned} \tag{II.9}$$

by the transformation

$$\begin{aligned} z &= \begin{bmatrix} y \\ s \end{bmatrix} \\ \hat{f}(z) &= \begin{bmatrix} f(s, y) \\ 1 \end{bmatrix}. \end{aligned}$$

We want to know if the result of an explicit Runge-Kutta method stays the same when we apply the Runge-Kutta method to the autonomous system.

By applying the explicit Runge-Kutta method to the autonomous sytem (II.9), we obtain

$$z_{i+1} = z_i + \Delta t \sum_{j=1}^{s} b_j \hat{r}_j,$$

or more precisely

$$\begin{bmatrix} y_{i+1} \\ t_{i+1} \end{bmatrix} = \begin{bmatrix} y_i \\ t_i \end{bmatrix} + \Delta t \sum_{j=1}^{s} b_j \begin{bmatrix} 1 \\ r_j \end{bmatrix}.$$

We will first just consider the lower part of the last equation, i.e.

$$t_{i+1} = t_i + \Delta t \sum_{j=1}^{s} b_j = t_i + \Delta t,$$

so it is necessery that

$$\sum_{j=1}^{s} b_j = 1.$$

Further, we have that

$$\hat{r}_j = \begin{bmatrix} r_j \\ 1 \end{bmatrix} = \hat{f}\left( z_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} \hat{r}_\nu \right)$$
$$= \hat{f}\left( \begin{bmatrix} y_i \\ t_i \end{bmatrix} + \Delta t \sum_{\nu=1}^{s} a_{j\nu} \begin{bmatrix} r_j \\ 1 \end{bmatrix} \right),$$

such that[2]

$$r_j = f\left( t_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} \cdot 1,\ y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} r_\nu \right),$$

so our second condition is

$$\sum_{\nu=1}^{s} a_{j\nu} = c_j.$$

We summarize our result in the following lemma.

**Lemma II.13.** The explicit Runge-Kutta method $(A, b, c)$ is invariant to making it autonomous if and only if it holds

(i)
$$\sum_{j=1}^{s} b_j = 1$$

(ii)
$$\sum_{\nu=1}^{s} a_{j\nu} = c_j$$

for all $j = 1, \ldots, s$.

We are heading towards the implicit Runge-Kutta methods and try to motivate them.

Since in the implicit case the matrix $A$ has more non-zero entries, we have more conditions to our method. We seek to find out if it is possible to use the additional coefficients of an implicit Runge-Kutta method to obtain a higher order for a given number $s$ of stages. In the explicit case, theorem II.12 shows that for the order of consistency $p$ it holds $p \leq s$, so we want to find out if it is possible that $p > s$ for a implicit Runge-Kutta method.

Our idea is to use collocation method, i.e. choose (simple) functions from some function space (e.g. polynomials $\mathbb{P}$) and a set of collocation points (pairwise different points) and set the free coefficients of the (simple) function such that the problem function (e.g. $f$) holds in the collocation points. Consider the ODE

$$y'(t) = f(t, y(t)),\ y(t_0) = y_0.$$

We assume that the numerical solution has been carried up to the point $(t_i, y_i)$. We now seek for a recipe to advance it to $(t_{i+1}, y_{i+1})$, where $t_{i+1} = t_i + \Delta t$. To do so, we choose $s$ collocation points $c_1, \ldots, c_s \in [0, 1]$ and then seek for a $s$-th degree polynomial $u_s \in \mathbb{P}_s$, such that it holds

$$u_s(t_i) = y_i \tag{II.10}$$
$$u_s'(t_i + c_j \Delta t) = f\Big(t_i + c_j \Delta t,\ u_s(t_i + c_j \Delta t)\Big)$$

for $j = 1, \ldots, s$.

A **collocation method** consists of finding such a $u_s$ and setting $y_{i+1} = u(t_{i+1})$.

But we do not yet know what the relation to a Runge-Kutta method is.

[2]Remember that in the non-autonomous case the structure of $r_j$ is

$$r_j = f\left( t_i + c_j \Delta t,\ y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} r_\nu \right).$$

**Lemma II.14.** The collocation method defined by (II.10) is equivalent to a $s$ stage implicit Runge-Kutta method with the coefficients

$$a_{ji} = \int_0^{c_j} L_i(\tau)\ d\tau \tag{II.11}$$

$$b_j = \int_0^1 L_j(\tau)\ d\tau, \tag{II.12}$$

where $L_j(\tau)$ is the Lagrangian interpolation polynomial

$$L_j(\tau) = \prod_{\substack{\ell=1 \\ \ell \neq j}}^{s} \frac{\tau - c_\ell}{c_j - c_\ell}.$$

*Proof.* By the collocation polynomial $u_s(t)$ we define $r_j := u_s'(t_i + c_j\Delta t)$. By the Lagrange interpolation formula, for any $\tau \in [0,1]$ we have that

$$u_s'(t_i + \tau\Delta t) = \sum_{\ell=1}^{s} L_\ell(\tau) r_\ell.$$

Integration gives

$$u_s(t_i + c_j\Delta t) = u_s(t_i) + \Delta t \cdot \sum_{\ell=1}^{s} r_\ell \int_0^{c_j} L_\ell(\tau)\ d\tau$$

for all $j = 1, \ldots, s$. Inserting for (II.11) gives

$$u_s(t_i + c_j\Delta t) = u_s(t_i) + \Delta t \sum_{\ell=1}^{s} r_\ell a_{j\ell}.$$

Since

$$r_j = u_s'(t_i + c_j\Delta t) \overset{\text{(II.10)}}{=} f\Big(t_i + c_j\Delta t,\ u_s(t_i + c_j\Delta t)\Big),$$

we obtain

$$u_s(t_i + c_j\Delta t) = u_s(t_i) + \Delta t \sum_{\ell=1}^{s} a_{j\ell} f\Big(t_i + c_j\Delta t,\ u_s(t_i + c_j\Delta t)\Big), \tag{$*$}$$

or

$$u_s(t_i + \Delta t) = u_s(t_i) + \Delta t \sum_{j=1}^{s} b_j f\Big(t_i + c_j\Delta t,\ u_s(t_i + c_j\Delta t)\Big). \tag{$**$}$$

Since in the collocation method we set $u_s(t_i) = y_i$ and $u_s(t_{i+1}) = y_{i+1}$, we have the Runge-Kutta method in $(*)$ and $(**)$. □

We do not know yet if every Runge-Kutta method originates in collocation. In general, this is not true, as the following example demonstrates: Consider the two stage Runge-Kutta method with $c_1 = 0,\ c_2 = \frac{2}{3}$. Computing $L_1, L_2$ and $a_{ji}, b_i$ via (II.11) and (II.12) will give

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}.$$

Given that every choice of collocation points corresponds to an unique collocation method (Lagrange interpolation, Numerical Mathematics I) we deduce, that the implicit Runge-Kutta method

$$\begin{array}{c|cc} 0 & \frac{1}{4} & -\frac{1}{4} \\ \frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\ \hline & \frac{1}{4} & \frac{3}{4} \end{array}$$

with order $p \geq 3$ has no collocation counterpart.

**Example.** For $s = 1$ the polynomial defined by

$$u(t) = y_i + (t - t_i) f(t_i + c_1 \Delta t,\ u(t_1 + c_1 \Delta t))$$

gives a collocation method like in (II.10). For $c_1 = 0$ it gives the explicit Euler method, for $c_1 = 1$ it gives the implicit Euler method and for $c_1 = \frac{1}{2}$ the midpoint rule is given. The drawing is missing.

We want to know how the order $p$ of an implicit Runge-Kutta method depends on the choice of the collocation method. To answer this question reconsider (II.1), i.e. the ODE

$$y'(t) = f(t, y(t))$$

and let $v$ be a "candidate" solution. The defect is defined as

$$d(t, v) = v'(t) - f(t, v(t)). \tag{II.13}$$

When the defect is small, the error $\|y(t) = v(t)\|$ is small too, because if $d(t, v) = 0$ our candidate $v$ is the solution.

Let us consider the linear case

$$y' = \Lambda y\ , y(t_0) = y_0.$$

Then the defect is given by

$$D(t) = v'(t) - \Lambda v(t).$$

Plugging in for our candidate solution $v$ we thus arrive at the linear inhomogeneous ODE

$$v'(t) = \Lambda v(t) + D(t),\ t \geq t_0,\ v(t_0) \text{ given.}$$

The exact solution of this ODE (theorem I.12) is given by

$$v(t) = e^{(t-t_0)\Lambda} + \int_{t_0}^{t} e^{(t-\tau)\Lambda} D(\tau)\ d\tau$$

wheras the solution of (II.13) is given by

$$y(t) = e^{(t-t_0)\Lambda} y_0\ t \geq t_0.$$

Hence

$$v(t) - y(t) = e^{(t-t_0)\Lambda}(v_0 - y_0) + \int_{t_0}^{t} e^{(t-\tau)\Lambda} D(\tau)\ d(\tau)$$

for $t \geq t_0$. Thus, the error can be controlled by the observables $v_0, y_0$ and the defect $D(\tau)$.

**Theorem II.15** (Alekseenz-Gröbner-Lemma)**.** Let $v$ be a smoothly differentiable function that obeys the initial conditions $v(t_0) = y_0$. Then

$$v(t) - y(t) = \int_{t_0}^{t} \Xi(t, \tau) d(\tau, v)\ d\tau, \quad t \geq t_0,$$

where $\Xi$ is the matrix of the partial derivatives of the solution of the ODE

$$w' = f(t, w),\ w(\tau) = v(\tau).$$

*Proof.* The proof can be found in *Deuflhard-Bornemann, Numerical Mathematics, Chapter III, Theorem 3.4* and is not given in this lecture. □

**Excursion II.2 (Quadrature)**

**Excursion Definition 3.** A quadrature rule $Q_{n+1}[f]$ is exact of order $n \in \mathbb{N}$ if

$$R_n[p] = I[p] - Q_{n+1}[p] = 0$$

holds for all $p \in \mathbb{P}_{n+1}$.

Recall that the scalar product of to functions is given by

$$\langle f, g \rangle = \int_a^b f(x)g(x)w(x)\ dx,$$

where $w(x)$ is a weight-function.

**Excursion Theorem 4.** Let $r = \mathbb{P}_s$ obey

$$\langle r, p \rangle = 0$$

for every $p \in \mathbb{P}_s$ and

$$\langle r, t^s \rangle \neq 0.$$

Let $c_1, \ldots, c_s$ be the zeros of the polynomial $r$ and choose $b_1, \ldots, b_s$ such that

$$\sum_{j=1}^{s} b_j c_j^{s-1} = \int_a^b \tau^{s-1}\ d\tau.$$

Then the quadrature rule has exactness of order $2s$.

**Theorem II.16.** An implicit Runge method $(A, b, c)$ generated by the collocation method (II.10) has for $f \in C^p(\mathbb{R} \times \mathbb{R}^d, \mathbb{R}^d)$ the consistency order $2s$ if and only if the quadrature formula defined by the nodes $c$ and weights $b$ meet the requirements of the previous excursion theorem, i.e. if it is of order $2s$.

*Proof.* We only give a scetch of the proof. Use theorem II.15 and $u(t_i) = y_i$, where $y_{i+1}$ is the solution of the collocation method and $y(t_{i+1})$ is the true solution. We obtain

$$y_{i+1} - y(t_{i+1}) = \int_{t_i}^{t_{i+1}} \Xi(t_{i+1}, \tau) d(\tau, u)\ d\tau.$$

Applying a quadrature rule with $w(t) = 1$ and the quadrature points $t_i + c_1\Delta t, \ldots, t_i + c_s\Delta t$ leads to

$$y_{i+1} - y(t_{i+1}) = \sum_{j=1}^{s} b_j d(t_i + c_j\Delta t, u) + \text{error of quadrature from (??)}.$$

According to (II.10) it holds

$$\begin{aligned} d(t_i + c_j\Delta t, u) &= u'(t_i + c_j\Delta t) - f(t_i + c_j\Delta t, u(t_i + c_j\Delta t)) \\ &\overset{\text{(II.10)}}{=} 0. \end{aligned}$$

Thus, according to excursion theorem the order of quadrature with the weight functions $w(t) = 1,\ 0 \le t \le 1$ and $c_1, \ldots, c_s$ is $2s$. □

We have seen that for an explicit Runge-Kutta method it is not possible to choose $p > s$. However, when dealing with implicit Runge-Kutta methods, by a suitable choice of $c_i, a_{ji}$ the implicit Runge-Kutta method can be reffered to as a collocation method. The theorem of Alekseenz-Gröbner tells us that

$$y_{i+1} - y(t_{i+1}) \le \text{quadrature error with certain exactness.}$$

So if we choose the $c_j$ such that they refer to an implicit Runge-Kutta method and also have an quadrature error of order $2s$, then we can obtain up to $p \le 2s$.

**Example** (Gauß-Legendre Runge-Kutta method)**.** $s$-stage, order $2s$ methods, choose $c_1, \ldots, c_s \in (0, 1)$ as zeros of the Legendre polynomial $p_s$ given by

$$p_s(t) = \frac{(s!)^2}{(2s)!} \cdot \sum_{k=0}^{s} (-1)^k \binom{s}{k} \binom{s+k}{k} t^k.$$

For $s = 1$ we get

$$p_1(t) = t - \tfrac{1}{2},$$

so $c_1 = \frac{1}{2}$ and the Butcher table is

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}.$$

For $s = 2$ we have

$$p_1(t) = t^2 - t + \tfrac{1}{6},$$

so $c_1 = \frac{1}{2} - \frac{\sqrt{3}}{6}$ and $c_2 = \frac{1}{2} + \frac{\sqrt{3}}{6}$ and by (II.11) and (II.12) from theorem II.14 we have the Butcher table given by

$$\begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} - \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}.$$

From Lemma II.10 we already know that Butcher trees are a method for finding the conditions on $(A, b, c)$ for reaching a certain order of an explicit Runge-Kutta method. The order of a tree is defined as the number of nodes in the tree. It grows exponentially in the treesize. The tree drawings are missing. We have the polynomial $\Phi$ defined by

$$\Phi(t) = \sum_{i,j} b_i c_i^2 a_{ij} c_j^2$$

and $\gamma(t) \in \mathbb{N} \setminus \{0\}$ by ... We obtain the following table:

| tree | order | $\gamma$ | condition |
|---|---|---|---|
| (a) | 1 | 1 | $\sum_i b_i = 1$ |
| (b) | 2 | 2 | $\sum_i b_i c_i = \frac{1}{2} = \frac{1}{\gamma(t)}$ |
| (c) | 3 | 3 | $\sum_i b_i c_i^s = \frac{1}{3}$ |

.

**Stability (A Stability)**

Looking at the explicit Euler method for the solution of the IVP

$$y'(t) = -20y(t),\ y(0) = 1.2$$

one can observe oscillation for large $t$. This motivates the definition of stability of a numerical scheme.

Consider a vector valued IVP given by

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(0) &= \hat{y}_0 \end{aligned} \tag{II.15}$$

We want to know what happens with small perturbations of $u_i$ at some $t_i$. We have

$$(y+u)'(t) = f(t, (y+u)(t))$$

with

$$(y+u)(t_i) = y(t_i) + u_i$$

for $t \geq t_i$. Using Taylor expansion gives

$$\begin{aligned} u'(t) &= (y+u)'(t) - y'(t) \\ &= f(t, (y+u)(t)) - f(t, y(t)) \\ &\approx f(t, y(t)) + \frac{\partial f}{\partial y}(t, y(t)) \cdot u(t) - f(t, y(t)) \\ &= \frac{\partial f}{\partial y}(t, y(t)) \cdot u(t). \end{aligned}$$

Freezing the matrix $\frac{\partial f}{\partial y}(t, y(t))$ at $t_i$ we obtain a linear system of ODEs with constant coefficients:

$$u'(t) = \frac{\partial f}{\partial y}(t, y(t_i)) \cdot u(t)$$

or

$$u'(t) = Au(t),$$

where $A$ is a matrix.

Remember, that for an $A \in \mathbb{C}^{n,n}$ the ODE

$$y'(t) = Ay(t) \tag{$*$}$$

has for $n = 1$ the solution

$$y(t) = y(t_0)e^{at},$$

where $a = A$. Using this as an Ansatz for $(*)$ we obtain

$$y(t) = ve^{\lambda t},\ \lambda \in \mathbb{C}, v \in \mathbb{C}^n.$$

Hence, $y(t) = ve^{\lambda t}$ solves $(*)$ if and only if $\lambda$ is an eigenvalue of $A$ and $v$ is an eigenvector of $A$, since

$$\lambda v e^{\lambda t} = Ave^{\lambda t}.$$

From Analysis II, recall that if $A$ is diagonalizable, then the matrix

$$Y = (v_1 e^{\lambda_1}, \dots, v_n e^{\lambda_n})$$

is the fundamental matrix of $y'(t) = Ay(t)$.

Heading back, this means that

$$u(t) = c_1 v_1 e^{\lambda_1 t} + \cdots + c_n v_n e^{\lambda_n t},$$

where $\lambda_i$ are eigenvalues and $v_i$ are eigenvectors. If $\mathrm{Re}(\lambda_i) < 0$, then it holds

$$\lim_{t\to\infty} u(t) = 0.$$

In this case, we consider the IVP as moderate, since small local perturbations fall off.[3]

We thus consider the scalar test problem

$$\begin{aligned} y'(t) &= \lambda y(t) \\ y(0) &= 1. \end{aligned} \tag{II.D}$$

This problem is called the **Dahlquist Test-Problem**.

**Definition II.17** (A-stability)**.** We call a numerical method **A-stable** (absolute stable) if its approximates $y_i$ of (II.D) for each $\lambda \in \mathbb{C}^- = \{\lambda \in \mathbb{C} \mid \mathrm{Re}(\lambda) < 0\}$ with arbitrary but fixed $\Delta t$ are contractive, i.e. if

$$|y_{i+1}| < |y_i| \tag{II.16}$$

for all $i \in \mathbb{N}$

**Example.**

(a) The explicit Euler method was given by $y_{i+1} = y_i + \Delta t f(t_i, y(t_i))$. Applying this on (II.D) we obtain

$$\begin{aligned} y_{i+1} &= y_i + \Delta t \cdot \lambda \cdot y_i \\ &= (1 + \Delta t\lambda) y_i, \end{aligned}$$

so we have to choose $\lambda\Delta t \in \{z \in \mathbb{C} \mid |z + 1| < 1\}$, since otherwise $(1+\Delta t\cdot\lambda) \not< 1$. Hence, the explicit Euler method is **not stable**.

(b) The implicit Euler method was given by $y_{i+1} = y_i + \Delta t f(t_{i+1}, y_{i+1})$. Applying this to (II.D) yields

$$y(t_{i+1}) = y_i + \Delta t \cdot \lambda y_{i+1},$$

or equivalently

$$y_{i+1} = \frac{1}{(1 - \Delta t\lambda)} y_i.$$

Hence, the implicit Euler method is A-stable.

[3]Note, that for sakes of simplicity we assumed pairwise different eigenvalues. However, above standing considerations also hold for multiple eigenvalues.

(c) For the Runge-Kutta methods we have

$$y_{i+1} = y_i + \Delta t \sum_{j=1}^{s} b_j f(t_i + c_j \Delta t, k_j).$$

Applying this to (II.D) yields

$$y_i + \Delta t \lambda \sum_{j=1}^{s} b_j k_j.$$

Remember that $k_j$ was given by

$$k_j = y_i + \Delta t \sum_{\nu=1}^{s} a_{j\nu} f(t_i + c_j \Delta t, k_\nu) \overset{\text{(II.D)}}{=} y_i + \Delta t \lambda \sum_{\nu=1}^{s} a_{j\nu} b_j.$$

Rewriting with

$$k = \begin{bmatrix} k_1 & \cdots & k_s \end{bmatrix}^T,\ e = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T \in \mathbb{R}^s$$

gives $k = y_i e + \Delta t A k$, where $A = [a_{j\nu}]$. This is equivalent to

$$(I_s - \Delta t \lambda A) k = y_i e.$$

By assuming that $\frac{1}{\Delta t k} \notin \sigma(A)$, where $\sigma(A)$ is the spectrum of $A$, it holds

$$k = (I_s - \Delta t \lambda A)^{-1} y_i e.$$

With this, we obtain

$$\begin{aligned} y_{i+1} &= y_i + \Delta t \lambda b^T k \\ &= y_i + \Delta t \lambda b^T (I_s - \Delta t \lambda A)^{-1} y_i e \\ &= \underbrace{\left(1 + \Delta t \lambda b^T \left(I_s - \Delta t \lambda A\right)^{-1} e\right)}_{=: R(\lambda \Delta t)} y_i. \end{aligned}$$

The observation for Runge-Kutta methods motivates the following definition.

**Definition II.18** (Stability function of Runge-Kutta methods)**.** For $\hat{\sigma}(A) = \left\{\lambda \in \mathbb{C} \setminus \{0\} \mid \lambda^{-1} \in \sigma(A)\right\}$ we call the mapping

$$R : \mathbb{C} \setminus \hat{\sigma}(A) \to \mathbb{C}, \quad \xi \mapsto R(\xi) = 1 + \xi b^T (I_s - \xi A)^{-1} e$$

**stability function of the Runge-Kutta method** defined by $(A, b, c)$.

**Theorem II.19.** A Runge-Kutta method is A-stable if and only if $|R(\xi)| < 1$ for $\xi \in \mathbb{C}^-$, where $R$ is its stability function.

*Proof.* No proof. □

We want to know how exactly the stability function of a Runge-Kutta method looks like, if we consider explicit or implicit Runge-Kutta methods.

**Theorem II.20.** The stability function of a Runge-Kutta method given by $(A, b, c)$ is

(a) a polynomial of degree less or equal $s$ in the case of an explicit RKM and

(b) a rational function, which possibly contains the inverse of the eigenvalues of $A$ as poles in the case of an implicit RKM.

*Proof.*

ad (a). $A \in \mathbb{R}^{s,s}$ is a strictly lower triangular matrix, hence $A^s = 0$. We consider the Neumann series given by

$$(I_s - \xi A)^{-1} = I_s + \xi A + \cdots + \xi^{s-1} A^{s-1}$$

such that

$$R(\xi) = 1 + \xi b^T (I_s \xi A)^{-1} e = 1 + b^T (\xi I_s + \cdots + \xi^s A^{s-1}) e$$

is a polynomial.

ad (b). We solve the linear system $(I - \xi A)v = e$ by using Cramer's rule and obtain

$$\begin{aligned} v_j &= \frac{\det\Big((I - \xi A)_1, \cdots, e, \cdots, (I - \xi A)_s\Big)}{\det(I - \xi A)} \\ &= \frac{p_j(\xi)}{\det(I - \xi A)}, \end{aligned}$$

with $p_j \in \mathbb{P}_{s-1}$. Thus, we have

$$R(\xi) = 1 + \frac{\sum_{j=1}^{s} b_j p_j(\xi)}{\det(I - \xi A)} \cdot \xi,$$

what ends the proof.

□

**Theorem II.21.** There exists no A-stable explicit Runge-Kutta method.

*Proof.* In the explicit case we have $\lim_{|\xi| \to \infty} R(\xi) = \infty$. □

We have thus seen that for explicit RKM the stepsize can not be chosen arbitray. So we seek for $\Delta t$ such that $|R(\lambda \Delta t)| < 1$. This leads us to the next definition.

**Definition II.22** (Domain of stability)**.** The set $S := \{\xi \in \mathbb{C} \mid |R(\xi)| < 1\}$ is called the **domain of stability** of the RKM belonging to $R$.

There might be some scenarios where the usage of a constant $\Delta t$ might be suboptimal, since the solution can behave very differently at different times. Thus, in the following, we are not going to assume that $\Delta t$ is constant. Not only the discretization error, but also the **rounding error** is important. When, by rounding, we obtain $\tilde{y}_i$ instead of $y_i$, the rounding error is given by

$$|y_i - \tilde{y}_i| = O\left(\frac{\varepsilon}{\Delta t}\right).$$

The discretization error is given by

$$|y(t_i) - y_i| = O(\Delta t^p).$$

Since the rounding error rises as $\Delta t \to 0$ and the discretization error rises as $\Delta t \to \infty$, we want to know if there exists an optimal stepsize $t_{\text{opt}}$.

But how is it possible to find a suitable step size, when the exact solution is unknown? Let $y_i$ be an approximation of $O(\Delta t^p)$ and $\tilde{y}_i$ be an approximation of $O(\Delta t^q)$ where $q > p$. Then we have[4]

$$\|y(t_i) - y_i\| \leq \|y(t_i) - \tilde{y}_i\| + \|\tilde{y}_i + y_i\| = O(\Delta t^q) + \|\tilde{y}_i - y_i\|$$

and hence $O(\Delta t^q) + \|\tilde{y}_i - y_i\|$ is an estimator for the exact solution of order $q$.

Consider an equidistant grid of points and a one-step method

$$\begin{aligned} y_{\Delta t}(t + \Delta t) &= y_{\Delta t}(t) + \Delta t \Phi(t, y_{\Delta t}(t), \Delta t) \\ y_{\Delta t}(t_0) &= y_0 \end{aligned} \tag{II.16}$$

[4]Note that $y(t_i)$ is unknown.

with order of consistency $p \in \mathbb{N}$ and local error (Def II.2)

$$\eta(t+\Delta t) = y(t) - \Delta t \Phi\big(t, y(t), \Delta t\big) - y(t+\Delta t) = O(\Delta t^{p+1}). \tag{II.17}$$

If $\Phi$ is sufficiently often differentiable, we can expand the error and obtain

$$\eta(t, \Delta t) = \sum_{i=p+1}^{N+1} g_i(t) \Delta t^i + O(\Delta t^{N+2}). \tag{II.18}$$

The next theorem shows that a similar expansion is possible for the global error $e(t, \Delta t)$.

**Theorem II.23** (Gregg)**.** Let $f(t,y)$ and $\Phi$ be sufficiently often continuously differentiable on $S = \big\{(t,y) \mid t_0 \le t \le t_{\text{end}},\ y \in \mathbb{R}^d\big\}$ and let us assume that the local error $\eta$ has the expansion (II.18). Then the global error $e(t, \Delta t)$ after $n$-steps with stepsize $\Delta t$ has in $t^* = t_0 + n\Delta t$ has an asymptotic expansion of the form

$$e(t, \Delta t) = e_p(t^*)\Delta t^p + e_{p+1}(t^*)\Delta t^{p+1} + \cdots + e_N(t^*)\Delta t^N + E_{N+1}(t^*, \Delta t)\Delta t^{N+1},$$

where the remainder $E_{N+1}(t^*, \Delta t)$ is bounded for $0 < \Delta t < \Delta t_0$.

*Proof.* Observe that

$$y(t+\Delta t) = y(t) + \int_t^{t+\Delta t} f(\tau, y(\tau))\ d\tau \rightsquigarrow y_{\Delta t}(t+\Delta t) = y(t) + \Delta t \Phi(t, y(t), \Delta t),$$

where

$$\Phi(t, y(t), \Delta t) = \frac{1}{\Delta t} \int_t^{t+\Delta t} f(\tau, y(\tau))\ d\tau - \frac{\eta(t, \Delta t)}{\Delta t}.$$

We now set

$$\Phi_y(t, y(t), \Delta t) = f_y(t, y) d\tau + O(\Delta t). \tag{II.19}$$

We construct a method of order $p+1$ such that

$$y(t) - y_{\Delta t}(t) = e_p(t)\Delta t^p + O(\Delta t^{p+1}).$$

We set

$$\bar{y}_{\Delta t}(t) = e_p(t)\Delta t^p + y_{\Delta t}(t). \tag{II.20}$$

For the "new" one step method we thus have

$$\begin{aligned} \bar{y}_{\Delta t}(t+\Delta t) &= \bar{y}_{\Delta t}(t) + \Delta t \bar{\Phi}\left(t, \bar{y}_{\Delta t}(t), \Delta t\right) \\ \bar{y}_{\Delta t}(t_0) &= y_0. \end{aligned} \tag{II.21}$$

We then have

$$\begin{aligned} \bar{y}_{\Delta t}(t+\Delta t) &\overset{\text{(II.20)}}{=} e_p(t+\Delta t)\Delta t^p + y_{\Delta t}(t+\Delta t) + O(\Delta t^{p+1}) \\ &\overset{\text{(II.16)}}{=} e_p(t+\Delta t)\Delta t^p + y_{\Delta t}(t) + \Delta t \Phi(t, y(t), \Delta t) + O(\Delta t^{p+1}) \\ &= \ e_p(t+\Delta t)\Delta t^p \overset{\text{(II.20)}}{+} \bar{y}_{\Delta t}(t) - e_p(t)\Delta t^p \\ &\quad + \Delta t \Phi\Big(t, \bar{y}_{\Delta t}(t) - e_p(t)\Delta t, \Delta t\Big) + O(\Delta t^{p+1}) \\ &= \ \bar{y}_{\Delta t}(t) + \Delta t \Phi\Big(t, \bar{y}_{\Delta t}(t) - e_p(t)\Delta t, \Delta t\Big) \\ &\quad + \Big(e_p(t+\Delta t) - e_p(t)\Big) \cdot \Delta t^p + O(\Delta t^{p+1}). \end{aligned}$$

Comparing (II.21) with this result gives

$$\Phi\Big(t, y_{\Delta t}(t) - e_p(t)\Delta t, \Delta t\Big) + \Big(e_p(t+\Delta t) - e_p(t)\Big)\Delta t^{p-1} = \bar{\Phi}(t, y_{\Delta t}(t), \Delta t).$$

For $\bar{\eta}$ it holds

$$\begin{aligned}
\bar{\eta}(t,\Delta t) &= y(t) + \Delta t\bar{\Phi}(t,y(t),\Delta t) - y(t+\Delta t)\\
&= y(t) + \Delta t\Phi\Big(t, y(t) - e_p(t)\Delta t^p, \Delta t\Big)\\
&\quad + \Big(e_p(t+\Delta t) - e_p(t)\Big)\Delta t^p - y(t+\Delta t) + O(\Delta t^{p+1})\\
&= y(t) + \Delta t\Phi\Big(t, y(t) - e_p(t)\Delta t^p, \Delta t\Big) + e_p'\Delta t^{p+1} - y(t+\Delta t) + O(\Delta t^{p+1})\\
&= y(t) + \Delta t\Phi\big(t, y(t), \Delta t\big) - f_y(t,y)\cdot e_p(t)\Delta t^{p+1}\\
&\quad + e_p'(t)\Delta t^{p+1} - y(t+\Delta t) + O(\Delta t^{p+1})\\
&= \Delta t^{p+1}\Big(g_{p+1}(t) + e_p'(t) - f_y(t,y(t))e_p(t)\Big) + O(\Delta t^{p+2})
\end{aligned}$$

and thus the method (II.21) has order of consistency $p+1$, if $e_p(t)$ solves the IVP

$$\begin{aligned}
e_p(t) &= f_y(t,y(t))e_p(t) + g_{p+1}(t)\\
e_p(t_0) &= 0.
\end{aligned}$$

For the global error of (II.21) it holds (according to theorem II.7)

$$y(t^*) - \bar{y}_{\Delta t}(t^*) = E_{p+1}(t,\Delta t)\Delta t^{p+1}$$

and with (II.20) it holds

$$e(t^*,\Delta t) = e_p(t^*)\Delta t^p + E_{p+1}(t^*,\Delta t)\Delta t^{p+1}.$$

For the next step $e_{p+1}(t)$ we start the proof at (II.21). □

With theorem II.23 in our hands, we can

(a) construct a method of higher order and

(b) construct an error estimator for the global error.

ad (a): Let

$$\begin{aligned}
y_{i+1}^{\Delta t},&\ \text{stepsize } \Delta t,\ 1 \text{ step}\\
y_{i+1}^{\Delta t/2},&\ \text{stepsize } \frac{\Delta t}{2},\ 2 \text{ step}
\end{aligned}$$

be methods on $[t_i, t_{i+1})$. Then, according to theorem II.23 it holds

$$\begin{aligned}
y(t_{i+1}) &= y_{i+1}^{\Delta t} + e_p(t_{i+1})\Delta t^p + O(\Delta t^{p+1})\\
y(t_{i+1}) &= y_{i+1}^{\Delta t/2} + e_p(t_{i+1})\frac{\Delta t^p}{2} + O(\Delta t^{p+1}).
\end{aligned} \tag{II.23}$$

Thus, substracting $(\frac{1}{2})^p$ times the first line from the second line, we get

$$y(t_{i+1}) - \left(\frac{1}{2}\right)^p y(t_{i+1}) = y_{i+1}^{\Delta t/2} - \left(\frac{1}{2}\right)^p y_{i+1}^{\Delta t} + O(\Delta t^{p+1})$$

and hence

$$y(t_{i+1}) = \frac{y_{i+1}^{\Delta t/2} - \left(\frac{1}{2}\right)^p y_{i+1}^{\Delta t}}{1 - \left(\frac{1}{2}\right)^p} + O(\Delta t^{p+1}).$$

Thus,

$$\tilde{y}_{i+1}^{\Delta t} = \frac{y_{i+1}^{\Delta t/2} - \left(\frac{1}{2}\right)^p y_{i+1}^{\Delta t}}{1 - \left(\frac{1}{2}\right)^p} + O(\Delta t^{p+1})$$

is an $O(\Delta t^{p+1})$ approximation of $y(t_{i+1})$.

ad (b): Using (II.23) and substracting the second line from the first line, we get

$$O(\Delta t^{p+1}) + \frac{y_{i+1}^{\Delta t} - y_{i+1}^{\Delta t/2}}{1 - \left(\frac{1}{2}\right)^p} = e_p(t_i)\Delta t^p.$$

Setting

$$\mathrm{EST}_{i+1}^{(\mathrm{g})} = \frac{\left\| y_{i+1}^{\Delta t} - y_{i+1}^{\Delta t/2} \right\|}{1 - \left(\frac{1}{2}\right)^p}, \tag{II.24}$$

for some given tolerance TOL of the global error, we can choose a suitable stepsize, by computing $y_{i+1}^{\Delta t}$ and $y_{i+1}^{\Delta t/2}$ and then

- increase $\Delta t$, if $\mathrm{EST}_{i+1}^{(\mathrm{g})} \leq \mathrm{TOL}$ and
- decrease $\Delta t$, if $\mathrm{EST}_{i+1}^{(\mathrm{g})} \geq \mathrm{TOL}$.

To derive an estimator of the local error, first observe that $\Delta t$ is not constant, i.e. we write $\Delta t_i$ instead of $\Delta t$. Hence, we want to adjust the time step size according to the local behaviour of the solution. This is why we need an estimate.

Consider a explicit one step method $\Phi$, once with stepsize $\Delta t$ (we denote the outcome by $y_{i+1}^{\Delta t}$) and once with stepsize $\Delta t/2$ (we denote the outcome by $\bar{y}_{i+1}^{\Delta t}$. Then,

$$\bar{y}_{i+1}^{\Delta t} = y_i^{\Delta t} + \frac{\Delta t}{2}\Phi\left(t_i, y_i^{\Delta t}, \frac{\Delta t}{2}\right) + \frac{\Delta t}{2}\Phi\left(t_i + \frac{\Delta t}{2}, y_i^{\Delta t} + \Delta t\Phi\left(t_i, y_i^{\Delta t}, \frac{\Delta t}{2}\right), \frac{\Delta t}{2}\right).$$

By (II.18), we know that the local error $\eta(t, \Delta t) = O(\Delta t^{p+1})$ can be written as

$$\eta(t, \Delta t) = g_{p+1}(t)\Delta t^{p+1} + O(\Delta t^{p+2}).$$

Thus, it holds

$$\bar{\eta}(t_i, \Delta t) = g_{p+1}(t_i) \cdot \left(\frac{1}{2}\right)^{p+1} + g_{p+1}\left(t_i + \frac{\Delta t}{2}\right)\left(\frac{\Delta t}{2}\right)^{p+1} + O(\Delta t^{p+2})$$

and using a Taylor-expansion in the middle part gives

$$g_{p+1}\left(t_i + \frac{\Delta t}{2}\right)\left(\frac{\Delta t}{2}\right)^{p+1} = g_{p+1}(t_i)\left(\frac{\Delta t}{2}\right)^{p+1} + O(\Delta t^{p+2}).$$

Hence, we have

$$\bar{\eta}(t_i, \Delta t) = 2g_{p+1}(t_i)\left(\frac{\Delta t}{2}\right)^{p+1} + O(\Delta t^{p+1}) = \left(\frac{1}{2}\right)^p \underbrace{g_{p+1}(t_i)\Delta t^{p+1}}_{=\eta(t,\Delta t)} + O(\Delta t^{p+2}). \quad (*)$$

So it holds

$$y(t_{i+1}) = \bar{y}_{i+1} + \bar{\eta}(t_i, \Delta t) + O(\Delta t^{p+2})$$
$$y(t_{i+1}) = y_{i+1} + \eta(t_i, \Delta t) + O(\Delta t^{p+2})$$

and plugging in $(*)$ gives

$$y(t_{i+1}) = \bar{y}_{i+1} + \left(\tfrac{1}{2}\right)^p g_{p+1}(t_i) + \Delta t^{p+1} + O(\Delta t^{p+2})$$
$$y(t_{i+1}) = y_{i+1} + g_{p+1}(t_i) + \Delta t^{p+1} + O(\Delta t^{p+2}).$$

We can now define

$$\mathrm{EST}_{i+1}^{(\mathrm{l})} = \frac{\left\| \bar{y}_{i+1}^{\Delta t} - y_{i+1}^{\Delta t} \right\|}{\left(\frac{1}{2}\right)^p - 1}$$

as an estimate.

This gives the **Embedding-Stategy for step size control**. Let us assume that we have two one step methods with incremental functions $\Phi_p$ and $\Phi_q$ such that $\Phi_p$ has order of consistency $p$, i.e. $\eta_p(t,\Delta t) = O(\Delta t^{p+1})$ and $\Phi_q$ has order of consistency $q$, i.e. $\eta_q(t,\Delta t) = O(\Delta t^{q+1})$. Without loss of generality, we can assume $q > p$. Then

$$\begin{aligned}\eta_p(t,\Delta t) &= y(t) + \Delta t\Phi_p - y(t+\Delta t) + \Delta t\Phi_q - \Delta t\Phi_q \\ &= y(t) + \Delta t\Phi_q - y(t+\Delta t) + \Delta t\left(\Phi_p - \Phi_q\right) \\ &= \eta_q(t,\Delta t) + \Delta t\left(\Phi_p - \Phi_q\right).\end{aligned}$$

If we now **assume**[5] that the error $\eta_q$ is negligible with respect to $\eta_p$, we have

$$|\Phi_p - \Phi_q| \approx O(\Delta t^p)$$

and so

$$\eta_p(t,\Delta t) \approx \Delta t(\Phi_p - \Phi_q).$$

We define

$$\varepsilon(\Delta t) = \frac{\eta_p(t,\Delta t)}{\Delta t}.$$

With this, the task of finding a new stepsize $\Delta t_{\text{new}}$ is realized by

$$\varepsilon(\Delta t_{\text{new}}) = \varepsilon_{\text{goal}}.$$

With the assumption

$$\eta_p(t,\Delta t) = c\cdot\Delta t^{p+1}$$

we obtain

$$\frac{\varepsilon_{\text{goal}}}{\Delta t_{\text{new}}^p} \approx \frac{\eta_p(t,\Delta t_{\text{new}})}{\Delta t_{\text{new}}^{p+1}} \approx c \approx \frac{\eta_p(t,\Delta t)}{\Delta t^{p+1}} \approx \frac{\varepsilon(\Delta t)}{\Delta t^p}$$

so we can use

$$\Delta t_{\text{new}} = \Delta t\sqrt[p]{\frac{\varepsilon_{\text{goal}}}{\varepsilon(\Delta t)}}$$

as an iteration. Hence, we choose $\Delta t$ courser, if $\varepsilon_{\text{goal}} > \varepsilon(\Delta t)$ and else finer.[6]

The embedding strategy can be realized in a Runge-Kutta scheme. Consider the RKM given by the Butcher table

| | | | | |
|---|---|---|---|---|
| 0 | | | | |
| 1/2 | 1/2 | | | |
| 3/4 | 0 | 3/4 | | |
| 1 | 2/9 | 1/3 | 4/9 | |
| $p=2$ | 7/24 | 1/4 | 1/3 | 1/8 |
| $p=3$ | 2/9 | 1/3 | 4/9 | |

This means

$$\begin{aligned}r_1 &= f(t_i, y_i)\\ r_2 &= f\left(t_i + \tfrac{\Delta t}{2},\ y_i + \tfrac{\Delta t}{2}r_1\right)\\ r_3 &= f\left(t_i + \tfrac{3}{2}\Delta t,\ y_i + \tfrac{3}{2}\Delta t r_2\right)\\ r_4 &= f\left(t_i + \Delta t,\ y_i + \tfrac{2}{9}\Delta t r_1 + \tfrac{1}{3}\Delta t r_2 + \tfrac{4}{9}\Delta t r_3\right),\end{aligned}$$

and

$$\begin{aligned}y_{i+1} &= y_i + \tfrac{7}{24}r_1 + \tfrac{1}{4}r_2 + \tfrac{1}{3}r_3 + \tfrac{1}{8}r_4\\ y_{i+1} &= y_i + \tfrac{2}{9}r_1 + \tfrac{1}{3}r_2 + \tfrac{4}{9}r_3.\end{aligned}$$

Hence, it is possible to gain a higher order by simply changing the weights $b_i$. This RKM is due to *Bobacki-Shampine* and can be found in `MATLAB` as `ode23`. Note, that the RKM of higher order does not use $r_4$. The Fehlberg-Trick is known by

$$r_4^i = f\left(t_i + \Delta t, y_i + \Delta t\left(\tfrac{2}{9}r_1^i + \tfrac{1}{3}r_2^i + \tfrac{4}{9}r_3^i\right)\right) =: r_1^{i+1}.$$

This allows for higher efficiency and is often used in programming.

[5] This is not always given!

[6] Warning! We had $\eta_p(t,\Delta t) = O(\Delta t^{p+1})$, for $\Delta t \to 0$, so $\eta_p(t,\Delta t) \le c\Delta t^{p+1}$ only is a reasonable assumption for small $\Delta t$.

## II.3 Multistep Methods

In one step methods, computing $y_{i+1}$ only depends on $y_i$. However, in multistep methods $y_{i+1}$ can also depend on the previous steps $y_{i-1}, y_{i-2}$ and so on.

**Example.** Let $y_i, y_{i+1}$ be given. Applying the midpoint rule gives

$$y(t_{i+2}) - y(t_i) = \int_{t_i}^{t_{i+1}} y'(\tau)\ d\tau = 2\Delta t f(t_{i+1}, y_{i+1}) + O(\Delta t^2)$$

and we thus obtain

$$y_{i+2} = y_i + 2\Delta t + f(t_{i+1}, y_{i+1}).$$

**"Adam's Family"**

Analogue to the last example, we have

$$y(t_{i+m}) - y(t_{i+m-r}) = \int_{t_{i+m-r}}^{t_{i+m}} f(\tau, y(\tau))\ d\tau.$$

Replacing $f$ by an interpolation polynomial $q$ we obtain

$$y(t_{i+m}) - y(t_{i+m-r}) = \int_{t_{i+m-r}}^{t_{i+m}} q(\tau)\ d\tau.$$

We now give examples of explicit and implicit multistep methods.

**Example** (Adams-Bashorth Method)**.** Let $r = 1$ and $q \in \mathbb{P}_{m-1}$ with

$$q(t_{i+j}) = f(t_{i+j}, y_{i+j})$$

for $j = 0, \dots, m-1$.

For different $m \in \mathbb{N}$, we obtain the methods

$m = 1$: $y_{i+1} = y_i + \Delta t f(t_i, y_i)$ and

$m = 2$: $y_{i+2} = y_{i+1} + \frac{\Delta t}{2}(3f_{i+1} + f_i)$.

**Example** (Adams-Moulton Method)**.** Let $r = 1$ and $q \in \mathbb{P}_m$ with

$$q(t_{i+j}) = f(t_{i+j}, y_{i+j})$$

for $j = 0, \dots, m$. We obtain the methods

$m = 1$: $y_{i+1} = y_i + \frac{\Delta t}{2}(f_{i+1} + f_i)$,

$m = 2$: $y_{i+2} = y_{i+1} + \frac{\Delta t}{12}(5f_{i+2} + 8f_{i+1} + f_i)$ and

$m = 3$: $y_{i+3} = y_{i+2} + \frac{\Delta t}{24}(9f_{i+3} + 19f_{i+2} - 5f_{i+1} + f_i)$.

**Definition II.24** (Multistep Method)**.** A method for solving the IVP (II.1) and (II.2) of the form

$$\sum_{j=0}^{m} \alpha_j y_{i+j} = \Delta t \Phi(t_i, y_i, \dots, y_{i+m}, \Delta t)$$

with coefficients $\alpha_j \in \mathbb{R},\ j = 0, \dots, m,\ \alpha_m \neq 0$ and given starting values $y_0, \dots, y_{m-1}$ at time steps $t_0, \dots, t_{m-1}$ and incremental function $\Phi : [a, b] \times \mathbb{R}^m \times \mathbb{R}^+ \to \mathbb{R}$ is a **$m$-step/multistep method**. We call it **explicit**, if $\Phi$ does not depend on $y_{i+m}$ and **implicit** otherwise.

**Remark.**

- In the following, we confine ourself to linear multistep methods, i.e. to the case

$$\Phi(t_i, y_i, \dots, y_{i+m}, \Delta t) = \sum_{j=0}^{m} \beta_j f(t_{i+j}, y_{i+j})$$

and claim $|\alpha_0| + |\beta_0| > 0$ such that the data at $t_i$ are used for computing $y_{i+m}$.

- We need more than one starting values, however often only $y_0$ is given. We thus have to compute the $m-1$ starting values. This is called *starting* of *initialization phase*.

**Definition II.25.** A multistep method is called **consistent of order** $p \in \mathbb{N}$ according to the ODE II.1 if for a solution $y$ the **local discretization error**

$$\eta(t, \Delta t) = \sum_{j=0}^{m} \alpha_j y(t + j\Delta t) \; - \Delta t \Phi\Big(t, y(t), y(t + \Delta t), \ldots, y(t + m\Delta t), \Delta t\Big)$$

for $t \in [a, b]$ and $0 < \Delta t < \frac{b-a}{m}$ meets

$$\eta(t, \Delta t) = O(\Delta t^{p+1}), \;\; \Delta t \to 1.$$

In the case of $p = 1$ we call it **consistent**.

**Theorem II.26.** A linear multistep method has order of consistency $p \in \mathbb{N}$ if and only if it holds

$$\sum_{j=0}^{m} \alpha_j = 0 \text{ and } \sum_{j=0}^{m} a_j j^q = q \cdot \sum_{j=0}^{m} \beta_j j^{q-1} \tag{II.25}$$

for $q = 1, \ldots, p$.

*Proof.* The expansions

$$y(t + j\Delta t) = \sum_{q=0}^{p} \frac{(j\Delta t)^q}{q!} y^{(q)}(t) \; + O(\Delta t^{p+1})$$

and

$$y'(t + j\Delta t) = \sum_{q=1}^{p} \frac{(j\Delta t)^{q-1}}{(q-1)!} y^{(q)}(t) \; + O(\Delta t^{p+1})$$

gives

$$\begin{aligned}
\eta(t, \Delta t) &= \sum_{j=0}^{m} \Big(\alpha_j y(t + j\Delta t) - \Delta t \beta_j f\big(t + j\Delta t, y(t + j\Delta t)\big)\Big) \\
&= \sum_{j=0}^{m} \Big(\alpha_j y(t + j\Delta t) - \Delta t \beta_j y'(t + j\Delta t)\Big) \\
&= \sum_{j=0}^{m} \left(\alpha_j \left(\sum_{q=0}^{p} \frac{(j\Delta t)^q}{q!} y^{(q)}(t)\right) - \Delta t \beta_j \left(\sum_{q=1}^{p} \frac{(j\Delta t)^q}{(q-1)!} y^{(q)}(t)\right)\right) \; + O(\Delta t^{p+1}) \\
&= \ldots
\end{aligned}$$

This ends the proof. □

For the further analysis of linear multistep methods, we need the first and second characteristic polynomials. They are given by

$$\varrho(\xi) = \sum_{j=0}^{m} \alpha_j \xi^j$$

and

$$\sigma(\xi) = \sum_{j=0}^{m} \beta_j \xi^j.$$

They are also called **generating polynmials**. We can thus rewrite the conditions (II.25) in theorem II.26 as

$$\varrho(1) = \sum_{j=1}^{m} \alpha_j, \quad \sigma(1) = \sum_{j=1}^{m} \beta_j, \quad \varrho'(\xi) = \sum_{j=1}^{m} j \alpha_j \xi^{j-1}.$$

For $q = p = 1$ we hence have to check

$$\varrho(1) = 0 \text{ and } \varrho'(1) = \sigma(1) \tag{II.26}$$

to obtain consistency.

**Definition II.27.** A multistep method with starting values $y_j = y(t_j) + O(\Delta t^p),\ \Delta t \to 0$ for $j = 0, \dots, m-1$ is convergent of order $p \in \mathbb{N}$ with respect to (II.1) and (II.2) if for a time step $\Delta t$ the produced approximation $y_i$ of $y(t_i)$ for $t_i = a + i\Delta t,\ t_i \in [a, b]$ the **global error**

$$e(t_i, \Delta t) = y(t_i) - y_i$$

for all $t_i$ meets the condition

$$e(t_i, \Delta t) = O(\Delta t^p),\ \Delta t \to 0.$$

If the above conditions also hold for $o(1)$ instead of $O(\Delta t^p)$, we call the method **convergent**.

In case of $\beta_j = 0$ for all $j$, we have $\Phi = 0$ and thus, the linear multistep method becomes a linear **homogeneous difference equation**, i.e.

$$\sum_{j=0}^{m} \alpha_j y_{i+j} = 0. \tag{II.27}$$

**Theorem II.28.** If the first characteristic polynomial

$$\varrho(\xi) = \sum_{j=0}^{m} \alpha_j \xi^i \tag{II.28}$$

of (II.27) has only pairwise different roots $\xi_1, \dots, x_m \in \mathbb{C}$, then the solution sequence $(y_n)_{n\in\mathbb{N}}$ of (II.27) is of the form

$$y_n = \sum_{k=1}^{m} \gamma_k \xi_k^n, \tag{II.29}$$

where $\gamma_k \in \mathbb{C}$ for all $k$.

*Proof.* We only give a sketch. We need to show

(a) (II.29) is a solution of (II.27) and

(b) the solution space has dimension $m$ and $(\xi_k^n)_{n\in\mathbb{N}}$ is a basis.

ad (a). By plugging in (II.29) we obtain

$$\sum_{j=0}^{m} \alpha_j y_{i+j} = \sum_{j=0}^{m} \alpha_j \sum_{k=1}^{m} \gamma_k \xi_k^{i+j} = \sum_{k=1}^{m} \gamma_k \xi_k^i \cdot \underbrace{\sum_{j=0}^{m} \alpha_j \xi_k^j}_{=\varrho(\xi_k)=0} = 0.$$

ad (b). The main idea is to consider that each solution sequence $(y_n)_{n\in\mathbb{N}}$ needs starting values $y_0, \dots, y_{m-1} \in \mathbb{C}$. This means the set of starting vectors $\left\{s^{(1)}, \dots, s^{(m)}\right\}$ is a basis of $\mathbb{C}^m$. Each starting vector $s = \begin{bmatrix} y_0 & \dots & y_{m-1} \end{bmatrix}^T$ can be written as

$$s = \sum_{j=1}^{m} \gamma_j s^{(j)}.$$

Then, one can show that out of $(s_n)_{n\in\mathbb{N}}$ it is possible to build $(y_n)_{n\in\mathbb{N}}$.[7]

[7] This is the long part of the proof.

□

**Example.** We consider the ODE

$$y'(t) = 0,\ y(0) = 0.1$$

and apply

$$y_{i+2} + 4y_{i+1} - 5y_i = \Delta t\big(4f(t_{i+1}, y_{i+1}) + 2f(t_i, y_i)\big). \tag{II.30}$$

This method is consistent of order 3 with respect to $y'(t) = f(t, y(t))$. Applying it to (II.30) gives

$$y_{i+2} + 4y_{i+1} - 5y_i = 0.$$

According to theorem II.28 we have

$$y_n = \gamma_1 \xi_1^n + \gamma_2 \xi_2^n, \tag{II.31}$$

where we have determined $\gamma_1, \gamma_2$ by the initial conditions and $\xi_1$ and $\xi_2$ are the roots of the first characteristic polynomial. We have

$$\varrho(\xi) = \xi^2 + 4\xi - 5 = (\xi - 1)(\xi + 5),$$

so it holds $\xi_1 = 1$ and $\xi_2 = -5$. Setting

$$y_0 = 0.1 \text{ and } y_1 = 0.1 + \varepsilon,$$

where $\varepsilon > 0$ is a small perturbation, with (II.31) we obtain

$$0.1 = \gamma_1 + \gamma_2 \text{ and } 0.1 + \varepsilon = \gamma_1 - 5\gamma_2,$$

so it holds

$$\gamma_1 = 0.1 + \frac{\varepsilon}{6} \text{ and } \gamma_2 = -\frac{\varepsilon}{6}.$$

Hence

$$y_n = 0.1 + \frac{\varepsilon}{6} - \frac{\varepsilon}{6}(-5)^n.$$

The solution of (II.30) is $y(t) = 0.1$. For a fixed $T$ and $\Delta t = \frac{T}{n}$, where $n \in \mathbb{N}$, we have

$$\lim_{n\to\infty} |y(T) - y_n| = \lim_{n\to\infty} |\frac{\varepsilon}{6} - \frac{\varepsilon}{6}(-5)^n| = \infty.$$

Even for $y_0 = y_1 = 0.1$ we end up with $|y_{50}| \approx 6.5 \cdot 10^{16}$.

**Definition II.29.** A multistep method

$$\sum_{j=0}^{m} \alpha_j y_{i+j} = \Delta t \Phi(t, y_i, \ldots, y_{i+m}, \Delta t)$$

is called **zero stable**, if the corresponding first characteristic polynomial

$$\varrho(\xi) = \sum_{j=0}^{m} \alpha_j \xi^j$$

meets the *Dahlquist root condition*, i.e. all roots of the polynomials lie inside of the closed complex unit circle and the roots on the boundary are simple.

**Remark.** In the beginning of theorem II.26 we had $\varrho(1) = 0$. This is why we want the roots on the boundary when they are simple.

**Theorem II.30.** A convergent linear multistep method is necessarily consistent and zero stable.

*Proof.* We will show

(i) zero stability (by contradiction) and

(ii) consistency $\varrho(1) = 0$ and $\varrho'(1) = \sigma(1)$.

ad (i): Let us assume, that the method is convergent but not zero stable. Apply the linear multistep method on

$$y'(t) = 0,\ y(0) = 0,\ t \in [0,1]$$

with solution $y(t) = 0$. For $i = 0, 1, \ldots$ we have

$$\alpha_m y_{i+m} + \cdots + \alpha_0 y_0 = 0. \tag{II.32}$$

Since the method is **not** zero stable, we have either a simple root $\xi_1$ with $|\xi_1| > 1$ or a multiple root $\xi_2$ with $|\xi_2| = 1$. Our goal is to generate initial values (meeting the definition for convergence) and show that (II.32) diverges.

For the case of $|\xi_1| > 1$, we know that according to theorem II.28 a solution of (II.32) is given as

$$y_n = \sqrt{\Delta t}\ \xi_1^n.$$

The first $m$ elements meet

$$\lim_{\Delta t \to 0} |y_j - y(t_j)| = \lim_{\Delta t \to 0} \left| \sqrt{\Delta t}\ \xi_1^j \right| = 0$$

for $j = 0, \ldots, m-1$ and thus the definition for the convergence. For some fixed $T = n\Delta t \in (0,1)$ with $n > m$ we obtain

$$\lim_{\Delta t \to 0} |y_n - y(T)| = \lim_{\Delta t \to 0} \left| \sqrt{\Delta t}\ \xi_1^{T/\Delta t} \right| = \infty,$$

which is a contradiction to the assumed convergence.

On the other hand, if $|\xi_2| = 1$, since $\xi_2$ is a multiple root it holds

$$0 = \varrho'(\xi_2) = \alpha_1 + 2\alpha_2\xi_2 + \cdots + m\alpha_m\xi_2^{m-1}.$$

We set

$$y_n = n \cdot \sqrt{\Delta t}\ \xi_2^{n-1},$$

such that

$$\sum_{j=0}^{m} \alpha_j y_{i+j} = \sum_{j=0}^{m} \alpha_j (i+j)\sqrt{\Delta t}\ \xi_2^{i+j-1} = \sqrt{\Delta t}\ i\xi_2^j \sum_{j=0}^{m} j\alpha_j \xi_2^{j-1} = 0$$

is a solution of the difference equation (II.32), where the starting values $y_0, \ldots, y_{m-1}$ with

$$\lim_{\Delta t \to 0} |y_j - y(t_j)| = \lim_{\Delta t \to 0} \left| j \cdot \sqrt{\Delta t} \right| = 0$$

meet the definition for convergence. For a fixed $T = n\Delta t \in (0,1)$ with $n \geq m$ we have

$$\lim_{\Delta t \to 0} |y_n - y(T)| = \lim_{\Delta t \to 0} |n\sqrt{\Delta t} \cdot 1| = \lim_{\Delta t \to 0} \left| \frac{T}{\sqrt{\Delta t}} \right| = \infty.$$

This again is a contradiction. Thus, zero stability is necessary for convergence.

ad (ii): For $\varrho(1) = 0$ consider the ODE

$$y'(t) = 0,\ y(0) = 1,\ t \in [0,1]$$

with exact solution $y = 1$. The corresponding difference equation is identical to the one before. For the starting values $y_0, \ldots, y_{m-1}$ and for $t = m \cdot \Delta t$ we have

$$\begin{aligned} 0 &= \lim_{\Delta t \to 0} \alpha_m (y(m\Delta t) - y_n) \\ &= \lim_{\Delta t \to 0} \alpha_m (1 - y_m) \\ &= \sum_{j=0}^{m} \alpha_j (1 - y_j) \\ &= \sum_{j=0}^{m} \alpha_j - \underbrace{\sum_{j=0}^{m} \alpha_j y_j}_{=0} = \varrho(1). \end{aligned}$$

For showing $\varrho'(1) = \sigma(1)$ we consider

$$y'(t) = 1,\ y(0) = 0,\ t \in [0, 1].$$

The exact solution is given by $y(t) = t$. This gives the inhomogeneous difference equation

$$\sum_{j=0}^{m} \alpha_j y_{i+j} = \Delta t \sum_{j=0}^{m} \beta_j$$

for all $i$. Since $\varrho(1) = 0$, we know that $\xi = 1$ is a simple root.[8] Hence, $\varrho'(1) \neq 0$ and we can set

$$M := \frac{\sigma(1)}{\varrho'(1)}$$

and

$$y_n := n \cdot \Delta t \cdot M$$

for all $n \in \mathbb{N}$. The first $m$ elements meet

$$\lim_{\Delta t \to 0} |y_j - y(j\Delta t)| = \lim_{\Delta t \to 0} |j \cdot \Delta t \cdot (m-1)| = 0$$

for $j = 0, \ldots, m-1$.[9] Further, it holds

$$\begin{aligned} \sum_{j=0}^{m} \alpha_j y_{i+j} - \Delta t \cdot \sum_{j=0}^{m} \beta_j &= \sum_{j=0}^{m} \alpha_j (i+j) \Delta t \underbrace{M}_{=\frac{\sigma(1)}{\varrho'(1)}} - \Delta t \cdot \underbrace{\sum_{j=0}^{m} \beta_j}_{=\sigma(1)} \\ &= \Delta t \cdot \left( \frac{\sigma(1)}{\varrho(1)} \left( i \cdot \sum_{j=0}^{m} \alpha_j + \sum_{j=0}^{m} j \alpha_j \right) - \sigma(1) \right) \\ &= 0 \end{aligned}$$

which means that $y_n = n \cdot \Delta t \cdot M$ is a solution of the difference equation. Since we have convergence for $T = n \cdot \Delta t \in (0, 1)$ fixed, it holds

$$0 = \lim_{\Delta t \to 0} |y_n - y(T)| = \lim_{\Delta t \to 0} |n \cdot \Delta t \cdot M - T| = \lim_{\Delta t \to 0} |T - MT| = |T(1 - M)|.$$

Hence, since $T$ is nonzero, $1 - M$ is zero, so $\sigma(1) = \varrho'(1)$. This ends the proof.

□

[8] This follows by the zero stability shown before.

[9] This means that the starting conditions are met by the first $m$ elements.

We now aim to show that

$$\text{zero stability } + \text{ consistency } \Rightarrow \text{ convergence.} \qquad (**)$$

This leads us to the next definition.

**Definition II.31.** A multistep method with incremental function $\Phi$ is **Lipschitz-continuous in** $(t, y(t))$ **with Lipschitz constant** $L > 0$, if there exists a neighbourhood $U$ of $(t, y(t))$ and a constant $H > 0$ such that

$$|\Phi(t_i, u_0, \ldots, u_m, \Delta t) - \Phi(t_i, v_0, \ldots, v_m, \Delta t)| \leq L \cdot \sum_{k=0}^{m} |u_k - v_k|$$

for all time step sizes $0 < \Delta t \leq H$ and all $(t, u_k), (t, v_k) \in U(t, y(t)),\ k = 0, \ldots, m$.

For the proof of our goal we need two technical lemma.

**Lemma II.32** (technical)**.** For $A \in \mathbb{C}^{n,n}$ the sequence of matrices $(A^k)_{k\in\mathbb{N}}$ is bounded if and only if the spectral radius $\varrho(A) \leq 1$ and for each eigenvalue $\lambda \in \mathbb{C}$ of $A$ with $|\lambda| = 1$ the algebraic and geometric multiplicity are the same.

*Proof.* We use the Jordan-decomposition $T^{-1}AT = J = \operatorname{diag}(J_1, \ldots, J_r)$.[10] In each $J_j$ with $m_j \geq 2$, we know that $|\lambda_j| < 1$. We now find a $\varepsilon > 0$ such that

$$|\lambda_j| + \varepsilon < 1.$$

an then we multiply $J$ by $D = \operatorname{diag}(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{k-1})$ and obtain

$$\tilde{J} = D^{-1}JD.$$

We arrive at the Jordan blocks

$$\tilde{J}_j = \begin{bmatrix} \lambda_j & \varepsilon & & 0 \\ & \ddots & \ddots & \\ & & \lambda_j & \varepsilon \\ 0 & & & \lambda_j \end{bmatrix}.$$

For $k \in \mathbb{N}$ it holds

$$\left\|\tilde{J}^k\right\| \leq \left\|\tilde{J}\right\|^k \leq \varrho(\tilde{J})^k \leq 1. \qquad (*)$$

Setting $S = D^{-1}T$, we have $A^k = (S\tilde{J}S^{-1})^k = (S\tilde{J}^kS^{-1})$. Using $(*)$ we have

$$\left\|A^k\right\|_\infty \leq \|S\|_\infty \left\|S^{-1}\right\|_\infty \left\|\tilde{J}^k\right\|_\infty \leq \kappa_\infty(S) < \infty$$

for all $k \in \mathbb{N}$.

The reverse direction is much shorter and is an exercise. □

**Lemma II.33** (discrete Gronwall-Lemma)**.** Let $\Delta t_0, \ldots, \Delta t_{r-1} \in \mathbb{R}_{>0}$ and $\delta, \gamma \in \mathbb{R}_{\geq 0}$ be given. If the numbers $e_0, \ldots, e_r \in \mathbb{R}$ meet

$$|e_0| \leq \delta \text{ and } |e_\ell| \leq \delta + \gamma \sum_{j=1}^{\ell-1} \Delta t_j\ |e_j|$$

for $\ell = 1, \ldots, r$, then it holds

$$|e_\ell| \leq \delta \cdot \exp\left(\gamma \sum_{j=0}^{\ell-1} \Delta t_j\right)$$

for $l = 0, \ldots, r$.

[10] It states that for each matrix $A \in \mathbb{C}^{n,n}$ there exists a nonsingular transformation $T \in \mathbb{C}^{n,n}$ such that $T^{-1}AT = J$, where $J = \operatorname{diag}(J_1, \ldots, J_r)$, where $J_i$ a is a Jordan block.

*Proof.* No proof, since the proof is similar to the known Gronwall-Lemma. □

**Theorem II.34.** Let the multistep method

$$\sum_{j=0}^{m} \alpha_j y_{i+j} = \Delta t \Phi(t_i, y_i, \dots, y_{i+m}, \Delta t) \tag{II.33}$$

be Lipschitz and zero stable. Then there exists a stepsize bound $H > 0$: $0 < \Delta t = \frac{b-a}{n} \leq H$ such that

$$\max_{j=0,\dots,n} |e(t_j, \Delta t)| \leq K \left( \max_{k=0,\dots,m-1} |e(t_k, \Delta t)| + \max_{a \leq t \leq b - m\Delta t} \frac{\eta(t, \Delta t)}{\Delta t} \right) \tag{II.34}$$

with a consistent $K > 0$ being independent on the Lipschitz-constant.

*Proof.* For the left hand side of (II.34), let us define

$$e := \begin{bmatrix} e_0 & \dots & e_n \end{bmatrix}^T \in \mathbb{R}^{n+1},$$ [11]

where $e_i = e(t_i, \Delta t),\ i = 0, \dots, m$. Furthermore, for each $j = 0, \dots, n - (m-1)$ we define

$$e_j := \begin{bmatrix} e_j & \dots & e_{j+m-1} \end{bmatrix}^T \in \mathbb{R}^m$$

such that we can express the total error as

$$\|e\|_\infty = \dots$$

Furthermore, we set

$$\eta_i = \eta(t_i, \Delta t),\ i = 0, 1, \dots$$

and obtain

$$\begin{aligned} \sum_{j=0}^{m} \alpha_j e_{i+j} &= \sum_{j=0}^{m} \alpha_j \left( y(t_{i+j}) - y_{i+j} \right) \\ &= \Delta t \underbrace{\left( \Phi\Big(t_i, y(t_i), \dots, y(t_{i+m}), \Delta t\Big) - \Phi\Big(t_i, y_i, \dots, y_{i+m}, \Delta t\Big) \right)}_{=:\mu_i} + \eta_i. \end{aligned} \tag{II.35}$$

Without loss of generality, we set $\alpha_m = 1$. With (II.35) we obtain

$$e_{i+m} = \sum_{j=0}^{m-1} \alpha_j e_j + \mu_i + \eta_i.$$

Finally, we arrive at

$$\begin{bmatrix} e_{i+1} \\ \vdots \\ e_{i+m} \end{bmatrix} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ -\alpha_0 & \dots & \dots & -\alpha_{m-1} \end{bmatrix} \begin{bmatrix} e_i \\ \vdots \\ \vdots \\ e_{i+m-1} \end{bmatrix} + \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mu_i + \eta_i \end{bmatrix},$$

or

$$e_{i+1} = A e_i + f_i.$$

We thus can rewrite

$$e_i = A e_{i-1} f_{i-1} = \dots = A^i e_0 + \sum_{k=0}^{i-1} A^{i-1-k} f_k. \tag{II.36}$$

[11] Note that $e$ is a vector.

This means that we can express the error vector $e_i$ in terms of the initialization error. One can show that

$$p(\lambda) = \det(A - \lambda I) = -\sum_{j=0}^{m-1} \alpha_j \lambda^j = -\varrho(\lambda).$$

Due to the zero stability we know that for all eigenvalues $\lambda$ it holds $|\lambda| \leq 1$ and the eigenvalues with $|\lambda| = 1$ are simple. We can thus apply lemma II.32 such that $(A^k)_{k\in\mathbb{N}}$ is bounded, i.e.

$$\|A^p\|_\infty \leq C$$

for all $p \in \mathbb{N}$ with a $C \geq 1$. We thus rewrite (II.36) as

$$\|e_i\|_\infty \leq C \left( \|e_0\|_\infty + \sum_{k=0}^{i-1} \|f_k\|_\infty \right). \tag{II.37}$$

We still need to handle $\|f_k\|_\infty$. Our idea is to use the Lipschitz condition

$$\|\mu_k\| \leq \Delta t \cdot L \cdot \sum_{j=0}^{m} |y(t_{k+j}) - y_{k+j}| = \Delta t \cdot L \cdot \sum_{j=0}^{m} |e_{k+j}|.$$

Using that, we obtain

$$\begin{aligned}
\|f_k\|_\infty &= |\mu_k + \eta_k| \\
&\leq |\mu_k| + |\eta_k| \\
&\leq |\eta_k| + \Delta t \cdot L \cdot \sum_{j=0}^{m} |e_{k+j}| \\
&\leq n \cdot \max_{j=0,\ldots,n-m} |\eta_j| + \Delta t \cdot L \cdot m \cdot \|e_k\|_\infty + \Delta t \|e_{k+1}\|_\infty .
\end{aligned}$$

It further holds

$$\sum_{k=0}^{i-1} \|f_k\|_\infty \leq n \cdot \max_{j=0,\ldots,n-m} |\eta_j| + \Delta t \cdot L(m+1) \cdot \sum_{k=0}^{i-1} \|e_k\|_\infty + \Delta t \cdot L \cdot \|e_i\|_\infty .$$

Inserting these two results into (II.37) gives

$$\begin{aligned}
&\|e_i\|_\infty \\
&\leq C \left( \|e_0\|_\infty + n \cdot \max_{j=0,\ldots,n-m} |\eta_j| + \Delta t \cdot (m+1) \cdot L \sum_{k=0}^{i-1} \|e_k\|_\infty + \Delta t \cdot L \cdot \|e_i\|_\infty \right).
\end{aligned}$$

We thus have

$$(1 - C \cdot \Delta t \cdot L) \|e_i\|_\infty \leq C \left( \|e_0\|_\infty + n \cdot \max_{j=0,\ldots,n-m} |\eta_j| + \Delta t \cdot (m+1) \cdot L \sum_{k=0}^{i-1} \|e_k\|_\infty \right).$$

Since we bound the time step $\Delta t$ by a constant $H < \frac{1}{CL}$, we have

$$(1 - \Delta t \cdot C \cdot L) \geq 1 - HCL > 0.$$

We thus have

$$\|e_i\|_\infty \leq \delta + \gamma \cdot \sum_{k=0}^{i-1} \Delta t \|e_k\|_\infty$$

for $i = 0, \ldots, n-m$, where

$$\delta := \frac{C}{1 - HCL} \left( \|e_0\|_\infty + n \cdot \max_j |\eta_j| \right) \geq 0$$

and

$$\gamma := \frac{(m+1)\cdot L\cdot C}{1-HCL} \geq 0.$$

With Gronwall's lemma (lemma II.33), we obtain

$$\|e_i\|_\infty \leq \delta \cdot \exp\left(\gamma \cdot \sum_{k=0}^{i-1} \Delta t\right) \leq \delta \cdot \exp(\gamma(b-a)).$$

Setting

$$\tilde{k} := \frac{C}{1-HCL} \cdot \exp(\gamma(b-a)) \text{ and } n = \frac{b-a}{\Delta t},$$

we arrive at

$$\begin{aligned} |e_i| \leq \|e_i\|_\infty &\leq \tilde{k}\left(\|e_0\|_\infty + n \cdot \max_{j=0,\ldots,n-m} |\eta_j|\right) \\ &\leq \tilde{k}\left(\|e_0\|_\infty + n \cdot \max_{a\leq t\leq b-m\Delta t} |\eta(t,\Delta t)|\right). \end{aligned}$$

Thus, it holds

$$|e_i| \leq k \cdot \left(\max_{j=0,\ldots,m-1} |y(t_i)-y_i| + n \cdot \max_{a\leq t\leq b-m\Delta t} \frac{n(t,\Delta t)}{\Delta t}\right)$$

for $i = 0,\ldots,n-(m-1)$ with

$$k := \max\left\{\tilde{k}, \tilde{k}\cdot(b-a)\right\} \geq 0.$$

Because of

$$\max_{j=n-(m-1),\ldots,n} |y(t_j)-y_j| = \max_{n-(m-1),\ldots,n} |e_j| \leq \left\|e_{n-(m-1)}\right\|_\infty$$

we obtain

$$\max_{i=0,\ldots,n} |y(t_i)-y_i| \leq k\left(\max_{k=0,\ldots,m-1} |e(t_k,\Delta t)| + \max_{a\leq t\leq b-m\Delta t} \frac{|\eta(t,\Delta t)|}{\Delta t}\right),$$

what ends the proof. □

**Remark.**

- It holds $e(t_i,\Delta t) = y(t_i) - y_i$.
- Theorem II.34 in short states, that for a Lipschitz-continuous multistep method it holds $(**)$.
- Starting values need at least the order of consistency of the method.
- Compare $\frac{\eta(t_i,\Delta t)}{\Delta t}$ with the consistency definition.

### BDF Methods (Backwards Differentiation Formula Methods)

**Definition II.35** (BDF-scheme)**.** For $t_{i+j} = t_i + j\Delta t$, $j = 0,\ldots,m$ and nodes $(t_i,y_i),\ldots,(t_{i+m-1},y_{i+m-1}) \in \mathbb{R}^2$ let $p \in \mathbb{P}_m$ be given via

$$p(t_{i+j}) = y_{i+j} \text{ and } p'(t_{i+m}) = f(t_{i+m}, p(t_{i+m})) \tag{II.38}$$

for $j = 0,\ldots,m-1$. Then, the BDF scheme computes the approximation $y_{i+m}$ at $t_{i+m}$ by

$$y_{i+m} = p(t_{i+m}). \tag{II.39}$$

**Example** (BDF(2)-method)**.** For given two values $y_i, y_{i+1}$ we consider $p \in \mathbb{P}_2$ such that

$$p(t_i) = y_i,\ p(t_{i+1}) = y_{i+1} \text{ and } p'(t_{i+2}) = f(t_{i+2}, p(t_{i+2})). \qquad (*)$$

Using an expansion, we get the equations

$$p(t_{i+1}) = p(t_{i+2}) + \Delta t p'(t_{i+2}) + \frac{\Delta t^2}{2} p''(t_{i+2})$$

and

$$p(t_i) = p(t_{i+2}) + 2\Delta t p'(t_{i+2}) + 2\Delta t^2 p''(t_{i+2}).$$

Substracting the first equation from the second four times, we obtain

$$p(t_i) - 4p(t_{i+1}) = -3p(t_i) - 2\Delta t p'(t_{i+2}) + 0.$$

The conditions $(*)$ give

$$\frac{y_i - 4y_{i+2} + 3y_{i+2}}{2\Delta t} = p'(t_{i+2}) = f(t_{i+2}, y_{i+2}).$$

Rearranging leads to

$$\begin{aligned} \frac{3y_{i+2} - 4y_{i+1} + y_i}{2\Delta t} &= \frac{3p(t_{i+2}) - 4p(t_{i+1}) + p(t_i)}{2\Delta t} \\ &= p'(t_{i+2}) \\ &= f(t_{i+2}, p(t_{i+2})) \approx y'(t_{i+2}) \end{aligned}$$

**Remark.**

- Due to (II.38) and (II.39) all BDF methods are implicit.
- It can be shown, that BDF(2) is zero stable and consistent of order $p = 2$ and also convergent of order $p = 2$. The proof relies on Taylor expansion.

**Stiff ODEs**

**Example** (mass-spring system)**.** Let $m = 25\text{kg}$, spring sonstant $D = 5000\text{N/m}$, damping $\sigma = 4000\text{kg/s}$ and consider

$$mu''(t) + \sigma u'(t) + Du(t) = 0,\ u(0) = 0.05\text{m},\ u'(0) = 10\text{m/s}.$$

The roots of the characteristic polynomial are given by

$$\lambda_{1,2} = -80 \pm \sqrt{6200},$$

i.e.

$$\lambda_1 \approx -158.74 \text{ and } \lambda_2 \approx -1.26.$$

The solution is approximated by

$$u(t) \approx -0.064 \cdot \exp(-158.74t) + 0.114 \exp(-1.26t).^{12}$$

Remember that we wanted $|R(\Delta t \cdot \lambda)| < 1$. But since the first summand is very small, the second summand is dominant and the first summand is almos negligible. Regarding $\lambda_2$, the stability is easy to "buy" by choosing $\Delta t = 0.5$, but for $\lambda_1$ we need a much smaller $\Delta t$, despite the fact that its influence is much smaller.

The dominance of the $\lambda_2$ in combination with the small stepsize needed for the $\lambda_1$ part is our problem. Due to stability reasons, the whole interval needs fine discretization. Hence, the eigenvalue with the largest absolute value is important for the stability.

[12] We ignore the units.

In the literature, one can find different measures for the stiffness. A popular one is

$$S := \frac{\max_j |\mathrm{Re}(\lambda_j)|}{\min_j |\mathrm{Re}(\lambda_j)|}.$$

In our example, we have $S \approx 100$.

Solving our problem in `MATLAB` needs 1957 time steps when using `ode45`[13] and 72 time steps when using `ode23s`. The reason behind this is that `ode23s` uses a more stable method, the *Rosenbrock method*.

**Rosenbrock-Method**

Consider

$$y' = f(y),\ J := f'(y_n),\ y' = f(f(y_n) + J(y - y_n)) + (f(y) - f(y_n)) - J(y - y_n),$$

i.e. a linearization of $f$ at $y_n$ and assume

$$y' = Jy(t) + f(y(t)) - Jy(t).$$

We consider the slopes $r_j$[14]

$$r_j = J\left(y + \Delta t \sum_{\nu=1}^{j} \tilde{a}_{j\nu} r_\nu\right) + \left(f\left(y + \Delta t \sum_{\nu=}^{j-1} a_{j\nu} r_\nu\right) - J\left(y + \Delta t \sum_{\nu=1}^{j-1} a_{j\nu} r_\nu\right)\right).$$

Hence, $r_j$ appears only linear on the right hand side. Further, $r_1, \ldots, r_{j-1}$ appear only in the nonlinear part of $f$. This means that we can compute $r_j$ via a linear system:

(i) Compute $J = f(y_n)$.

(ii) Compute

$$(I - \Delta t \tilde{a}_{jj} J) r_j = \Delta t \sum_{\nu=1}^{j-1} (\tilde{a}_{j\nu} - a_{j\nu})\, J \cdot r_0 \ + \ f\left(y_n + \Delta t \sum_{\nu=1}^{j-1} a_{j\nu} r_\nu\right).$$

(iii) Compute

$$y_{n+1} = y_n + \Delta t \sum_{j=1}^{s} b_j r_j.$$

[13] This is a lot.

[14] Compare with the Runge-Kutta-scheme.

# III Numerics of Boundary Value Problems (BVPs)

In IVPs, we had initial conditions. In BVPs, there are conditions on the solution, usually on the boundary of $I = [x_0, x_{\text{end}}] \subseteq \mathbb{R}$.

# Numerical Mathematics II: Boundary Value Problem

Konstantin Fackeldey

TU Berlin

June 17, 2019



## Definition

In the following we consider a
**non-linear two point boundary value problem (BCP)**:

$$y'(x) = f(x, y(x))$$
$$R(y(a), y(b)) = 0$$

with $R : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$



## Different kinds of BVP

- **higher order BVP**

$$y^{(m)}(x) = f(t, y(x), ..., y^{(m-1)}(x))$$
$$R(y(a), y(b)) = 0$$

- **linear BVP**

$$y'(x) = A(x)y(x) + q(x); \quad a \le x \le b$$
$$R(y(a), y(b)) = B_a y(a) + B_b y(b) - c = 0, \qquad B_a, B_b \in \mathbb{R}^{n \times n}$$

- **BVP with separate boundary values**

$$R_1(y(a)) = 0, \quad R_1 : \mathbb{R}^n \to \mathbb{R}^p$$
$$R_2(y(b)) = 0, \quad R_2 : \mathbb{R}^n \to \mathbb{R}^q, \qquad p + q = n$$



## Different kinds of BVP (cont'd)

- **free BVPs**

$$y'(x) = f(x, y(x)), \quad a \le x \le b, \; b \text{ free}$$
$$R(y(a), y(b)) = 0$$

- **multiple BCs**

$$y'(x) = f(x, y(x)), y \in \mathbb{R}^n \quad a \le x \le b$$
$$0 = R(y(x_0), ..., y(x_s))$$
where $a = x_0 < x_1 < ... < x_s = b$



## Picard Lindelöf for BVPs ?

$$y'(x) = f(x, y(x)), \quad R(y(a), y(b)) = 0 \qquad \textbf{(III.1)}$$

**Definition (III.1)**

A sloution $y(x)$ of the BVP (III.1) is **local unique**, if there ex itst a neighbourhood around $y(x)$, such that $y(x)$ is the only solution, i.e.

$$\exists \delta > 0 : \forall u(x) \not\equiv y(x) : \{\|u - y\| \ge \delta \vee u \text{ is no soution of } \textbf{(III.1)}\}$$



## Isolated Solution & Variational Problem

$$y'(x) = f(x, y(x)), \quad R(y(a), y(b)) = 0 \qquad \textbf{(III.1)}$$

**Definition (III.2)**

A sloution $y(x)$ of the BVP (III.1) is **isolated**, if the *variational problem*

$$z' = \frac{\partial f(x, y(x))}{\partial y} z \qquad \textbf{(III.2)}$$
$$\frac{\partial R(y(a), y(b))}{\partial y(a)} z(a) + \frac{\partial R(y(a), y(b))}{\partial y(b)} z(b) = 0 \qquad \textbf{(III.3)}$$

has the unique solution $z(x) \equiv 0$.

## Theorem (III.3)

*Let $f \in C^2$, such that (III.2) is well defined. If then $y$ is an isolated solution of (III.1), then it is also unique.*

reminder:

## Definition (III.2)

A sloution $y(t)$ of the BVP (III.1) is **isolated**, if the *variational problem* (**III.2+3**)

$$z' = \frac{\partial f(t, y(x)}{\partial y} z$$
$$\frac{\partial R(y(a), y(b))}{\partial y(a)} z(a) + \frac{\partial R(y(a), y(b))}{\partial y(b)} z(b) = 0$$

has the unique solution $z(x) \equiv 0$.



need to show:

$$z' = \frac{\partial f(x, y(x)}{\partial y} z$$
$$\underbrace{\frac{\partial R(y(a), y(b))}{\partial y(a)}}_{:=B_a} z(a) + \underbrace{\frac{\partial R(y(a), y(b))}{\partial y(b)}}_{:=B_b} z(b) = 0$$

has a unique solution.



little more general:

## Theorem (III.4)

*Let $A$ and $q$ be continuous on $[a, b]$. The linear BVP*

$$y'(x) = A(x)y(x) + q(x); \quad a \le x \le b$$
$$R(y(a), y(b)) = B_a y(a) + B_b y(b) - c = 0,$$

*has a unique solution, iff*

$$Q := B_a Y(a) + B_b Y(b)$$

*is non singular, where $Y(x)$ is a fundamental matrix of the homogeneous ODE*

$$\mathbf{y}'(x) = \mathbf{A}(x)\mathbf{y}(x)$$

**Example.** Consider the ODE $-u''(x) = u(x)$ with boundary conditions $u(0) = 0, u(\pi/2) = 1$. Then the unique solution is given by $u(x) = \sin(x)$. When choosing the boundary conditions $u(0) = 0, u(\pi) = 0$, we get infinitely many solutions of the form $u(x) = c_1 \cdot \sin(x)$. Furthermore, choosing $u(0) = 0, u(\pi) = 1$ there is no solution.

In general, for each BVP

$$y' = f(x, y),\ R(y(a), y(b)) = 0$$

we can give an IVP

$$y' = f(x, y),\ y(a) = s$$

with still unknown parameter $s$. If these initial conditions give unique solutions

$$w(x, s) := y(x; a, s),$$

then there exists a unique function

$$F(x) = R(s, w(b, s)).$$

Thus, the solution is uniquely solvable if $F(s) = 0$ has exactly one solution.

Let $(u_i)_{i\in\mathbb{N}}$ be a sequence of functions such that

$$\|u_i - y\|_\infty = \delta_i,\ \delta_i \to 0$$

and instead of $y(a) = s^*$ we have $u_i(a) = s^* + \alpha_i$ with $\alpha_i \to 0$ and

$$u_i(b) = w(b, s^*) + \beta_j,\ \beta_j \to 0.$$

In other words, $(u_i)_{i\in\mathbb{N}}$ is a sequence of functions approaching the solution. We use the expansion

$$\frac{F(s^* + \alpha_i)}{\|\alpha_i\|} = \big(F(s^*) + F_s(s^*)\alpha_i + O(\alpha_i^2)\big) \cdot \frac{1}{\|\alpha_i\|}.$$

Since $F(s^*) = 0$, we only need to consider the part $F_s(s^*) \cdot \alpha_i$. In other words,

$$\frac{F(s^* + \alpha_i)}{\|\alpha_i\|} = 0 \Rightarrow F_s(s^*)\alpha_i \to 0.$$

If $F_s$ is continuous, then $F_s(s^*) \cdot \alpha = 0$. But since

$$F(s) = R(s, w(b, s))$$

it holds

$$F_s(s^*) \cdot \alpha = \frac{\partial R(s, w(b, s))}{\partial y(a)}\alpha + \frac{\partial R(s, w(b, s))}{\partial y(b)} \cdot \frac{\partial w(b, s)}{\partial s} \cdot \alpha.$$

If the solution is not local unique, then $F_s$ is singular and there exists a direction $\alpha$ for varying the $s^*$ such that

$$\frac{\partial R(s, w(b, s))}{\partial y(a)} \cdot \alpha + \frac{\partial R(s, w(b, s))}{\partial y(b)} \cdot \underbrace{\frac{\partial w(b, s)}{\partial s}}_{=:\beta} \cdot \alpha = 0.$$

## III.1 Finite Elements

We consider

$$\begin{aligned} -u''(x) + b(x)u'(x) + c(x)u(x) &= f(x) \\ 0 < x < 1,\ u(0) &= u(1) = 0 \end{aligned} \tag{III.4}$$

and expect that there exists a solution

$$u \in C^2(0, 1) \cap C^0([0, 1]).$$

Such a solution is called *classical solution*.

We seek for a minimization problem, but it can happen that we can find a suitable minimization, but it has solution or it has only a solution by assuming that the solution is suitable smooth.

Let

$$X := \left\{ v \in C^1(0,1) \cap C^0([0,1]) \mid v(0) = v(1) = 0 \right\}.$$

Multiplying (III.4) by some $v \in X$ gives

$$\int_0^1 (-u'' + bu' + cu)\, v \, dx = \int_0^1 fv \, dx.$$

With integration by parts on $-u''v$ gives

$$\int_0^1 \underbrace{-u''v + bu'v + cuv}_{=:a(u,v)} \, dx = \underbrace{-u'(x)v(x)\Big|_0^1}_{=0} + \int_0^1 u'v' + bu'v + cuv \, dx = \underbrace{\int_0^1 fv \, dx}_{=:f(v)=\langle f,v\rangle}$$

for all $v \in X$.

**Definition III.5.** The closure (*Abschluss*) of a subset $A$ of $V$ with respect to $V$ is denoted by $\overline{A}^V$. By

$$\operatorname{supp} v := \overline{\{X \in (0,1) \mid v(x) \neq 0\}}^{\mathbb{R}}$$

we define the **support**.

**Remark.**

- For an example, let $(X, \|\cdot\|_X)$ be some normed space and $A \subseteq X$. Then $x \in \overline{A}^{\|\cdot\|_X}$ if there exists $(x_n)_{n\in\mathbb{N}} \subseteq A$ such that

$$\lim_{n\to\infty} \|x_n - x\|_X = 0.$$

- We are interested in functions with small support.

**Definition III.6.** We define

$$C_0^\infty(0,1) := \{v \in C^\infty \mid \operatorname{supp} v \subseteq (0,1)\}$$

and

$$L^2(0,1) := \left\{ v : [0,1] \to \mathbb{R} \mid \int_0^1 v^2(x) \, dx < \infty \right\}$$

and equip $L^2(0,1)$ with the norm

$$\|u\|_{L^2(0,1)}^2 = \int_0^1 u^2(x) \, dx$$

and inner product (*Skalarprodukt*)

$$\langle u, v\rangle_{L^2(0,1)} = \int_0^1 uv \, dx$$

**Remark.**

The sets $L^p(\Omega)$ (equipped with the respective norm) are for $1 \leq p \leq \infty$ are separable Banach spaces.

**Definition III.7.** We call $w \in L^2(0,1)$ the **weak derivative** of $u \in L^2(0,1)$ if for all $v \in C_0^\infty(0,1)$ it holds

$$\int_0^1 wv \, dx = -\int_0^1 uv' \, dx.$$

In this case, we write $w = u'$.

**Definition III.8.** We call the space

$$H^1(0,1) := \left\{ v \in L^2(0,1) \mid \exists v' \in L^2(0,1) \right\}$$

**sobolev space of the functions on** $(0,1)$. Further, we set

$$H_0^1(0,1) = \overline{C_0^\infty(0,1)}^{H^1(0,1)}.$$

But what does $\overline{C_0^\infty(0,1)}^{H^1(0,1)}$ mean?

**Remark.**

(1) The set $H_0^1(0,1)$ is the closure of $C_0^\infty(0,1)$ with respect to $H^1(0,1)$.

(2) It holds $u \in H_0^1(0,1)$ if and only if there exists a sequence of functions $(u_n)_{n\in\mathbb{N}} \in C_0^1(0,1)$ such that

$$\|u - u_n\|_{H^1(0,1)} \overset{n\to\infty}{\longrightarrow} 0$$

**Definition III.9.** The solution of the problem

"Find $u \in H_0^1(0,1)$ such that $a(u,v) = f(v)$ holds for all $v \in H_0^1(0,1)$." (III.5)

is called **weak solution** of (III.1)

**Minimization Problem**

Consider

$$u''(x) = f(x)\ \forall x \in (0,1),\ u(0) = u(1) = 0.$$

Analogue to (III.5) we have

$$a(u,v) = f(v) \text{ and } a(u,v) = \int_0^1 u'v'\ dx.$$

This gives the minimization problem

"Find $u \in V$ such that $(u,v) = \langle f, v\rangle$ holds for all $v \in V$".

For this setting, we consider

$$J(v) = \frac{1}{2}(u,v) - \langle f, v\rangle.$$

Let us assume that we have a minimizer $u$, i.e.

$$J(u) = \min_{v\in V} J(v). \qquad (*)$$

That means

$$J(u) \le J(u + tw)\ \forall w \in V, \forall t \in [0,1].$$

We have

$$\begin{aligned} J(u+tw) &= \frac{1}{2}\left((u+tw, u+tw)\right) - \langle f, u+tw\rangle \\ &= \frac{1}{2}\left((u,u) + 2t(u,w) + t^2(w,w)\right) - \langle f,u\rangle - t\langle f,w\rangle \\ &= J(u) + \frac{t^2}{2}\|w\|^2 + t(u,w) - t\langle f,w\rangle. \end{aligned}$$

Thus, $(*)$ is equivalent to the problem

"Find $u \in V$ such that $\frac{t}{2}\|w\|^2 + (u,w) - \langle f,w\rangle \ge 0$ for all $w \in V$ and $t \in [0,1]$."

which is equivalent to finding $u \in V$ such that

$$(u,w) \ge \langle f, w\rangle$$

for all $w \in V$.[15]

[15] Since $w \in V$ implies $-w \in V$ we have an equality here.

**Ritz-Galerkin Problem**

We want to solve the problem

$$\text{"Find } u \in V \text{ such that } a(u,v) = f(v) \text{ for all } v \in V\text{."} \tag{III.6}$$

with a bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ and a linear functional $f : V \to \mathbb{R}$ in some Hilbert space $V$.

Choose $V_n \subseteq V$ with $\dim(V_n) = n < \infty$ and solve

$$\text{"Find } u_n \in V_n \text{ such that } a(u_n,v) = f(v) \text{ for all } v \in V_n\text{."} \tag{III.7}$$

Let $\{\varphi_i\}_{i\in\mathbb{N}}$ be a basis of $V_n$. Then (III.7) can be rewritten as

$$\text{"Find } u_n \in V_n \text{ such that } a(u_n,\varphi_i) = f(\varphi_i) \text{ for all } i \in \mathbb{N}\text{."}$$

with

$$u_n = \sum_{i=1}^{n} \overline{u_i}\varphi_i.$$

We arrive at a linear system for the coefficients $\overline{u_i}$, i.e.

$$a(u_n,\varphi_i) = a\left(\sum_{j=1}^{n} \overline{u_j}\varphi_j(x), \varphi_i(x)\right) = \sum_{j=1}^{n} \overline{u_j} a(\varphi_j(x),\varphi_i(x)) = f(\varphi_i)$$

or

$$A_n U_n = F_n,$$

where

$$U_n := \begin{bmatrix}\overline{u_1} & \dots & \overline{u_n}\end{bmatrix}^T,\ A_n := [a_{ij}],\ a_{ij} := a(\varphi_i,\varphi_j) \text{ and } F_n := \begin{bmatrix} f(\varphi_1) & \dots & f(\varphi_n)\end{bmatrix}^T.$$

The next theorem gives the answer to the question whether this linear system has always a solution.

**Theorem III.10.** Let $V$ be a normed space, $V_n \subseteq V$ a subspace with $\dim V_n = n < \infty$ and $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ an $V$-elliptic (coercive) bilinear form, i.e. there exists a $\gamma > 0$ such that

$$a(v,v) \geq \gamma \|v\|_V^2$$

for all $v \in V$ and $f : V \to \mathbb{R}$ linear and continuous, i.e. there exists $k > 0$ such that

$$|f(v)| \leq k \|v\|_V$$

for all $v \in V$. Then,

(i) the linear system $A_n U_n = F_n$ has a unique solution and

(ii) it holds $\|u_n\|_V \leq \frac{k}{\gamma}$.

*Proof.* Let $W_n = \begin{bmatrix}\overline{w_1} & \dots & \overline{w_n}\end{bmatrix}^T \neq 0$ and

$$w = \sum_{i=1}^{n} \overline{w_i}\varphi_i.$$

Then

$$A_n W_n W_n = a(w_n,w_n) \geq \gamma \|w\|_V^2 > 0,$$

hence $A_n W_n \neq 0$ for all $W_n \neq 0$. With this, it holds

$$\gamma \|u_n\|_V^2 \leq a(u_n,u_n) = f(u_n) \leq k \|u_n\|_V.$$

□

Now, we are concerned with the relation between the finite dimensional and the variational problem and how it is possible to construct our scheme such that $A_n U_n = F_n$ is numerically "easy" to solve.

**Theorem III.11.** (Lemma of Cea) Let $V$ be a Hilbert space, $V_n \subseteq V$ be a subspace with $\dim V_n = n < \infty$ and

$$a(\cdot,\cdot) : V \times V \to \mathbb{R}$$

be a $V$-elliptic and continuous bilinear form, where continuity means

$$\exists \Gamma > 0 : \ a(u,v) \leq \Gamma \|u\|_V \|v\|_V \ \ \forall u, v \in V.$$

Then, it holds

$$\|u - u_n\|_V \leq \frac{\Gamma}{\gamma} \inf_{u_n \in V_n} \|u_n - u\|_V .$$

*Proof.* Since $V_n \subseteq V$, we have

$$u \in V : \ a(u, v_n) = f(v_n) \ \forall v_n \in V_n$$

and

$$u_n \in V_n : \ A(u_n, v_n) = f(v_n) \ \forall v_n \in V_n.$$

Substracting the second equation from the first, we get

$$a(u - u_n, v_n) = 0 \ \forall v_n \in V_n.$$

This is called Galerkin-orthogonality. For all $u_n \in V_n$, we have

$$\begin{aligned} \gamma \|u - u_n\|_V^2 &\leq a(u - u_n, u - u_n) \\ &= a(u - u_n, u - v + v - u_n) \\ &= a(u - u_n, u - v) + \underbrace{a(u - u_n, v - u_n)}_{=0} \\ &\leq \Gamma \|u - u_n\|_V \|u - v\|_V . \end{aligned}$$

This completes the proof. □

**Remark.** If $V_n$ has polynomials of degree $k$, then the approximation quality of the interpolation of $u$ in $\mathbb{P}_k$ can be transferred tothe approximation quality of $u - u_n$ in $V_n$.

**Choice of basis functions**

Consider

$$-u'' = f \text{ in } (0,1), \ u(0) = u(1) = 0$$

with

$$a(u,v) = \int_0^1 u'v' \, dx$$

and

$$a_{ij} = a(\varphi_i, \varphi_j) = \int_0^1 \varphi_i' \varphi_j' \, dx.$$

For instance, choose basis functions

$$\varphi_j(x) = \sin(j\pi x), \ j = 1, \ldots, n^{16}$$

and set the meshsize $h = \frac{1}{n}$ and

$$V_n = \operatorname{span}\{\varphi_i\}_{i=1}^n = \left\{ v \mid v(x) = \sum_{j=1}^n c_j \varphi_j(x) \right\} \subseteq V$$

[16] One can show that this is a basis.

It can then be shown, that

$$\lim_{h\to 0}\left(\inf_{v\in V_n}\|u-v\|\right)=0$$

for any given $u\in V$. We then have

$$a(\varphi_i,\varphi_j)=\begin{cases}\pi^2\cdot\frac{j^2}{2} & \text{if } i=j\\ 0 & \text{if } i\neq j\end{cases}.$$

For the right hand side, we obtain

$$F_i=\int_0^1 f(x)\varphi_i(x)\,dx$$

and it can then be shown, that the solution

$$\sum_{j=1}^{n}a(\varphi_i,\varphi_j)\cdot\overline{u_j}=f(\varphi_j)\ \forall i=1,\dots,n$$

is

$$\overline{u_j}=\frac{2F_i}{\pi^2 j^2}.$$

The Galerkin approximation is then

$$u_n(x)=\sum_{j=1}^{n}\overline{u_j}\sin(j\pi x).$$

But having an explicit solution is not given in general. This case is exceptional, since the basis functions $u_j$ are the eigenfunctions of the differential operator.

Consider the same problem as before, but now choose

$$V_n:=\operatorname{span}\left\{\frac{1}{i}x^i\right\}_{i=1}^{n}.$$

The Galerkin approximation is then

$$u_n(x)=\sum_{i=1}^{n}\frac{\overline{u_i}}{i}x^i.$$

The entries of $A_n$ in $A_nU_n=F_n$ are then

$$a_{ij}=a(\varphi_i,\varphi_j)=\frac{1}{i+j-1}.$$

The matrix $A_n$ is called a Hilbert matrix, and has extremely bad condition. For example, for $n=10$, we have $\operatorname{cond}(A)=10^{13}$.

Our demands on the choice of the basis functions are

- The computation of the matrix elements $a_{ij}=a(\varphi_i,\varphi_j)$ and the right hand side $f(\varphi_j)$ should be "cheap".
- In general, we will have $n\gg 100$ and hence $A$ should be sparse.
- The matrix $A$ should not be bad conditioned, i.e. $\operatorname{cond}(A)\approx O(n)$ or $O(n^2)$ would be okay, but $\operatorname{cond}(A)\approx O(n^4)$ or $O(e^n)$ would not be okay.

We take $V_n=\operatorname{span}\{\varphi_i\}_{i=1}^{n}$ as the space of the linear functions

$$\varphi_j(x)=\begin{cases}\frac{x-x_{j-1}}{h} & \text{if } x\in(x_{j-1},x_j]\\ \frac{x_{j+1}-x}{h} & \text{if } x\in(x_j,x_{j+1})\\ 0 & \text{otherwise}\end{cases}.$$

Here $\{x_j\}_{j=0}^n$ is an equidistant mesh of $(0,1)$ with meshsize $h$, i.e. $x_j = j \cdot h$ for $j = 0, \dots, N$ with $h = \frac{1}{N}$.[17] It then follows, that

$$a(\varphi_i, \varphi_j) = \begin{cases} \frac{2}{h} & \text{if } i = j \\ -\frac{1}{h} & \text{if } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases},$$

so $A_n$ is a tridiagonal matrix.

The basis functions hence should have a small support, since this has an impact on the bandwidth of the matrix $A_n$.

### Design of Finite Element spaces

**Definition III.12.** Let $\Omega \subseteq \mathbb{R}^2$ have a polygonal boundary. A set $\mathcal{T}$ of triangles is called a **triangulation of** $\Omega$ if it holds

(i)
$$\bigcup_{t \in \mathcal{T}} = \overline{\Omega}$$
and

(ii) the intersection of two triangles is either a common edge, a point, or empty.

**Example.** The drawings are missing.

**Definition III.13.** We define

$$S^{(m)} := \left\{ v \in C(\overline{\Omega}) \mid v|_t \in \mathbb{P}_m \ \forall t \in \mathcal{T} \right\}$$

as the **space of the (polynomial) finite elements of order** $m$.

**Remark.**

- In general, we have $S^{(m)} \not\subseteq C^1(\Omega)$.
- In the case of $m = 1, 2, 3$ we call the finite elements *linear*,*quadratic* or *cubic* finite elements.

It is unclear how to construct a basis for $S^{(m)}$. The drawings for $m = 1, 2, 3$ are missing. The following lemma tells us how to define basis functions.

**Lemma III.14.** Let $m \geq 0$. In a triangle $t$ we have on $m+1$ parallel lines

$$s = 1 + 2 + \cdots + (m+1) = \frac{(m+1)(m+2)}{2}$$

nodes $\mathcal{N}^{(m)}(t) = \{z_1, \dots, z_s\}$. Then, the interpolation problem

"Find $p \in \mathbb{P}_m$ such that $p(z_i) = p_i$ holds for all $i = 1, \dots, s$."

has a unique solution for all node values $p_i$.

*Proof.* The proof will be given later in the lecture. □

Putting the nodes of all triangles together gives us the following corollary.

**Corollary III.15.** Let

$$\mathcal{N}^{(m)} := \bigcup_{t \in \mathcal{T}} \mathcal{N}^{(m)}(t)$$

be the collection of all nodes. Then, the interpolation problem

"Find $v \in S^{(m)}$ such that $v(p) = v_p$ holds for all $p \in \mathcal{N}^{(m)}$."

has a unique solution for all node values $v_p$.

[17] Notice, that we have $\varphi_i(x_j) = \delta_{ij}$. This property usually is called the *Kronecker-δ-property*.

**Definition III.16.** Solving

$$\text{"Find } \varphi_p \in S^{(m)} \text{ such that } \varphi_p(q) = \delta_{pq} \text{ holds for all } q \in \mathcal{N}^{(m)}\text{."}$$

for each $p \in \mathcal{N}^{(m)}$ gives us the **nodal basis**

$$\Lambda^{(m)} := \left\{ \varphi_p \mid p \in \mathcal{N}^{(m)} \right\}$$

of $S^{(m)}$. If $p$ is a vertex of $t$, then $\varphi_p|_t \in \mathbb{P}_m$ is called a **form function**. We also define

$$S_h := \left\{ v \in S^{(m)} \mid v|_{\partial\Omega} = 0 \right\},$$

where

$$h := \max_{t \in \mathcal{T}} \operatorname{diam}(t).$$

The next theorem gives a characterization of the space $S_h$.

**Theorem III.17.** The space $S_h$ is a closed subspace of $H_0^1(\Omega)$.

*Proof.* No proof. □

This theorem is good news, since it tells us that $S_h$ is the finite dimensional subspace we were looking for earlier, i.e. we can choose $V_n := S_h$.

Slides are missing.

Notice, that $h$ refers to the size of the triangles, the number of basis functions and maybe also the approximation quality. To precisely measure the approximation quality, we need error estimates.

- Discretization error: $\|u - u_h\|_{H^1(\Omega)}$, where $u$ is the true solution.

If $H_0^1(\Omega)$ is replaced by $S_h$, then we know by Cea's lemma that we only have to consider the *approximation error*:

- Approximation Error: $\inf_{v \in S_h} \|u - v\|_{H^1(\Omega)}$.

We now consider the easiest case, i.e. $m = 1$ (linear polynomials), $\Omega = (a, b) \subseteq \mathbb{R}$ and

$$\mathcal{T}: \ a = x_0 < x_1 < \cdots < x_N = b, \ t_i := [x_{i-1}, x_i], \ h_i := x_i - x_{i-1}, \ h := \max_{1 \le i \le N} h_i$$

where $S_h$ is the space of the continuous and piecewise linear functions with zero bounding conditions.

**Theorem III.18.** For each $u \in H_0^1(\Omega)$ it holds

$$\inf_{v \in S_h} \|u - v\|_{H^1(\Omega)} \overset{h \to 0}{\longrightarrow} 0$$

*Proof.* No proof. □

But now, we are interested in how fast the approximations approach $u$.

We now assume $u \in H^2(a, b)$. By Sobolev's Embedding theory it holds

$$H^2(a, b) \subseteq C[a, b].$$

This also holds if $\Omega \subseteq \mathbb{R}^2$ or $\Omega \subseteq \mathbb{R}^3$, but not for $\Omega \in \mathbb{R}^d$ for $d \ge 4$.

Now denote $\|\cdot\|_{1,\Omega} := \|\cdot\|_{H^1(\Omega)}$ and consider

$$u''(x) = f(x), \ x \in (a, b) =: \Omega, \ u(a) = u(b) = 0.$$

We now want to compute the discretization error, which we reduce to the estimatia of the interpolation error.

Step 1. It holds

$$\|u - I_h u\|_{1,\Omega}^2 = \sum_{i=1}^{N} \|u - I_h u\|_{1,t_i}^2 .$$

Step 2. We transform onto the unit interval. For each fixed $i$ we transform the unit interval $T = [0,1]$ affine onto $t_i$, i.e.

$$\begin{aligned} t_i &= F_i(T), \\ x &= F_i(\xi) = x_{i-1} + h_i \xi \\ \xi_i &:= F^{-1}(x) = h^{-1}(x - x_i) \end{aligned}$$

and we set

$$\hat{v}(\xi) := v(F_i(\xi)) = v(x)$$

for all $v \in H^2(t_i)$. The drawing illustrating the transformation is missing. By the transformation rule for integrals, it holds[18]

$$\|v\|_{0,t_i}^2 = \int_{x_{i-1}}^{x_i} v(x)^2 \, dx = h_i \cdot \int_0^1 \hat{v}(\xi)^2 \, d\xi = h_i \, \|\hat{v}\|_{0,T}^2 .$$

The chain rule gives[19]

$$\hat{v}'(\xi) = \frac{d\hat{v}}{d\xi}(\xi) = \frac{dv}{dx}(x) \cdot \frac{dx}{d\xi} = h_i v'(x)$$

and thus

$$\|v'\|_{0,t_i}^2 = h_i^{-1} \, \|\hat{v}'\|_{0,T}^2 .$$

For $h_1 \leq 1$ it holds

$$\|v\|_{1,t_i}^2 \leq h_i^{-1} \, \|\hat{v}\|_{1,T}^2 \tag{III.8}$$

for all $v \in H^2(t_i)$.

Step 3. We compute the local interpolation error. Applying (III.8) onto $u - J_h u$ gives

$$\|u - I_h\|_{1,t_i}^2 \leq h^{-1} \, \|\hat{u} - I_h \hat{u}_i\|_{1,T}^2 ,$$

where

$$\hat{I}v(x) = v(0) + (v(1) - v(0)) \cdot x$$

is the interpolation operator on $T$. It can then be shown, that

$$\left\| v - \hat{I}v \right\|_{1,T}^2 \leq c \, \|v''\|_{0,T}^2 \tag{III.9}$$

with $c = \frac{1}{3}(8 + 2\sqrt{3})$ holds for all $v \in H^2(T)$.

Step 4. We do the backtransformation. Analogue to step 2, it can be shown by the chain rule, that

$$\|\hat{u}_i''\|_{0,T}^2 = h_i^3 \, \|u''\|_{0,t_i}^2 .$$

[18] In the literature, it is common to denote

$$\|\cdot\|_{L^2} = \|\cdot\|_{H^0} = \|\cdot\|_0 .$$

[19] Note that we use the chain rule for the strong derivative and transfer it back to the weak derivative by a density argument.

Summing up, we have

$$\begin{aligned}
\|u - I_h u\|_{1,\Omega}^2 &= \sum_{i=1}^{N} \|u - I_h u\|_{1,t_i}^2 \\
&\leq \sum_{i=1}^{N} h_i^{-1} \left\| \hat{u}_i - \hat{I}\hat{u}_i \right\|_{1,t_i}^2 \\
&\leq c \cdot \sum_{i=1}^{N} h_i^{-1} \|u_i''\|_{1,T}^2 \\
&= c \cdot \sum_{i=1}^{N} h_i^{-1} h_i^3 \|u''\|_{1,t_i}^2 \\
&\leq ch^2 \|u''\|_{0,\Omega}^2 .
\end{aligned}$$

We have thus shown, that for $u \in H^2(a,b)$ the a priori error estimate is given by

$$\|u - u_h\|_1 \leq \tilde{c} h \|u''\|_0 \leq \tilde{c} h \|u\|_2$$

with $\tilde{c} = \sqrt{c} \cdot \frac{\Gamma}{\gamma}$.

**Remark** (FE in higher dimensions, i.e. $d = 2, 3$)**.** This remark is missing, since it heavily relies on illustrations. Just note, that it can happen that we choose bad triangles (the angles are important, "angle condition") for which the transformation gets bad. The importance of the angles marks a meeting point between geometry and numerical analysis. Summing up, the quality of the grid influences the approximation quality.

**Remark.** We are using triangles, since the transformation $F$ gets harder when using different shapes (e.g. squares) since one has to be careful not to flip points.

## III.3 Finite Differences

Our idea is to replace the derivative of the differential equation by suitable differences.

Consider the interval $[0,1]$ and $x_i = i \cdot h$, $h := \frac{1}{N}$, $i = 0, \ldots, N$ and the grid $w_h := \{x_i \mid i = 0, \ldots, N\}$.

**Definition III.19** (Gridfunction)**.** A vector $u_h = \begin{bmatrix} u_0 & \ldots & u_N \end{bmatrix}^T \in \mathbb{R}^{N+1}$ assigning each grid point a function value is called **grid function**.

The restriction $R_h u$ of a function $u \in C^1[0,1]$ onto a grid function is given as

$$R_h u = \begin{bmatrix} u(x_0) & \ldots & u(x_N) \end{bmatrix}^T .$$

**Example.** Consider the grid $\{0.00, 0.25, 0.50, 0.75, 1.00\}$, $u(x) = x^2$ and $R_h u = \begin{bmatrix} 0 & \frac{1}{16} & \frac{1}{4} & \frac{9}{16} & 1 \end{bmatrix}^T$.

Caution: Consider $u(x) = \sin(4\pi x)$ and $v(x) = 0$. Then

$$R_h u = R_h v = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T .$$

This means that different functions can have the same grid function. In this example, we would need a finer mesh for $u$.

**Definition III.20.** Let $v(x)$ be a sufficiently smooth function. Denote $v_i := v(x_i)$, where $x_i$ is a node of the grid. Then, we call

(i)
$$D^+v(x_i) = \frac{v_{i+1} - v_i}{h}$$
the **forward difference**,

(ii)
$$D^-v(x_i) = \frac{v_i - v_{i-1}}{h}$$
the **backward difference**,

(iii)
$$D^0v(x_i) = \frac{v_{i+1} - v_{i-1}}{2h}$$
the **central difference** and

(iv)
$$D^+D^-v(x_i) = \frac{v_{i+1} - 2v_i + v_{i-1}}{h^2}$$
the **second difference**.

**Definition III.21.** Let $L$ be a differntial operator. Then the difference operator $L_h : \mathbb{R}^{N+1} \to \mathbb{R}^N$ is **consistent with $L$ of order $k$** if

$$\max_{0\le i\le N} \|(Lu)(x_i) - (L_hu_h)_i\| =: \|(Lu)(x_i) - (L_hu_h)_i\|_{\infty,d} = O(h^k)$$

holds. We call $\|\cdot\|_{\infty,d}$ the **discrete maximum norm in the space of the grid functions**.

**Remark.** Consistency is a measure for the approximation quality.

**Example.** The operators $D^+, D^-, D^0$ are consistent of order 1 to $L = \frac{d}{dx}$ and $D^+D^-$ is consistent of order 2 to $L = \frac{d^2}{dx^2}$.

**Example.** Consider
$$Lu = \frac{d}{dx}\left(k(x) \cdot \frac{du}{dx}\right),$$
where $k(x)$ is continuously differentiable. We define $L_h$ by
$$\begin{aligned}(L_hu_h)_i &= D^+(aD^-u(x_i))\\ &= \frac{1}{h}\left(a_{i+1}D^-u(x_{i+1}) - a_iD^-u(x_i)\right)\\ &= \frac{1}{h}\left(a_{i+1}\frac{u_{i+1} - u_i}{h} - a_i\frac{u_i - u_{i-1}}{h}\right),\end{aligned}$$
where $a$ is a grid function, that has to be chosen suitable (cp $k(x)$). Using the chain rule, we obtain
$$(Lu)_i = k'(x_i)(u')_i + k(x_i) + (u'')_i.$$
A Taylor expansion gives
$$(L_hu_h)_i = \frac{a_{i+1} - a_i}{h}(u')_i + \frac{a_{i+1} - a_i}{2}(u'')_i + \frac{h(a_{i+1} - a_i)}{6}(u''')_i + O(h^2).$$
For the difference we obtain
$$\begin{aligned}&(Lu)_i - (L_hu_h)_i\\ &= \left(k(x_i) - \frac{a_{i+1} - a_i}{h}\right)(u')_i + \left(k(x_i) - \frac{a_{i+1} - a_i}{2}\right)(u'')_i\\ &\quad - \frac{h(a_{i+1} - a_i)}{6}(u''')_i + O(h^2).\end{aligned}$$

$L_h$ is consistent with $L$ of order 2, if

$$\frac{a_{i+1}-a_i}{h} = k'(x_i) + O(h^2) \quad \text{and} \quad \frac{a_{i+1}-a_i}{2} = k(x_i) + O(h^2).$$

Possible choises are

$$a_i = \frac{k(x_i)-k(x_{i+1})}{2},\ a_i = k\left(x_i - \tfrac{h}{2}\right) \quad \text{and} \quad a_i = \sqrt{k(x_i)\cdot k(x_{i-1})}.$$

Note, that the choice $a_i = k(x_i)$ gives only order of consistency 1.

Now consider

$$Lu = u'' + b(x)u' + c(x)u = f(x),\ x \in (0,1),\ u(0) = u(1) = 0 \tag{III.10}$$

where we write $u = u(x)$ for short. We assume, that $b$ and $c$ are sufficiently smooth and that $c(x) \geq 0$ for all $x \in [0,1]$.

**Definition III.22.** The central difference scheme for (III.10) is given as

$$-D^+D^-u_i + b_iD^0u_i + c_iu_i = f_i,\ i = 1,\dots,N,\ u_0 = u_N = 0.$$

**Remark.** This leads to a tridiagonal matrix system of linear equations

$$r_iu_{i-1} + s_iu_i + t_iu_{i+1} = f_i,\ i = 1\dots,N,\ u_0 = u_N = 0.$$

with

$$r_i = -\frac{1}{h^2} - \frac{1}{2h}b_i,\ s_i = c_i + \frac{2}{h^2} \text{ and } t_i = -\frac{1}{h^2} + \frac{1}{2h}b_i.$$

Transferring the consistency of Definition III.21 to (III.10), we obtain

$$L_hu_h := R_h(Lu).$$

We assume, that the boundary conditions are imposed such that the first row and the last row of $L_h$ are $\begin{bmatrix}1 & 0 & \dots & 0\end{bmatrix}^T$ and $\begin{bmatrix}0 & \dots & 0 & 1\end{bmatrix}^T$, respectively. It holds

$$(R_h(Lu))_0 = u_0,\ (R_h(Lu))_N = u_N.$$

We thus call (III.10) consistent of order $k$, if

$$\|L_hR_hu - R_h(Lu)\|_{\infty,d} \leq c\cdot h^k,$$

where $c$ and $k$ are independent of $h$.[20]

**Definition III.23.** A difference scheme $L_hu_h = f_h$ is called **stable**, if there exists a stability constant $c_s$, which is not depending on the meshsize $h$, such that

$$\|u_h\|_{\infty,d} \leq c_s\|L_hu_h\|_{\infty,d}$$

holds for all grid functions $u_h$.

**Definition III.24.** A difference scheme for (III.10) is **convergent of order** $k$ if there exist constants $c, k > 0$ (being independent of $h$) such that

$$\|u_h - R_hu\|_{\infty,d} \leq c\cdot h^k.$$

For a consistent and stable scheme we have

$$\begin{aligned}
\|u_h - R_hu\|_{\infty,d} &\leq c_s\|L_h(u_h - R_hu)\|_{\infty,d}\\
&= c_s\|L_hu_h - L_hR_hu\|_{\infty,d}\\
&= c_s\|f_h - L_hR_hu\|_{\infty,d}\\
&= c_s\|R_hf - L_hR_hu\|_{\infty,d}\\
&= c_s\|R_hLu - L_hR_hu\|_{\infty,d}\\
&\leq K\cdot h^k.
\end{aligned}$$

We have thus shown the following theorem.

[20] For $u \in C^4[0,1]$ it can be shown, that the difference scheme for (III.10) has order $k = 2$.

**Theorem III.25.** A consistent and stable difference scheme is convergent and the order of consistency and convergence are the same.

**Remark.** The consistency is usually shown by Taylor expansion. The stability is in general harder to show, dealing with functions and matrices.

**Definition III.26.** Let $x, y \in \mathbb{R}^n$ and $A = [a_{ij}] \in \mathbb{R}^{n,n}$. Then we define

$$\begin{aligned} x \leq y &:\Leftrightarrow x_i \leq y_i \ \forall i = 1, \dots, n \\ x \geq 1 &:\Leftrightarrow x_i \geq 1 \ \forall i = 1, \dots, n \\ A \geq 0 &:\Leftrightarrow a_{ij} \geq 0 \ \forall i, j = 1, \dots, n. \end{aligned}$$

We call $A$ **inverse monotone**, if $A^{-1}$ exists and $A^{-1} \geq 0$.

Let us assume, that $A$ is inverse monotone and $Av \leq Aw$ for $v, w \in \mathbb{R}^n$. Then

$$A(v - w) = b \leq 0$$

and thus

$$v - w = A^{-1}b \leq 0$$

so $v \leq w$.

Now, consider the BVP

$$\begin{aligned} (Lu)(x) = &-(pu')'(x) + q(x)u'(x) + r(x)u(x) = f(x), \quad x \in I := [a, b] \\ &u(a) = \alpha, \ u(b) = \beta, \end{aligned} \tag{III.11}$$

the point grid

$$a = x_0 < x_1 < \cdots < x_N = b, \ I_i = [t_{i-1}, t_i)$$

and the difference scheme

$$\begin{aligned} L_h u_i := &- D^0_{h/2}(p_i D^0_{h/2} u_i) + q_i \cdot D^0_h u_i + r_i u_i = f_i, \ i = 1, \dots, N-1 \\ &u_0 = \alpha, \ u_N = \beta, \end{aligned}$$

where

$$D^0_h u(x) = \frac{u(x+h) - u(x-h)}{2h} \text{ and } g_i = g(x_i).$$

This is equivalent to

$$\tilde{A}_h \tilde{u}_h = \tilde{f}_h \tag{III.12}$$

where $\tilde{A}_h$ and $\tilde{f}_h$ are the result of $u_0 = \alpha, u_N = \beta$ and

$$-\left(p_{i-1/2}u_i - (p_{i-1/2} + p_{i+1/2})y_i + p_{i+1/2}u_{i+1} + \tfrac{h}{2}q_i(u_{i+1} - u_{u-1}) + h^2 r_i y_i\right) = h^2 f_i.$$

$A_h u_h = f_h$, where

$$A_h = \frac{1}{h^2}\begin{bmatrix} p_{1/2} + p_{3/2} + h^2 r_1 & -p_{3/2} + \frac{1}{2}hq_1 & & \\ -p_{3/2} + \frac{1}{2}hq_N & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \end{bmatrix}$$

and

$$f_h = \left(f_1 + h^{-2}p_{1/2}\alpha + \tfrac{1}{2}h^{-1}q\alpha, \ f_2, \dots, f_{N-1} + h^{-1}p_{N-1/2}\beta - \tfrac{1}{2}h^{-1}q_N\beta\right),$$

where $A_h$ is triadiagonal.

**Remark.**

- It can be shown that this scheme for (III.11) has order of consistency 2.

- Stability in difference schemes is in general harder to show. The similarity of $A_h$ to $L_h$ is missing.

**Definition III.27.** A matrix $A = [a_{ij}] \in \mathbb{R}^{N\times N}$ is called a **M-matrix** if

(a) $a_{ij} \leq 0$ for $i \neq j$

(b) $A^{-1}$ exists with $A^{-1} \geq 0$.

Let us assume, that in (III.11) we have $p > 0, q = 0$ and $r \geq 0$. Then the matrix $A_h$ is given by

$$A_h = \frac{1}{h^2}\begin{bmatrix} p_{1/2}+p_{3/2}+h^2r_1 & -p_{3/2} & & \\ -p_{i-1/2} & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \end{bmatrix}.$$

This matrix is then *strictly diagonal dominant*, i.e. it holds for at least one $s \in \{1,\dots,N\}$ that

$$\sum_{j\neq i}|a_{ij}| \leq |a_{ii}|,\ 1 \leq i \leq N, \quad \text{and} \quad \sum_{j\neq s}|a_{sj}| < |a_{ss}|.$$

Further, it is *irreducable*, i.e. for each two indices $i, j \in \{1,\dots,N\}$ there exists a sequence of indices $j_1,\dots,j_n$ such that

$$a_{j_1,i} \neq 0,\ a_{j_2,j_1} \neq 0,\dots,\ a_{j,j_m} \neq 0.$$

**Theorem III.28.** It can be shown, that for (III.11) with $q = 0, r \geq 0$ the difference scheme (III.12) is stable.

[21]

In case of $q \neq 0$, we have

$$A_h = \frac{1}{h^2}\begin{bmatrix} p_{1/2}+p_{3/2}+h^2r_1 & -p_{3/2}+1/2hq_1 & & \\ -p_{i-1/2}-1/2hq_i & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \end{bmatrix}.$$

This matrix is unsymmetric and only diagonal dominant, if

$$h \leq 2 \min_{1\leq i\leq N-1}\left\{\frac{\min\{p_{i-1/2}, p_{i+1/2}\}}{|q_i|}\right\}.$$

We now consider the problematic case $|q_i| \gg |p_i|$.

**Example.** Let $I = [0,1],\ f = 0,\ q = 1,\ r = 0,\ 0 < p = \varepsilon \ll 1$ and consider the problem

$$\begin{aligned} L^\varepsilon u(x) = -\varepsilon u''(x) + u'(x) = 0,\ x \in I \\ u(0) = 1, u(1) = 0. \end{aligned}$$

This problem has the analytical unique solution

$$u^\varepsilon(x) = \frac{e^{1/\varepsilon} - e^{x/\varepsilon}}{e^{1/\varepsilon} - 1}.$$

For $x = 1-\delta,\ \delta > 0$, we have

$$u^\varepsilon(1-\delta) = \frac{e^{1/\varepsilon}}{e^{1/\varepsilon}-1}\left(1 - e^{-\delta/\varepsilon}\right) \approx 1.$$

[21] If interested, the books *R. Plato, "Numerische Mathematik - kompakt"* and *Stör/Bullirsch, "Numerical Mathematics II"* can give further information and proofs regarding difference methods.

Further, for $\varepsilon = 0$ we have $u_0 = 1$, which does **not** meet the boundary condition for $x = 1$.

The drawing illustrating the problem with the solution is missing.

Approximating this problem by the above difference scheme with equidistant step size $h = \frac{1}{N}$ we have

$$\begin{gathered} -(\varepsilon + tfrac12h)u_{i-1} + 2\varepsilon u_i - (\varepsilon - \tfrac{1}{2}h)u_{i+1} = 0,\ 1 \le i \le N-1 \\ u_0 = 1, u_N = 0 \end{gathered} \tag{III.13}$$

The corresponding system matrix is only non-negative and diagonal dominant, if $h \le 2\varepsilon$.

For the solution of the difference equation (III.13) we now take the approach $u_i = \lambda^i$ (cp. with characteristic polynomials of chapter II). We are seeking for the roots of the system. It holds

$$\lambda^i + \frac{2\varepsilon}{\frac{h}{2}\varepsilon}\lambda - \frac{\frac{h}{2}+\varepsilon}{\frac{h}{2}-\varepsilon} = 0,\ u_0 = 1, u_N = 0.$$

We take the ansatz

$$u_i = c_1 \cdot \lambda_1^i + c_2\lambda_2^i,$$

where $\lambda_1, \lambda_2$ are the roots of the polynomial. It can then be shown, that the solution is given as

$$u_i = \frac{\lambda_1^N \cdot \lambda_2^i - \lambda_2^N \cdot \lambda_1^i}{\lambda_1^N - \lambda_2^N}$$

and the roots are

$$\lambda_1 = 1 \text{ and } \lambda_2 = \frac{\varepsilon + \frac{h}{2}}{\varepsilon - tfrach2},$$

so for $\varepsilon \ll \frac{h}{2}$ we have $\lambda_2 \approx -1$. In this case, we would have an oscillating solution

$$u_i = \frac{\lambda_2^i - \lambda_2^N}{1 - \lambda_2^N}$$

We can get rid of this, by

(1). Compute $u'(t)$ by $D_h^+$ or $D_h^-$. Because for

$$D_h^- u(t) = \frac{u(t+h) - u(t)}{h}$$

we arrive at

$$-(\varepsilon + h)u_{u-1} + (2\varepsilon + h)u_i - \varepsilon u_{i+1} = 0.$$

The corresponding system matrix is diagonal dominant for all $h > 0$ and is a M-matrix. With $u_i = \lambda^i$ we arrive at the polynomial

$$\lambda^2 - \frac{2\varepsilon + h}{\varepsilon}\lambda + \frac{\varepsilon + h}{\varepsilon} = 0,$$

which has the roots

$$\lambda_1 = \frac{\varepsilon + h}{\varepsilon} \quad \text{and} \quad \lambda_2 = 1,$$

so the solution is

$$u_i = \frac{\lambda_1^N - \lambda_1^i}{\lambda_1^N - 1}.$$

This is called the $D_h^-$ upwind discretization (*Rückwärtsdiskretisierung*).

(2). Artificial diffusion. We set $\hat{\varepsilon} := \varepsilon + \delta h$ and obtain

$$-(\hat{\varepsilon} + \tfrac{h}{2})u_{i-1} + 2\hat{\varepsilon}u_i - (\hat{\varepsilon} - \tfrac{h}{2})u_{i+1} = 0,$$

which works for $\delta \ge 1/2$, $A_h$ is a M-matrix.