Numerical Mathematics II
SS 2019

Lecture by Konstantin Fackeldey

May 15, 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

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

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

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

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

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

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.

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

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 $i, j \in \{1, \ldots, n\},\ n \in \mathbb{N},\ y : I \to \mathbb{C}^n$ and $f : I \to \mathbb{C}^n$ is a linear system of ODEs.

The function f is called inhomogeneity.

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

Theorem I.6. Let y_1, y_n be two solutions of the homogeneous system

Proof. Incomplete. □

I.3 Qualitative Behaviour of ODEs

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

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

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 $[t_0, t_0 + a_0 + a_1 + \ldots]$ 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

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

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

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

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

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

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, Φ) 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. $\square$

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. Δt_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*

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

[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.

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

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

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 Φ depends not on y_{i+1} and **implicit** otherwise.

Example. For the explicit Euler-Method the incremental function Φ is

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

For the implicit Euler-Method the incremental function Φ 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 $0 \leq \Delta t \leq b - t$ 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

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

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 $t_i,\ i=1,\dots,n$ 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 $i=0,\dots,m-1$ and $z_m\in\mathbb{R}$ and it holds

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

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

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

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

Proof. We prove the statement by induction on i. For z_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 Φ 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 $i = 0,\dots,n-1$, where

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

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

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

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

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

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

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

Inserting this into $(\star)$ gives the result. $\square$

Theorem II.7. If a one-step method with Lipschitz condinitions (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 ξ_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 $i = 1,\dots,s$. This is 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

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

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

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

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 on intermediate steps. This can be very useful, since in many cases the function f is hard or expensive to evaluate.

Reconsider the (maybe non-linear) system (II.$\star$) 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 solveability of the system (II.$\star$) 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 $j = 1, \ldots, s$, 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 $r_1^{(0)}, \ldots, r_s^{(0)}$ 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

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

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 at least order of consistency at least $p = 1$ 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 $j = 1, \dots, s$.

(b) The method has order of consistency at least $p = 2$ 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 at least $p = 3$ 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. The proof is an exercise on the fourth sheet. □

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

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

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 $j = 1, \dots, s$.

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

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 $c_1, \dots, c_s \in [0, 1]$ and then seek for a s-th degree polynomial $u_s \in \mathbb{P}_s$, such that it holds

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

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

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

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

or

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

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 ($\star$) and ($\star\star$). $\square$

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.