

# Numerical Mathematics II

## SS 2019

Lecture by Konstantin Fackeldey

May 6, 2019

### Contents

<b>I</b>	<b>Basic Facts on Ordinary Differential Equations</b>	<b>2</b>
I.3	Qualitative Behaviour of ODEs . . . . .	3
I.4	Stability and Flow . . . . .	6
<b>II</b>	<b>Numerics of ODEs</b>	<b>9</b>
II.1	Two different schemes . . . . .	9
II.2	One-Step Methods . . . . .	10

# 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

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

where  $y : I \rightarrow \mathbb{C}^n$ ,  $y \in C^1(I)$  and  $f : I \times \mathbb{C}^n \rightarrow \mathbb{C}^n$ . 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  $f : I \times (\mathbb{C}^n)^n \rightarrow \mathbb{C}^n$  is a function.

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

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

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

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

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

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

Then the IVP  $y'(t) = F(t, y(t))$ ,  $y(t_0) = y_0$  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  $A(t) = (a_{ij}(t))_{ij} \in \mathbb{C}^{n,n}$ ,  $a_{ij} : I \rightarrow \mathbb{C}$  for  $i, j \in \{1, \dots, n\}$ ,  $n \in \mathbb{N}$ ,  $y : I \rightarrow \mathbb{C}^n$  and  $f : I \rightarrow \mathbb{C}^n$  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 autonomous system of first order

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

where  $f : D \rightarrow \mathbb{R}^n$ ,  $D \subset I \times \mathbb{R}^n$ ,  $t_0 \in I$ ,  $I \subset \mathbb{R}$ . 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

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

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 a 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

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

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  $f : X \rightarrow Y$  is local Lipschitz in  $x \in X$  if there 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) interval  $I_0 = [t_0 - a_0, t_0 + a_0]$  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 continuation 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  $f : G \rightarrow \mathbb{R}^n$  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

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

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

**Remark.** In case of  $t_+ < \infty$  the maximum solution approaches for  $t \rightarrow t_+$ , 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  $g : I \rightarrow \mathbb{R}$  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

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

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,  $f : D \rightarrow \mathbb{R}^n$  continuous and local Lipschitz with respect to  $y$  and  $(t_0, y_0) \in D$ . If the solution of

$$\begin{aligned} y'(t) &= f(t, y(t)) \\ y(t_0) &= y_0, \quad y_0 \in \mathbb{R}^n \end{aligned}$$

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

$$\begin{aligned} z'(t) &= f(t, z(t)) \\ z(t_0) &= z_0, \quad z_0 \in \mathbb{R}^n \end{aligned}$$

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

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

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  $y : I \rightarrow \mathbb{R}^n$  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

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

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  $f : D \rightarrow \mathbb{R}^n$  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

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

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

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

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  $\Phi : T \times S \rightarrow S$  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

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

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

- (i)  $\Phi(t_0, y_0) = y_0$  for all  $y_0 \in \mathbb{R}^n$
- (ii)  $\Phi(t_1 + t_2, \cdot) = \Phi(t_2, \Phi(t_1, \cdot))$  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  $t_0$ .
- $\Phi : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$  is differentiable, i.e.  $\Phi(t, y_0)$  is a  $C^1$ /function and it holds

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

**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

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

and

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

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  $t \rightarrow \Phi(t, y_0)$ . 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 predators the 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 proportional 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 Lotka-Volterra equations.

The origin  $(0, 0)$  is the only equilibrium point on the boundary of the state space  $\mathbb{R}_{\geq 0}^2$ . In the interior of  $\mathbb{R}_{\geq 0}^2$  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

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

and

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

By adding up, we obtain

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

or (using the method of *scharf hinsehen*)

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

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

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

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  $V : D \rightarrow \mathbb{R}$ ,  $D \subseteq \mathbb{R}^n$  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

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

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



**Theorem 1.24** (Ljapunov-Stability). Let  $\bar{y} \in D \subseteq \mathbb{R}^n$  be an equilibrium point of  $y' = f(y)$ . Let further  $V : D \rightarrow \mathbb{R}$  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

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

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 1.24 is called Ljapunov-function.

## II Numerics of ODEs

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

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

with initial condition

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

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

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

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

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

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*

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

---

<sup>1</sup>We assume in (II.1) that  $f$  is sufficiently small, such that all necessary (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

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

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

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

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

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

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  $\Phi(t_i, y_i, \Delta t)$ .

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

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

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

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

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  $x \rightarrow a$ ” if  $\left| \frac{f(x)}{g(x)} \right|$  is bounded when  $x \rightarrow a$ . Furthermore it holds ” $f(x) = o(g(x))$  for  $x \rightarrow a$ ” if  $\lim_{x \rightarrow a} \frac{f(x)}{g(x)} = 0$ . 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

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

**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  $\Delta t \rightarrow 0$ , since  $y''$  is bounded in  $[t, t + \Delta t]$ .  $\square$

**Definition II.4** (Convergence of one-step methods). A one-step method with starting value  $y_0 = y(0) + O(\Delta t^p)$ ,  $\Delta t \rightarrow 0$  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

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

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.  $\square$

**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

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

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

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

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

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

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

*Proof.* Reconsider that

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

Rearranging gives

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

Consider now

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

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

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

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 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

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

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

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

which is also called the explicit midpoint rule.

### Excursion to quadrature

We know, that

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

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

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

There exist many quadrature rules. Here we consider **quadrature by interpolation**. Let us assume, that we know  $f$  only at  $(n + 1)$  points  $x_0, \dots, x_n$  and 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.