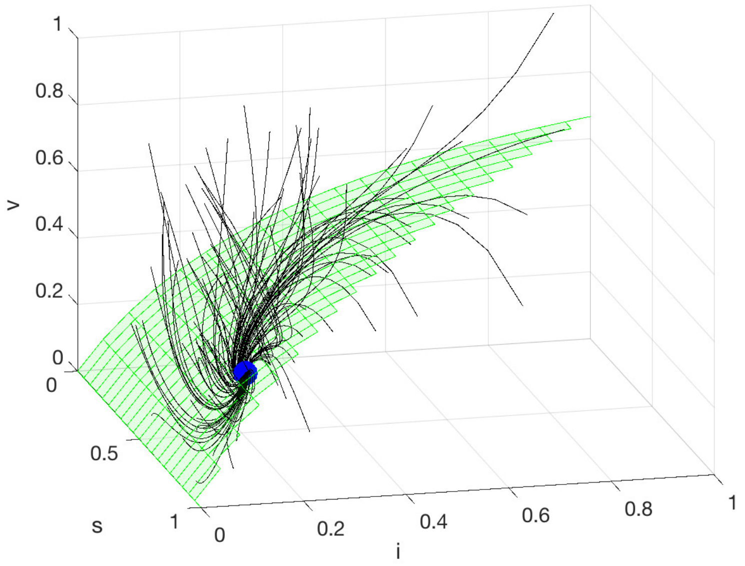


Lecture Notes

Applied Differential Equations

Thomas Götz

university of koblenz
Mathematics / Natural Sciences

Version: February 13, 2023

Contents

Bibliography

[Wal98] WALTER, *Ordinary Differential Equations*, Springer 1998, doi: 10.1007/978-1-4612-0601-9.

[DB02] DEUFLHARD, BORNEMANN *Scientific Computing with Ordinary Differential Equations*, Springer 2002, doi: 10.1007/978-0-387-21582-2

[DH03] DEUFLHARD, HOHMANN *Numerical Analysis in Modern Scientific Computing*, Springer 2003.

[Python] PYTHON, available for download, see www.python.org.

[LL16] LINGE, LANGTANGEN, *Programming for Computations – Python*, Springer 2016. doi: 10.1007/978-3-319-32428-9.

[Sun20] SUNDNES, *Introduction to Scientific Programming with Python*, Springer 2020. doi: 10.1007/978-3-030-50356-7.

The above list is just a small selection of literature and references suitable for self studying, intensive reading or exam preparation. This list does not claim to be complete.

Important Note: These lecture notes represent a summary of the contents of the works mentioned in the bibliography or similar works. The lecture notes are only intended for internal use within the University of Koblenz. The lecture notes do not claim to be free of errors or complete. No liability is assumed for the validity or content of external links. In case of doubt, please consult the works listed in the table of contents.

Chapter 1

Introduction

1.1 Modeling with Differential Equations

Example 1.1 (Discrete population model)**.** We consider the growth of a colony of bacteria. In a first step we consider *discrete time*, i.e. we look at the population y_k at discrete time points $t_k = t_0 + k \cdot \delta t$. Here, t_0 denotes the initial time of observation and $\delta t > 0$ denotes the (fixed) time interval. The number of bacteria observed at time t_k is called $y_k = y(t_k)$.
The population at the next time point $t_{k+1} = t_k + \delta t$ is given by

$$y_{k+1} = y_k + \text{Increase}$$

Assuming the increase to be proportional to both the current population y_k and the length of the time interval δt, we obtain

$$y_{k+1} = y_k + r \cdot \delta t \cdot y_k \ .$$

The constant $r \in \mathbb{R}$ is called the *growth rate* (in case of $r > 0$); it measures the number of newborns per individual and per unit time step. In case of $r < 0$, one calls the constant the *death* or *mortality rate*.

Given an initial population y_0 at initial time t_0, we can predict the population at any later time based on the above recursion or *difference equation*.

Python Example 1.1 (Simulation of discrete population growth)**.**

```
from numpy import *
import matplotlib.pyplot as plt

# Parameters
```

```
n  = 40
T  = 1.0
dt = T/n
t  = linspace(0,T, n+1)

y  = zeros(n+1)
r  = 1.2
y0 = 1000

y[0] = y0
for k in range(n):
    y[k+1] = y[k] + r*dt*y[k]

# Plot
plt.figure()
plt.plot(t,y, 'bo-')
plt.xlabel('Time $t$')
plt.ylabel('Population $y$')
plt.grid()
plt.show()
```

Example 1.2 (Exponential growth)**.** If the time interval δt shrinks to zero, we get

$$\frac{y(t+\delta t)-y(t)}{\delta t} = r \cdot y(t) \xrightarrow{\delta t \to 0} \frac{dy}{dt} = r \cdot y(t) \ ,$$

a so–called *differential equation*

$$y' = ry \tag{1.1}$$

with *initial value* $y(t_0) = y_0$.

Definition 1.1 (Initial value problem)**.** Let $I \subset \mathbb{R}$ be an interval and $t_0 \in I$. The problem: Find a (continuously diff'able) function $y : I \to \mathbb{R}$, satisfying the (ordinary) differential equation (ODE)

$$\frac{dy}{dt} = r \cdot y(t) \ ,$$

for all $t \in I$ and satisfying the initial condition $y(t_0) = y_0$ is called an *initial value problem* (IVP).

The constant $r \in \mathbb{R}$ determines the behavior of the population model (1.1)

▷ For $r = 0$ the population remains constant.

▷ For $r > 0$ the population grows in time.

▷ For $r < 0$ the population decreases in time.

An *educated guess* helps finding a solution to Eqn. (1.1). We use the *ansatz* $y(t) = c \cdot e^{\alpha t}$ with two yet undetermined real constants c and α. Plugging this ansatz into the ODE yields

$$y'(t) = c\alpha e^{\alpha t} \overset{!}{=} ry = rce^{\alpha t}$$

which leads us to

$$0 = ce^{\alpha t}\,(\alpha - r)\ .$$

It follows, that either $c = 0$ and hence $y \equiv 0$. Or, as second possibility, we have to consider $\alpha = r$. Using the initial value $y(t_0) = y_0$, we obtain

$$y(t_0) = ce^{rt_0} \overset{!}{=} y_0\ .$$

Therefore $c = y_0 e^{-rt_0}$ and finally

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

Example 1.3 (Logistic growth model)**.** We can improve the previous population model by assuming, that the reproduction rate r is not constant, but decreases with the current population, e.g. $r = K \cdot (M - y)$. The term $M - y$ can be interpreted as the *available resources*. For a population $y = M$, no resources are available and the reproduction rate equals zero. The threshold M is also called the *carrying capacity* of the model. The resulting ODE reads as

$$y' = K(M - y)\,y\ .$$

This equation is also known as *logistic (differential) equation*.

Introducing the *scalings* $t = s/(KM)$ und $y(t) = Mu(s)$ in the logistic equation, i.e. introducing the new time scale s and measuring the population y in fractions of the carrying capacity M, we get

$$0 = \frac{dy}{dt} - K(M - y)y = KM^2\frac{du}{ds} - K(M - Mu)\,Mu = KM^2\left[u' - (1 - u)u\right]\ .$$

The prototypic equation is given by

$$u' = u(1 - u)\ .$$

A solution to the above ODE can be obtained as follows (a systematic approach will be given later on)

$$1 = \frac{u'}{u(1-u)} = \frac{u'}{u} - (-1)\frac{u'}{1-u} = (\ln u)' - (\ln(1-u))' = \left(\ln \frac{u}{1-u}\right)' .$$

using partial fractions. Integration yields

$$\begin{aligned} t + c_0 &= \ln \frac{u}{1-u} + c_1 \\ ce^t &= \frac{u}{1-u} \\ u &= \frac{ce^t}{1+ce^t} = \frac{1}{1+\tilde{c}e^{-t}} . \end{aligned}$$

Python Example 1.2 (Numerical solution of the logistic equation)**.**

```python
from numpy import *
from scipy.integrate import solve_ivp
import matplotlib.pyplot as plt

u0 = 0.3
T  = 5

def rhs(t,u):
    return u*(1.-u)

Sol = solve_ivp(rhs, [0, T], [u0],
                t_eval=linspace(0,T,101))
t = Sol.t
u = Sol.y[0,:]

plt.figure()
plt.plot(t,u, 'b-')
plt.xlabel('$t$')
plt.ylabel('$u$')
plt.grid()
plt.show()
```

Example 1.4 (Predator–prey model)**.** We consider two interacting species: a predator and a prey. The prey u grows in the absence of predators according to the first population model and gets diminished by the predators. The number of removed individuals is assumed to be proportional to both the available prey population and the current predator population. The predators themselves die out in the absence of the prey. Their reproduction is assumed to be proportional to the available prey and to their own population. These assumptions lead to the following system of coupled ODEs for the prey u and predators v

$$\begin{aligned} u' &= \alpha u - \beta uv \ , \\ v' &= -\delta v + \gamma uv \ . \end{aligned}$$

Considering a stationary equilibrium, i.e. solutions $(u, v) = \text{const.}$, we obtain from $u' = 0 = v'$ the following conditions

$$\begin{aligned} u' &= u\,(\alpha - \beta v) \overset{!}{=} 0 & &\rightsquigarrow u = 0 \quad \text{or} \quad v = \alpha/\beta \\ v' &= v\,(\gamma u - \delta) \overset{!}{=} 0 & &\rightsquigarrow v = 0 \quad \text{or} \quad u = \delta/\gamma \end{aligned}$$

This leads to two possible equilibria or *stationary points*

- ▷ The trivial equilibrium $u = 0 = v$, i.e. no population at all.
- ▷ The *coexistence equilibrium* $(u^*, v^*) = \left(\frac{\delta}{\gamma}, \frac{\alpha}{\beta}\right)$.

How do small perturbations affect the coexistence equilibrium (u^*, v^*)?

Let us consider a *linearization* of the equations around (u^*, v^*). Using the ansatz

$$u = u^* + \text{small perturbation} = \frac{\delta}{\gamma} + x \quad \text{and} \quad v = \frac{\alpha}{\beta} + y$$

where $x, y \ll 1$, i.e. both functions $x(t)$ and $y(t)$ are assume to be that small, such that quadratic terms like x^2, xy, y^2 can be neglected compared to the linear ones. Plugging this into the ODE and keeping only terms up to first order, we arrive at

$$\begin{aligned} x' &= -\frac{\beta\delta}{\gamma} y = -k_1 y \\ y' &= \frac{\alpha\gamma}{\beta} x = k_2 x \ . \end{aligned}$$

Differentiating again leads to

$$x'' = -k_1 k_2 x = -\alpha\delta x \ .$$

Later on, we will see that this *differential equation of second order* describes the so–called *harmonic oscillator* $x'' + k^2 x = 0$ and its solutions are periodic functions of the form

$$x(t) = c_1 \cos kt + c_2 \sin kt \ .$$

This implies, that our solutions u and v oscillate periodically around the equilibrium u^*, v^*.

Drawing the two equilibria as well as the qualitative behaviour of the solutions, i.e. the signum of u' and v' in the uv–plane, we obtain what is called the *phase portrait* of the ODE.

Example 1.5 (Mathematical epidemiology)**.** Let us consider the spread of an infection like influenza in a population. We can subdivide the total population N into three sub–compartments:

- ▷ The susceptible individuals $S = u_1$.
- ▷ The infected individuals $I = u_2$.
- ▷ The recovered individuals $R = u_3$.

The transmission of the disease proceeds stepwise

- ▷ If a susceptible and an infected individual meet, the susceptible one get also infected at a rate β.
- ▷ An infected individual recovers at a rate γ.
- ▷ Recovered and hence immune individuals loose their immunity at a rate δ and get susceptible again.

In mathematical terms, this lead to the ODE–system (called SIR–model)

$$\begin{aligned} S' &= \overbrace{-\beta SI}^{\text{infection}} & & \overbrace{+\delta R}^{\text{loss of immunity}} \\ I' &= \beta SI & -\gamma I & \\ R' &= & \underbrace{+\gamma I}_{\text{recovery}} & -\delta R \end{aligned}$$

The total population $N = S + I + R$ remains constant, since

$$N' = 0\ .$$

Example 1.6 (Pendulum)**.** We consider a pendulum, i.e. a mass m fixed to a (massless) rod of length l and moving on a circular orbit of radius l around the point of fixation (the origin). The angle φ describes the deflection from the position at rest $\varphi = 0$. Gravitation exerts a force $F = -mg \sin\varphi$ and hence an acceleration $ml \cdot \frac{d^2\varphi}{dt^2}$ driving the mass back to its position at rest. Newton's law "Mass × Acceleration = Force" leads to the ODE

$$\varphi'' = -\frac{g}{l}\sin\varphi \ .$$

For small angles $\varphi \ll 1$ we linearize $\sin\varphi \approx \varphi$ and hence

$$\varphi'' = -\frac{g}{l}\varphi \ .$$

Again, we arrive at a differential equation of second order of the form

$$\varphi'' + k^2\varphi = 0$$

where $k^2 = g/l > 0$. A straight forward calculation shows, that the periodic functions

$$\varphi(t) = c_1 \cos kt + c_2 \sin kt$$

are solutions for $c_1, c_2 \in \mathbb{R}$. The constant $k = \sqrt{g/l}$ equals to the *frequency* of the *harmonic oscillator*, i.e. the pendulum.

1.2 Basic Notations

Definition 1.2 (Ordinary differential equation)**.** Let $I \subset \mathbb{R}$ be an interval. We call an equation

$$F(x,\, u,\, u',\, u'',\, \dots,\, u^{(n-1)},\, u^{(n)}) = 0 \quad \forall x \in I \ . \tag{1.2}$$

an *ordinary differential equation* (ODE) of *order* n. A function $u : I \to \mathbb{R}$ is called *solution* of (1.2) on the interval I, if u is n–times continuously differentiable in I and (1.2) holds for all $x \in I$.

Definition 1.3 (Explicit ODE)**.** The ODE (1.2) is called *explicit*, if the highest derivative is given explicitly in terms of the lower ones

$$u^{(n)} = f(x,\, u,\, u',\, u'',\, \dots,\, u^{(n-1)}) \ , \tag{1.3}$$

otherwise we call the ODE *implicit*.

Definition 1.4 (Linear ODE)**.** An ODE is called *linear*, if the function F is linear in $u,\, u',\, \dots,\, u^{(n)}$, i.e. (1.2) can be written in the form

$$a_n(x)u^{(n)} + \dots + a_1(x)u' + a_0(x)u + b(x) = 0 \ . \tag{1.4}$$

The coefficients a_k are allowed to be functions, i.e. $a_k : I \to \mathbb{R}$. If all the coefficients a_k are *constant*, i.e. $a_k = \text{const.}$ for all $k = 0 \dots n$, we call the equation a *linear ODE with constant coefficients*. The equation is called *homogeneous*, if $b(x) \equiv 0$, otherwise we call it *inhomogeneous*.

Definition 1.5 (Autonomous ODE)**.** An ODE of the form $F(u, u', \ldots, u^{(n)}) = 0$ is called *autonomous*; in this case F does not depend explicitly on x.

Definition 1.6. Let $I \subset \mathbb{R}$ be an interval. If the functions $\boldsymbol{u}$, $\boldsymbol{u}'$, $\ldots$, $\boldsymbol{u}^{(n)}$ and $\boldsymbol{F}$ in (1.2) are *vector–valued* (in $\mathbb{R}^m$), we call (1.2) a *system of m ODEs of order n.* A vector–valued function $\boldsymbol{u} : I \to \mathbb{R}^m$ is called *solution*, if it is n–times continuously diff'able and the ODE (1.2) holds for all $x \in I$.

The complex–valued case u, u', $\ldots$, $u^{(n)}, F \in \mathbb{C}^m$ can also be summarized under this definition using the identification $\mathbb{C} \simeq \mathbb{R}^2$.

Remark 1.1. In the script we write (most of the time) vectors $\boldsymbol{u} \in \mathbb{R}^n$ in boldface and matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ in bold upright letters.

Remark 1.2. In physical or technological applications where the independent variable t can be interpreted as *time*, the derivative is often written with a dot, i.e. $\dot{u} = \dfrac{du}{dt}$ and for the second derivative $\ddot{u} = \dfrac{d^2u}{dt^2}$. We will not use this notation.

Example 1.7 (Some ODEs)**.**

$$\begin{array}{rl}
u' = ku & \text{linear growth equation} \\
u'' + ku' + \omega^2 u = f(t) & \text{oscillator} \\
u'' + \mu(u^2+1)u' + u = 0 & \text{VAN DER POL's equation} \\
\left.\begin{array}{rcl} u' &=& (a-bv)u \\ v' &=& (-c+du)v \end{array}\right\} & \text{predator–pey–model} \\
y' = \mathbf{A}\boldsymbol{y} \quad \text{where } \mathbf{A} \in \mathbb{R}^{n\times n},\ \boldsymbol{u}(t) \in \mathbb{R}^n & \text{lin. hom. system of order 1} \\
y' + \ln y' = x & \text{implicit ODE, order 1}
\end{array}$$

But the LAPLACE equation $\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2} = 0$ is a *partial differential equation* (PDE).

1.3 The Directional Field of an ODE

Let $\Omega \subset \mathbb{R}^2$ be non–empty and $f : \Omega \to \mathbb{R}$. We consider the first order explicit ODE

$$u' = f(x, u)\ . \tag{1.6}$$

Let $I \subset \mathbb{R}$ be an interval (open or halfopen). A function $u : I \to \mathbb{R}$ is a solution of the ODE (1.6), if u is diff'able in I, $\operatorname{graph} u \subset \Omega$ and if (1.6) holds, i.e.

$$(x,\ u(x)) \in \Omega \quad \text{and} \quad u'(x) = f\,(x,\ u(x)) \qquad \text{for all } x \in I\ .$$

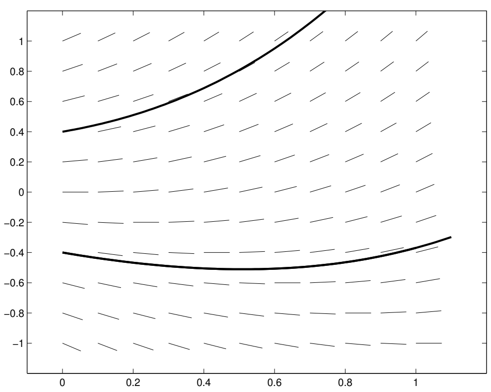


Figure 1.1: Directional field of the ODE $u' = x + u$.

Assume the solution curve (trajectory) passes through a point $(x_0,\ u_0) \in \Omega$, i.e. $u(x_0) = u_0$. Then the slope at that point is given by $u'(x_0) = f(x_0,\ u_0)$. The ODE determines the slopes of the trajectories at the points $(x,\ u)$.
Drawing for "all" points $(x_0,\ u_0) \in \Omega$ a line element with slope $f(x_0,\ u_0)$, we obtain the *directional field* of the ODE (1.6), see Fig. 1.1.
A solution to the ODE "fits" to the directional field. The direction of the tangents to the solution curve is given by the line elements of the directional field.

The above geometrical concept of the directional field raises the question:
Given an *arbitrary* point $(\xi,\ \eta) \in \Omega$. Does there exist a *unique solution* u passing through that point $(\xi,\ \eta)$?

This leads us to the following

Definition 1.7 (Initial value problem)**.** Let $I \subset \mathbb{R}$ be an interval, $\Omega = I \times \mathbb{R}$, $f : \Omega \to \mathbb{R}$ and $(\xi,\ \eta) \in \Omega$. The problem:
Find a function $u : I \to \mathbb{R}$ diff'able, such that

$$\begin{aligned} u'(x) &= f\,(x,\ u(x)) \qquad \forall x \in I \\ u(\xi) &= \eta \end{aligned} \tag{IVP}$$

is called an *initial value problem* (IVP). The second equation $u(\xi) = \eta$ is called the *initial condition.*

Python Example 1.3 (Numerical solution of an IVP)**.**

To solve the IVP $u' = f(t, u)$, $u(t_0) = u_0$ for a scalar–valued right hand side $f : [t_0, T] \to \mathbb{R}$ using PYTHON, we can use the following commands

```
from scipy.integrate import solve_ivp
def f(t,u):                    # Definition of rhs
    ...
Sol = solve_ivp(f, [t0, T], [u0])  # Solve
t = Sol.t                      # t-coordinates
u = Sol.y[0,:]                 # solution values
```

In chapter 2, Section 2.3 and 2.4 we will deal with the existence and uniqueness of solutions to initial value problems (IVP).

Example 1.8 ($u' = f(x)$)**.** We consider the IVP

$$u' = f(x), \qquad u(\xi) = \eta$$

where $f : I \to \mathbb{R}$ is continuous, $\xi \in I$ and we define $\Omega := I \times \mathbb{R}$.

According to the fundamental theorem of calculus, we get

$$\Phi(x) = \int_{\xi}^{x} f(s)\,\mathrm{d}s$$

as *a* solution of the ODE. We obtain *all* solutions in the form

$$u = u(x;\, C) := \Phi(x) + C$$

for arbitrary constants $C \in \mathbb{R}$, the so–called *general solution* of the ODE.

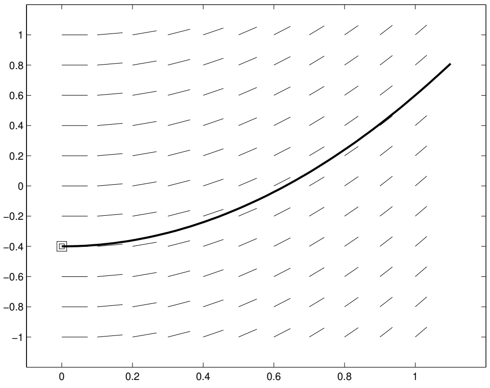


Figure 1.2: Directional field of $u' = 2x$. Initial value $(\xi,\, \eta) = (0,\, -0.4)$.

Using the initial condition $u(\xi) = \eta$ leads to $C = \eta$ and hence

$$u(x) = \eta + \int_{\xi}^{x} f(s)\,\mathrm{d}s\ .$$

The solution exists for all $x \in I$.

Remark 1.3. If I is *not compact* (i.e. not bounded and closed), f may be not bounded on I and hence not integrable over I. However, the integral $\Phi(x)$ exists for all $x \in I$, since $[\xi, x]$ is a compact interval and since f as a continuous function is integrable on this part.

Example 1.9 ($u' = g(u)$)**.** We consider the IVP

$$u' = g(u), \qquad u(\xi) = \eta$$

where g is continuous on an interval I.

Formal calculations yield for $g(u) \neq 0$

$$\frac{du}{dx} = g(u) \qquad \Longleftrightarrow \qquad \frac{du}{g(u)} = dx$$

and integation leads to the implict form

$$\int \frac{\mathrm{d}u}{g(u)} = \int \mathrm{d}x = x + C \ .$$

Using the initial condition, we obtain

$$x(u) = \xi + \int_{\eta}^{u} \frac{\mathrm{d}v}{g(v)} \ .$$

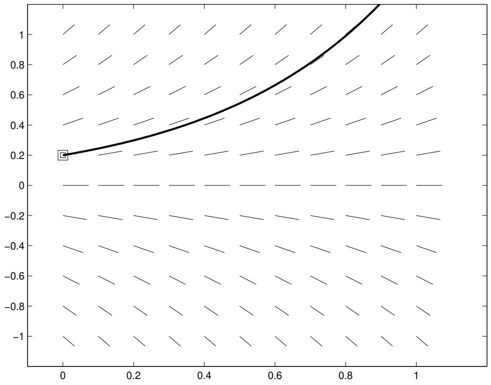


Figure 1.3: Directional field of $u' = 2u$. Initial value $(\xi,\ \eta) = (0,\ 0.2)$.

This is an *implicit* solution $x = x(u)$ of the IVP. Applying the inverse function theorem leads to the solution $u = u(x)$. Examples can be seen in the tutorials.

Example 1.10 (A non–unique ODE)**.** As a special case of the previous example 1.9 we consider the ODE

$$u' = \sqrt{|u|} \ . \tag{1.7}$$

Thanks to symmetry $v(x) := -u(-x)$ is also a solution, if $u(x)$ is a solution. Wlog we just consider positive solutions and obtain formally

$$\int \frac{\mathrm{d}u}{\sqrt{u}} = 2\sqrt{u} = x + C$$

$$u(x;\ C) = \frac{(x+C)^2}{4} \qquad \text{for } x > -C \in \mathbb{R} \ .$$

Furthermore $u \equiv 0$ is a solution and $-u(-x;\ C)$ are negative solutions. Using these three possibilities, we can construct even more solutions by concatenating them smoothly, i.e. such that they are at least diff'able.

Considering the initial value $u(2) = 1$, then all solutions are of the form

$$u(x;\ K) = \begin{cases} x^2/4 & \text{for } x > 0 \\ 0 & \text{for } K \leq x \leq 0 \\ -(x-K)^2/4 & \text{for } x < K \end{cases} ,$$

where $K \leq 0$ is arbitrary.

The IVP (1.7) allows for all initial values $u(\xi) = \eta$ *non unique* solutions. For $\eta = 0$ they branch immediately at $x = 0$, for $\eta \neq 0$ at some distance from the initial value. We say, that the IVP has *locally unique solutions* for $\eta \neq 0$. For $\eta = 0$ the solutions are not even locally unique.

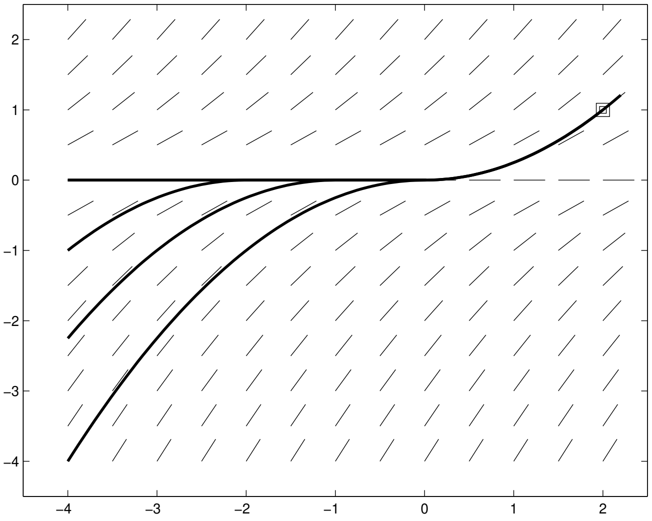


Figure 1.4: Directional field of the ODE $u' = \sqrt{|u|}$ and solution trajectories for the initial value $(\xi,\, \eta) = (2,\, 1)$.

1.4 Separation of Variables

We consider an initial value problem of the type

$$u'(x) = f(x)\, g(u), \qquad u(\xi) = \eta\,. \tag{1.8}$$

Example 1.11 (Separable ODE)**.** Examples for such *ODEs with separated variables* or *separable ODEs* are

(1) $u'(x) = f(x)$.

(2) $u'(x) = g(u)$.

(3) $u'(x) = f(ax + bu + c)$.
Here, one may introduce $v(x) = ax + bu(x) + c$ and derive an ODE for v.

(4) $u'(x) = f(u/x)$.
Let $v(x) = u(x)/x$ and consider an ODE for v.

(5) $u' = f\left(\frac{ax+bu+c}{\alpha x+\beta u+\gamma}\right)$.

(6) The linear first order ODE $u' + a(x)u = b(x)$.

We will treat those examples in the tutorials.

The following theorem deals with the solution of (1.8).

Theorem 1.1. *Let $I_x \subset \mathbb{R}$ and $I_u \subset \mathbb{R}$ be two intervals and let $f : I_x \to \mathbb{R}$, $g : I_u \to \mathbb{R}$ be continuous. Furthermore let $\xi \in I_x$ and $\eta \in \mathit{int}\, I_u$. Given that*

$$g(\eta) \neq 0 \ ,$$

then there exists a neighborhood $U_\xi \subset I_x$ of ξ (one–sided, if $\xi \in \partial I_x$) such that the IVP (1.8) admits an unique *solution $u : U_\xi \to \mathbb{R}$. This solution is given* implicitly *by*

$$\int_\eta^{u(x)} \frac{\mathrm{d}s}{g(s)} = \int_\xi^x f(t)\,\mathrm{d}t \ . \tag{1.9}$$

Proof. Assume (1.9) holds and define $G(u) = \displaystyle\int_\eta^u \frac{\mathrm{d}s}{g(s)}$ and $F(x) = \displaystyle\int_\xi^x f(t)\,\mathrm{d}t$. Then $G : I_u \to \mathbb{R}$ exists in a neighborhood of η, since $g(\eta) \neq 0$. Furthermore, G is diff'able in this neighborhood and it holds that $G'(u) = 1/g(u)$. Due to the inverse function theorem exists a function $H = G^{-1} : \mathbb{R} \to I_u$ where $u = H(G(u))$. Due to (1.9) we get $u = H(F(x))$ or $F(x) = G(u(x))$.

Differentiation w.r.t. x yields $G'(u) \cdot u' = F'$, thus $u' = f(x) \cdot g(u)$. So u satisfied the ODE (1.8).

Moreover $G(\eta) = 0$, $F(\xi) = 0$ and $H(0) = \eta$, therefore $u(\xi) = H(F(\xi)) = \eta$; hence also the initial condition of (1.8) holds.

Let v be another solution, i.e. $v'(x)/g(v(x)) = f(x)$. Integration w.r.t. x yields

$$\int_\xi^x f(t)\,\mathrm{d}t = \int_\xi^x \frac{v'(t)\,\mathrm{d}t}{g(v(t))} = \int_{v(\xi)}^{v(x)} \frac{\mathrm{d}s}{g(s)} \ .$$

Therefore $F(x) = G(v(x))$ and $v(x) = H(F(x)) = u(x)$. □

Question 1.1. What happens in case of $g(\eta) = 0$?
One solution is given by $u(x) \equiv \eta$. But do there exist other solutions?

Theorem 1.2. *Consider the same situation as in Thm. 1.1, but let $g(\eta) = 0$ and $g(u) \neq 0$ for $\eta < u \leq \eta + \alpha$ resp. $\eta - \alpha \leq u < \eta$ for $\alpha > 0$. Furthermore assume*

$$\int_\eta^{\eta+\alpha} \frac{\mathrm{d}s}{g(s)} = \infty \qquad \textit{or resp.} \qquad \int_{\eta-\alpha}^{\eta} \frac{\mathrm{d}s}{g(s)} = \infty \ .$$

Then there exists no *solution $u(x)$ approaching the constant $u \equiv \eta$ from above or below.*

Proof. See [Wal98, Chapter I, §1.VIII]. □

Corollary 1.3. *Let $u(x)$ be a solution of (1.8) where $u(x_0) \lesseqgtr \eta$ for some $x_0 \in I_x$. Then the estimate $u(x) \lesseqgtr \eta$ holds true for all $x \in I_x$.*

If $\eta \in \operatorname{int} I_u$ and both integrals in Thm. 1.2 diverge, then the IVP (1.8) admits a *locally unique* solution.
This is the case, if $g(s)$ has an isolated root at η and if g is Lipschitz–continuous around η, i.e. $|g(u) - g(\eta)| = |g(u)| \leq K\,|u - \eta|$; e.g. if $g'(\eta)$ exists.

Remark 1.4. The opposite direction of Thm.1.2 is in general **not true**, i.e.

$$\text{both integrals converge} \quad \not\Longrightarrow \quad \text{IVP } \textit{not} \text{ uniquely solvable.}$$

1.5 Linear First Order ODEs

A special case of an ODE with separated variables in given by the linear ODE of first order

$$u'(x) = a(x)u + b(x)\ . \tag{1.10}$$

To construct solutions of (1.10) we start with the *homogeneous* problem

$$u'(x) = a(x)u\ . \tag{1.11}$$

Using Thm. 1.1 we can write down the *general* solution

$$u(x) = Ce^{A(x)}\ , \tag{1.12}$$

where $A(x) = \int a(x)\,\mathrm{d}x$. The solution of the IVP (1.11) with $u(\xi) = \eta$ is given by

$$u(x) = \eta e^{A(x)}\ , \quad \text{where} \quad A(x) = \int_\xi^x a(x)\,\mathrm{d}x\ . \tag{1.13}$$

To construct the solution of the *inhomogeneous* problem (1.10), we use the *"variation of constants"*. This approach goes back to Lagrange and starts from the general solution (1.12). We replace the constant C by a function $C(x)$ and obtain

$$\begin{aligned} u(x) &= C(x)e^{A(x)} \\ u'(x) &= C'(x)e^{A(x)} + C(x)A'(x)e^{A(x)} \end{aligned}$$

Inserting the ODE (1.10) we get

$$b(x) = u'(x) - a(x)\,u = \left[C'(x) + C(x)A'(x) - a(x)C(x)\right]e^{A(x)}\ .$$

Since $A(x) = \int a(x)\,\mathrm{d}x$, i.e. $A'(x) = a(x)$, we derive

$$\begin{aligned} C'(x) &= b(x)e^{-A(x)} \\ C(x) &= C_0 + \int b(t)e^{-A(t)}\,\mathrm{d}t\ . \end{aligned}$$

The solution of the IVP $u'(x) = a(x)u + b(x)$, $u(\xi) = \eta$ can now be written as

$$u(x) = \left[\eta + \int_\xi^x b(t)e^{-A(t)}\,\mathrm{d}t\right] e^{A(x)} = \eta e^{A(x)} + \int_\xi^x b(t)e^{A(x)-A(t)}\,\mathrm{d}t$$

where $A(t) = \displaystyle\int_\xi^t a(\tau)\,\mathrm{d}\tau$ and especially $A(x) - A(t) = \displaystyle\int_t^x a(\tau)\,\mathrm{d}\tau$.

Chapter 2

Existence and Uniqueness

2.1 Some Tools from Functional Analysis

Definition 2.1 (Norm)**.** Let V be a real vector space. A function $\|\cdot\| :\ V \to \mathbb{R}$ is called *norm*, if

(1) (definiteness) $\|u\| \geq 0$ and $\|u\| = 0$ if and only if $u = 0$

(2) (homogenity) $\|\lambda u\| = |\lambda|\, \|u\|$

(3) (triangle inequality) $\|u + v\| \leq \|u\| + \|v\|$

hold for all $u, v \in V$ and $\lambda \in \mathbb{R}$.

Definition 2.2 (Norms on continuous functions)**.** Let $I \subset \mathbb{R}$ be a compact interval and $m \in \mathbb{N}$. We define the space of continuous scalar (or vector–valued) functions on I as

$$\mathcal{C} = \mathcal{C}(I, \mathbb{R}^m) := \{u :\ I \to \mathbb{R}^m,\ u \text{ continuous}\}\ .$$

One can check, that $\mathcal{C}$ is a real vector space.

We define the maximum (or infinity) norm $\|\cdot\|_\infty$ on $\mathcal{C}$ by

$$\|u\|_\infty := \max_{x \in I} |u(x)|_\infty = \max_{x \in I} \max_{i=1\ldots m} |u_i(x)|\ .$$

One can check, that $\|\cdot\|_\infty$ defines a norm on $\mathcal{C}$ and that $\|u\|_\infty < \infty$ holds for all $u \in \mathcal{C}$.

Let $p :\ I \to \mathbb{R}$ be a *weighting function* satisfying $0 < a \leq p(x) \leq b < \infty$ for all $x \in I$. Then

$$\|u\|_{p,\infty} := \|u\, p\|_\infty = \max_{x \in I} \left(p(x) \cdot |u(x)|_\infty\right)$$

is called *weighted maximum norm.* One can check that $\|\cdot\|_{p,\infty}$ also defines a norm on $\mathcal{C}$.

As a special case we consider the exponential weighting function $p(x) := e^{-\alpha|x-\xi|}$ for $\alpha > 0$ und $\xi \in I$. In the sequel we frequently make use of the norm induced by this weight and use the notation

$$\|u\|_{\alpha,\infty} := \max_{x\in I} |u(x)|_\infty \cdot e^{-\alpha|x-\xi|} .$$

Definition 2.3 (Banach space)**.** Let $(V,\ \|\cdot\|)$ be a normed vector space. We call V *complete* or Banach space, if all Cauchy sequences converge in V.

Lemma 2.1. *Let $I \subset \mathbb{R}$ be a compact interval and $m \in \mathbb{N}$. The vector space $\left(\mathcal{C}(I,\mathbb{R}^m), \|\cdot\|_{p,\infty}\right)$ is complete for any weighting function $p : I \to \mathbb{R}$.*

In particular, for any $\alpha > 0$, the spaces $(\mathcal{C}(I,\mathbb{R}^m), \|\cdot\|_\infty)$ and $\left(\mathcal{C}(I,\mathbb{R}^m), \|\cdot\|_{\alpha,\infty}\right)$ are Banach *spaces.*

Proof. As an exercise. □

Definition 2.4 (Operator)**.** Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be two real normed vector spaces. Let $D \subset V$ and $T : D \to W$ be a mapping (operator).

We call T a *functional*, if $W = \mathbb{R},\ \mathbb{C}$. Furthermore, we call T

linear, if D is a subspace of V and $T[\alpha u+\beta v] = \alpha T[u]+\beta T[v]$ for all $\alpha,\ \beta \in \mathbb{R}$ and all $u,\ v \in D$.

affine linear, if D is a subspace of V and $T[\lambda u+(1-\lambda)v] = \lambda T[u]+(1-\lambda)T[v]$ for all $\lambda \in \mathbb{R}$ and all $u,\ v \in D$.

continuous in $u_0 \in D$, if $T[u_n] \to T[u_0]$ for all sequences $(u_n)_{n\in\mathbb{N}} \to u_0$.

Lipschitz–cont. (L–cont.) in D with Lipschitz–constant $q \geq 0$, if $\|T[u] - T[v]\|_W \leq q\,\|u - v\|_V$ for all $u,\ v \in D$.

contraction, if T is L–continuous with L–constant $q < 1$.

Remark 2.1. Let T be a linear operator. Then the following holds true:

$$\begin{aligned} T \text{ cont. in } u_0 \in V \quad &\Longleftrightarrow \quad T \text{ cont. in } 0 \\ T \text{ L–cont.} \quad &\Longleftrightarrow \quad \|T[u]\|_W \leq q\,\|u\|_V \quad \forall u \in D \end{aligned}$$

Lipschitz–cont. operators are continuous. The minimal L–constant of a linear operator is called the *operator norm* of T.

Note, that the definition of continuity and L–continuity involve *two* different norms: $\|\cdot\|_V$ for the arguments and $\|\cdot\|_W$ for the images.

Example 2.1. Let $I \subset \mathbb{R}$ be a compact interval. Let $V = W = \mathcal{C}(I, \mathbb{R}^m)$. The integral operator

$$T[u](x) := \eta + \int_\xi^x u(s)\,\mathrm{d}s$$

is *affine linear* and *L–continuous* w.r.t. the $\|\cdot\|_{\alpha,\infty}$–norm in V and W with the L–constant $1/\alpha$. Check this as an exercise.

2.2 Banach Fixpoint Theorem

An important tool to prove the Picard–Lindelöf Theorem on existence and uniqueness of solutions to initial value problems for ordinary differential equations is the follwoing

Theorem 2.2 (Banach Fixpoint Theorem)**.** *Let B be a* Banach *space and let $D \subset B$ be* closed *and* non–empty. *Furthermore let $T : D \to B$ be a* contraction *and* self–mapping, *i.e. $T(D) \subset D$. Then there exists* a unique fixpoint u^* *of T in D, i.e. the fixpoint equation*

$$u = T[u]$$

has a unique solution $u^* \in D$. The fixpoint–iteration

$$u_{n+1} = T[u_n]$$

converges for all *initial guesses $u_0 \in D$ to the fixpoint u^* and*

$$\|u_n - u^*\| \leq \frac{1}{1-q}\,\|u_{n+1} - u_n\| \leq \frac{q^n}{1-q}\,\|u_1 - u_0\| \ ,$$

where $\|\cdot\|$ denotes the norm in B and $q < 1$ equals to the L–constant of T.

Proof. (1) Since T is a self–mapping, it holds that $u_n \in D$.

(2) We show

$$\|u_{n+1} - u_n\| \leq q^n\,\|u_1 - u_0\| \ . \tag{$*$}$$

For $n = 0$ this holds true and

$$\|u_{n+2} - u_{n+1}\| \leq \|T[u_{n+1}] - T[u_n]\| \leq q\,\|u_{n+1} - u_n\| \leq q^{n+1}\,\|u_1 - u_0\| \ .$$

(3) Let $u, v \in D$. Then

$$\begin{aligned}\|u - v\| &\leq \|u - T[u]\| + \|T[u] - T[v]\| + \|v - T[v]\| \\ &\leq \|u - T[u]\| + q\,\|u - v\| + \|v - T[v]\| \\ &\leq \frac{1}{1-q}\,[\|u - T[u]\| + \|v - T[v]\|]\end{aligned} \tag{$**$}$$

(4) **Uniqueness:** Let u, v be two fixpoints. Due to ($**$) we get $u = v$.

(5) **Existence:** Set $u = u_{n+p}$ and $v = u_n$ for $p, n \in \mathbb{N}$ in ($**$), then

$$\|u_{n+p} - u_n\| \leq \frac{1}{1-q} \left[\|u_{n+p+1} - u_{n+p}\| + \|u_{n+1} - u_n\|\right]$$

and using ($*$)

$$\leq \frac{1}{1-q} \left[(q^{n+p} + q^n) \|u_1 - u_0\|\right] \leq \frac{2\|u_1 - u_0\|}{1-q} q^n \, .$$

hence $(u_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $D \subset B$. Since B is a Banach space, the sequence u_n converges to some $u^* \in B$. The set $D \subset B$ is closed and $u_n \in D$, hence the limit u^* is also an element of D. The operator T is continuous, therefore $T[u_n]$ converges to $T[u^*]$ and we obtain the fixpoint equation $u^* = T[u^*]$.

(6) Let $u = u_n$ and $v = u^*$ in ($**$), then $\|u_n - u^*\| \leq \frac{1}{1-q} \|u_{n+1} - u_n\|$ and using ($*$) we obtain the estimate $\|u_n - u^*\| \leq \frac{q^n}{1-q} \|u_1 - u_0\|$.

□

2.3 Picard–Lindelöf Theorem

An important result is the following Theorem by Picard and Lindelöf. We consider the IVP

$$u' = f(x, u), \qquad u(\xi) = \eta \, . \tag{IVP}$$

Here, $I \subset \mathbb{R}$ is a compact interval, $x, \xi \in I$, $\eta \in \mathbb{R}^m$ and $f : I \times \mathbb{R}^m \to \mathbb{R}^m$ is assumed to be *continuous*. This means, we consider a scalar ($m = 1$), complex ($m = 2$, $\mathbb{C} \simeq \mathbb{R}^2$) or vectorial ($m > 1$) IVP for a *first order ODE*.

Theorem 2.3 (Picard–Lindelöf)**.** *Given (IVP) and assume f to be continuous w.r.t. x and L–continuous w.r.t. u on $I \times \mathbb{R}^m$, i.e.*

$$\|f(x, u) - f(x, v)\| \leq L \|u - v\| \qquad \forall u, v \in \mathbb{R}^m \, .$$

Then the initial value problem (IVP) admits a unique *solution on I.*

Proof. The IVP is equivalent to the fixpoint–problem

$$u(x) = T[u](x) := \eta + \int_\xi^x f(s, u(s)) \, \mathrm{d}s \, . \tag{FPP}$$

Let u be a continuous solution of (FPP), then u satisfies the initial condition. Since the right hand side of (FPP) is continuously diff'able, u is also cont. diff'able and hence $u' = f(x,u)$. On the other hand, let u be a diff'able solution of (IVP). Then the continuity of f implies the continuity of $\varphi(x) := f(x, u(x))$ and hence u is continuously diff'able. Due to the fundamental theorem of calculus the fixpoint equation (FPP) holds.

To apply the Banach Fixpoint Theorem, we let $B = \mathcal{C}(I,\ \mathbb{R}^m)$ and use the exponentially weighted norm $\|\cdot\|_{\alpha,\infty}$ on B for some $\alpha > 0$ to be specified later on.

- ▷ B is a Banach space w.r.t this norm, c.f Lemma 2.1,
- ▷ $D = B$ is closed and non–empty,
- ▷ $T(D) \subset D$, hence T is a self–mapping,
- ▷ **contraction:** Wlog we consider $x \geq \xi$. Let $u, v \in D$.

$$\begin{aligned}
|T[u](x) - T[v](x)|_\infty &= \left| \int_\xi^x f(s, u(s)) - f(s, v(s))\,\mathrm{d}s \right|_\infty \\
&\leq \int_\xi^x |f(s, u(s)) - f(s, v(s))|_\infty \,\mathrm{d}s \\
&\leq \int_\xi^x L\, |u(s) - v(s)|_\infty\, e^{-\alpha|s-\xi|} e^{\alpha|s-\xi|}\,\mathrm{d}s \\
&\leq L\, \|u - v\|_{\alpha,\infty} \frac{e^{\alpha|x-\xi|} - 1}{\alpha} \\
&\leq \frac{L}{\alpha} \|u - v\|_{\alpha,\infty}\, e^{\alpha|x-\xi|} \\
|T[u](x) - T[v](x)|_\infty\, e^{-\alpha|x-\xi|} &\leq \frac{L}{\alpha} \|u - v\|_{\alpha,\infty} \qquad \forall x \in I \\
\|T[u] - T[v]\|_{\alpha,\infty} &\leq \frac{L}{\alpha} \|u - v\|_{\alpha,\infty}
\end{aligned}$$

Choosing $\alpha = 2L$, we get

$$\|T[u] - T[v]\|_{\alpha,\infty} \leq q\, \|u - v\|_{\alpha,\infty} \quad \text{for } q = \frac{1}{2}\,.$$

Hence we can apply the Banach Fixpoint Theorem to (FPP) and obtain the desired result. □

In many cases f is not defined on the entire set $I \times \mathbb{R}^m$ but just on a subset $I \times G$.

Corollary 2.4. *Let $I \subset \mathbb{R}$ be a compact interval. Let $G \subset \mathbb{R}^m$ be compact and simply connected. Let $\Omega = I \times G$. Assume, that $f : \Omega \to \mathbb{R}^m$ be continuous w.r.t. $x \in I$ and* Lipschitz*–continuous w.r.t $u \in G$.*

Then the following holds:
The initial value problem (IVP) *admits a* unique *solution existing* at least *on an interval $J = I \cap \{|x - \xi| \leq \delta / \|f\|_\infty\}$ where $\delta = \operatorname{dist}(\eta, \partial G)$.*

Proof. Let $x \in J$. Then $|x - \xi| \leq \delta / \|f\|_\infty$ and

$$|T[u](x)|_\infty \leq |\eta|_\infty + \int_\xi^x |f(s, u(s))|_\infty \,\mathrm{d}s \leq |\eta|_\infty + |x - \xi| \, \|f\|_\infty \leq |\eta|_\infty + \delta$$

Hence $T[u](x) \in G$. We modifiy the proof of Picard–Lindelöf as follows:

- ▷ Let $B = \mathcal{C}(J, \mathbb{R}^m)$. Then B is again a Banach space.
- ▷ Let $D = \{u \in B : \operatorname{graph} u \subset \Omega\}$. Then D is non–empty and closed.
- ▷ We have already shown that $T[u](x) \in G$, hence $\operatorname{graph} T[u] \in \Omega$ for all $u \in D$ and therefore $T(D) \subset D$.

The contraction property of T remains unchanged. □

The following result deals with the *global* existence and uniqueness of solutions.

Corollary 2.5. *Let $\Omega \subset \mathbb{R}^{m+1}$ be a domain (non–empty, open and simply connected). Let $f : \Omega \to \mathbb{R}^m$ be continuous and* Lipschitz*–continuous w.r.t. $\boldsymbol{u}$ for all $(x, \boldsymbol{u}) \in \Omega$. Then:*
The initial value problem (IVP) *has a* unique *solution. This solution can be extended up to the boundary of Ω.*

Proof. See [Wal98, §10, Thm. VI]. □

Remark 2.2. The above Corollary states, that the solution u exists e.g. on an interval $\xi \leq x \leq b$ right of ξ, where $b = \infty$ is also possible. Depending on the value of b we have one of the following cases:

(1) $b = \infty$: The solution exists for all $x \geq \xi$, i.e. the solution can be extended up to the boundary of Ω in x–direction.

(2) $b < \infty$ and $\sup |u(x)| \to \infty$ for $x \to b$: The solution blows up, i.e. it can be extended up to the boundary of Ω in u–direction.

(3) $b < \infty$ and $\operatorname{dist}\left((x, u(x)),\, \partial\Omega\right) \to 0$ for $x \to b:$ The solution gets arbitrarily close to the boundary of Ω.

In particular this implies that the *maximal interval of existence* of the solution u is *open*. The same holds true for values of x left of ξ.

All these results also hold true for initial value problems for ordinary differential equations of order n. Consider an IVP of order n

$$F\left(x, \boldsymbol{u}, \boldsymbol{u}', \ldots, \boldsymbol{u}^{(n)}\right) = \boldsymbol{0}\ ,$$

with initial conditions

$$\boldsymbol{u}^{(i)}(\xi) = \boldsymbol{\eta}^{(i)} \quad \text{for} \quad i = 0, \ldots, (n-1)\ .$$

We introduce the auxiliary functions $\boldsymbol{y_i} = \boldsymbol{u}^{(i)}$ für $i = 0, \ldots, (n-1)$ and obtain the new IVP

$$\begin{aligned} \boldsymbol{y_i}' &= \boldsymbol{y_{i+1}} \quad \text{for} \quad i = 0, \ldots, (n-2)\ , \\ F\left(x, \boldsymbol{y_0}, \boldsymbol{y_1}, \ldots, \boldsymbol{y_{n-1}}, \boldsymbol{y}'_{\boldsymbol{n-1}}\right) &= \boldsymbol{0}\ , \end{aligned}$$

with the initial conditions given by

$$\boldsymbol{y_i}(\xi) = \boldsymbol{\eta}^{(i)} \quad \text{for} \quad i = 0, \ldots, (n-1)\,,$$

With this trick, we have transformed the scalar valued IVP of order n into a *system* of n *first–order IVPs*.

2.4 Peano's Theorem

The Picard–Lindelöf Theorem ensures the existence and *uniqueness* of a solution to an IVP provided the right hand side f is Lipschitz–continuous. But what happens, if f is no longer assumed to be Lipschitz–continuous?

As we have already seen in Example 1.10, the *uniqueness* may get lost. However, in this example solutions still exist. The Peano existence theorem ensures the existence of solutions provided the right hand side is at least *continuous*.

Theorem 2.6 (Peano Existence Theorem)**.** *Let $I \subset \mathbb{R}$ be a compact interval and $m \in \mathbb{N}$. Let $G \subset \mathbb{R}^m$ be closed and simply connected. Let $\Omega = I \times G$ and $f \in \mathcal{C}(\Omega)$ be bounded. Let $(\xi, \eta) \in \Omega$. Then the IVP*

$$u' = f(x, u) \qquad u(\xi) = \eta \tag{IVP}$$

has a least one solution *existing on an interval $J = I \cap \{|x - \xi| < \delta / \|f\|_\infty\}$ where $\delta = \operatorname{dist}(\eta, \partial G)$.*

Proof. The proof utilizes Schauder's fixpoint theorem instead of Banach. For details we refer to [Wal98, §7]. □

Corollary 2.7. *Let $\Omega \subset \mathbb{R}^{m+1}$ be a domain. Let $f \in \mathcal{C}(\Omega, \mathbb{R}^m)$. For any initial value $(\xi, \eta) \in \Omega$ there exists a solution to the IVP (IVP). This solution can be extended to the boundary of Ω.*

Proof. Without proof. □

2.5 Upper– and Lower Functions

In this section we consider just real and scalar ODEs. All functions appearing are assumed to be real–valued.

The concept of *upper*– and *lower functions* allows to derive estimates for the solutions to ODEs. Hence we can analyze the *qualitative* behavior of the solutions without solving the ODE explicitly.

Wlog we just consider solutions to the right of the initial value ξ. Mutatis mutandis the same holds true left of ξ.

Lemma 2.8. *Let Φ and Ψ be two diff'able functions on the interval $I_0 := \{\xi < x \leq \xi + a\}$. Let $\Phi(x) < \Psi(x)$ on $\xi < x < \xi + \varepsilon$ for some $\varepsilon > 0$. Then we are in one of the following cases.*

(1) Either $\Phi < \Psi$ on the entire interval I_0 or

(2) there exists $x_0 \in I_0$ such that $\Phi(x) < \Psi(x)$ on $\xi < x < x_0$ and $\Phi(x_0) = \Psi(x_0)$, $\Phi'(x_0) \geq \Psi'(x_0)$.

Proof. If Case 1 is not true, then due to the assumptions and the continuity of Φ and Ψ there exists $x_0 \in I_0$ such that $\Phi(x) < \Psi(x)$ for $\xi < x < x_0$ and $\Phi(x_0) = \Psi(x_0)$. In remains to show, that $\Phi'(x_0) \geq \Psi'(x_0)$. Consider the left sided difference quotient at x_0

$$\frac{\Phi(x_0) - \Phi(x)}{x_0 - x} > \frac{\Psi(x_0) - \Psi(x)}{x_0 - x}$$

where $\xi < x < x_0$. Passing to the limit $x \nearrow x_0$ yields $\Phi'(x_0) \geq \Psi'(x_0)$. □

Definition 2.5. Let $\Omega \subset \mathbb{R}^2$, $f \in \mathcal{C}(\Omega)$. We consider the IVP

$$u' = f(x, u), \qquad u(\xi) = \eta \tag{IVP}$$

on an interval $I = [\xi, \xi + a]$ for some $a > 0$.
A function v or resp. w is called a *lower*– or *upper function* of the IVP, if

$$v' < f(x, u(x)) \text{ on } I \text{ and } v(\xi) \leq \eta \quad \text{resp.} \quad w' > f(x, u(x)) \text{ on } I \text{ and } w(\xi) \geq \eta\,.$$

Proposition 2.9. *Let u be a solution of* (IVP) *and let v and w be lower– and upper functions. Then*

$$v(x) < u(x) < w(x) \qquad \text{on} \quad I_0 = (\xi, \xi + a)\ .$$

Proof. For v: If $v(\xi) < \eta$, then due to the assumption $v' < f(x, u(x))$it holds that $v(x) < u(x)$ for all $x \in I_0$.
If $v(\xi) = \eta$, then $v'(\xi) < f(\xi, u(\xi)) = f(\xi, \eta) = u'(\xi)$ implies the existence of $\tilde{\xi} < \xi$ where $v(\tilde{\xi}) < u(\tilde{\xi})$. Now we are in the situation of the first case. □

In applications the following theorem is widely used.

Theorem 2.10. *Let $I = [\xi, \xi + a]$, $R \subset \mathbb{R}$ be an interval and $\Omega = I \times R$. Let $f \in \mathcal{C}(\Omega)$ and $g, h \in \mathcal{C}(\Omega)$ be L–continuous. Furthermore let u be a solution of* (IVP) *and assume*

$$\begin{aligned} v' &= g(x, v)\ , & v(\xi) &\le \eta\ , & g(x, v) &\le f(x, v) & \forall x &\in I\ ,\\ w' &= h(x, w)\ , & w(\xi) &\ge \eta\ , & h(x, w) &\ge f(x, w) & \forall x &\in I\ . \end{aligned}$$

Then

$$v(x) \le u(x) \le w(x) \qquad \forall x \in I\ .$$

The functions v and w are again called lower– and upper function.

Proof. We consider the auxiliary problem $u_n'(x) = f(x, u) + \frac{1}{n}$, $u_n(\xi) = \eta + \frac{1}{n}$ for $n \in \mathbb{N}$. Since the assumptions of Proposition 2.9 are satisfied for all $n \in \mathbb{N}$, i.e. $v(x) < u_{n+1}(x) < u_n(x)$ we obtain $u(x) < u_{n+1}(x) < u_n(x)$ for all $x \in I$. The sequence $(u_n)_n$ is pointwise monotonically decreasing and bounded; hence it converges pointwise. Moreover it holds that

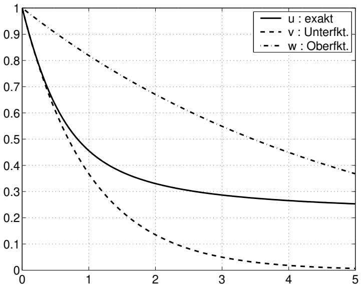
$$v(x) \le u(x) = \lim_{n \to \infty} u_n(x) \qquad \forall x \in I\ .$$

For w we proceed analogously. □

Remark 2.3. The detour of introducing the sequence u_n is necessary, since the equality sign in the assumptions does not allow for a direct proof, right?

Example 2.2. Let $I = [0, 2]$. We consider the IVP

$$u'(x) = -\frac{u}{1 + x^2}, \qquad u(0) = 1\ .$$



The ODE $v' = -v$ together with $v(0) = 1$ yields $v(x) = e^{-x}$ as a *lower function.*
The ODE $w' = -w/5$ with $w(0) = 1$ yields the *upper function* $w(x) = e^{-x/5}$.
The solution of the original IVP is given by $u(x) = \exp(-\arctan x)$.

Chapter 3

Linear Differential Equations

Example 3.1 (Semidiscretization of the heat equation)**.** We consider the one–dimensional heat equation

$$\frac{\partial u}{\partial t} - \kappa \frac{\partial^2 u}{\partial x^2} = f \;, \tag{3.1}$$

for $t \in \mathbb{R}_+$ and $0 < x < 1$ supplemented by the initial condition

$$u(t = 0, x) = u_0(x) \;,$$

and the boundary conditions

$$u(t, x = 0) = u_L(t), \quad u(t, x = 1) = u_R(t) \;.$$

The idea behind *semidiscretization* is the following: We introduce a spatial grid $x_i = ih$, $h = 1/n$ for $i = 0, \ldots, n$ and approximate the spatial derivative by a *finite difference*

$$\frac{\partial^2 u}{\partial x^2}(t, x_i) \approx \frac{u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)}{h^2}$$

using $u_i(t) = u(t, x_i)$. Plugging this approximation into the heat equation (3.1), we obtain a *system of ordinary differential equations*

$$u_i' - \frac{\kappa}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right) = f_i(t) \;, \tag{3.2a}$$

for $i = 1, \ldots, n-1$ and with initial conditions $u_i(0) = u_{0,i}$. For the nodes $i = 1$ and $i = n-1$ next to the boundary we use the boundary conditions for u_0 and u_n and obtain

$$u_1' - \frac{\kappa}{h^2}\left(u_2 - 2u_1\right) = f_1(t) + \frac{\kappa}{h^2} u_L(t) \;, \tag{3.2b}$$

$$u_{n-1}' - \frac{\kappa}{h^2}\left(-2u_{n-1} + u_{n-2}\right) = f_{n-1}(t) + \frac{\kappa}{h^2} u_R(t) \;. \tag{3.2c}$$

Introducing the vectors $\boldsymbol{u}(t) = (u_1(t), \dots, u_{n-1}(t))$ and $\boldsymbol{f}(t) = (f_1(t), \dots, f_{n-1}(t))$, we can write (3.2) as

$$\boldsymbol{u}' = \mathbf{A}\boldsymbol{u} + \boldsymbol{b} \,, \tag{3.3}$$

using the matrix

$$\mathbf{A} = \frac{\kappa}{h^2}\begin{pmatrix} -2 & 1 & & 0 \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ 0 & & 1 & -2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{b} = \boldsymbol{f} + \begin{pmatrix} \kappa/h^2\, u_L(t) \\ 0 \\ \vdots \\ 0 \\ \kappa/h^2\, u_R(t) \end{pmatrix} .$$

The dimension of this ODE system depends on the number n of spatial grid points; in applications one often works with $n = 10^6$ or even more nodes.

In this chapter we consider linear ODE systems of the following form

$$\begin{aligned} u_1' &= a_{11}(t)u_1 + \cdots + a_{1n}(t)u_n + b_1(t) \,, \quad u_1(\xi) = \eta_1 \,, \\ &\vdots \\ u_n' &= a_{n1}(t)u_1 + \cdots + a_{nn}(t)u_n + b_n(t) \,, \quad u_n(\xi) = \eta_n \,. \end{aligned} \tag{3.4}$$

Introducing the notations $\boldsymbol{u} = (u_1, \dots, u_n) : I \mapsto \mathbb{R}^n$, $\boldsymbol{b} = (b_1, \dots, b_n) : I \mapsto \mathbb{R}^n$, $\mathbf{A} = (a_{ij})_{i,j} : I \mapsto \mathbb{R}^{n\times n}$ and $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n) \in \mathbb{R}^n$, we can write (3.4) as

$$\boldsymbol{u}' = \mathbf{A}(t)\boldsymbol{u} + \boldsymbol{b}(t), \quad \boldsymbol{u}(\xi) = \boldsymbol{\eta} \,. \tag{3.5}$$

Definition 3.1. Let $\mathbf{A} \in \mathcal{C}^1(I, \mathbb{R}^{n\times n})$ be a continuously diff'able, *matrix valued* function on an interval $I \subset \mathbb{R}$. We introduce

$$\begin{aligned} \mathbf{A}'(t) &:= \big(a_{ij}'(t)\big)_{i,j} \\ \int \mathbf{A}(t)\,\mathrm{d}t &:= \left(\int a_{ij}(t)\,\mathrm{d}t\right)_{i,j} \\ \operatorname{tr} \mathbf{A}(t) &:= \sum_{i=1}^{n} a_{ii}(t) \qquad\qquad (\textit{trace} \text{ of the matrix}) \,. \end{aligned}$$

Definition 3.2 (Matrix norms)**.** Let $|\cdot|$ be a vector norm in $\mathbb{R}^n$ and $\|\cdot\|$ a matrix norm in $\mathbb{R}^{n\times n} \simeq \mathbb{R}^{n^2}$. The norm $\|\cdot\|$ is called *compatible* with $|\cdot|$, if

$$\begin{aligned} \|\mathbf{A}\mathbf{B}\| &\le \|\mathbf{A}\|\ \|\mathbf{B}\| && \text{(submultiplicative)} \,, \\ |\mathbf{A}\boldsymbol{x}| &\le \|\mathbf{A}\|\ |\boldsymbol{x}| && \text{(compatible)} \,. \end{aligned}$$

Example 3.2. The max–norm $|x| = |x|_\infty = \max_i |x_i|$ is compatible with the *row sum norm* $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$.

The L^1–norm $|x| = |x|_1 = \sum_i |x_i|$ is compatible with the *column sum norm* $\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}|$.

The Euclidean norm $|x| = |x|_2 = \left(\sum_i |x_i|^2\right)^{1/2}$ is compatible with the *spectral norm* $\|\mathbf{A}\|_2 = \rho(\mathbf{A}^H \mathbf{A})$. Here $\mathbf{A}^H = \overline{\mathbf{A}^T}$ denotes the conjugate transposed (Hermitean) matrix and $\rho(\mathbf{A})$ denotes the *spectral radius*, i.e. the magnitude of the largest eigenvalue.

Theorem 3.1. *Let $I \subset \mathbb{R}$ be a compact interval, $\mathbf{A} : I \mapsto \mathbb{R}^{n\times n}$, $\boldsymbol{b} : I \mapsto \mathbb{R}^n$ continuous and $\xi \in I$, $\boldsymbol{\eta} \in \mathbb{R}^n$. Then the IVP*

$$\boldsymbol{u}' = \mathbf{A}(t)\boldsymbol{u} + \boldsymbol{b}(t), \qquad \boldsymbol{u}(\xi) = \boldsymbol{\eta}\ . \tag{3.6}$$

has a unique solution $\boldsymbol{u}(t)$ existing on the entire interval I.

If the estimates

$$\|\mathbf{A}(t)\|_\infty \le L\ , \quad |\boldsymbol{b}(t)|_\infty \le \delta \quad \forall t \in I \quad \textit{and} \quad |\boldsymbol{\eta}|_\infty \le \gamma\ ,$$

hold, the solution $\boldsymbol{u}(t)$ of (3.6) is bounded *by*

$$|\boldsymbol{u}(t)|_\infty \le \gamma e^{L|t-\xi|} + \frac{\delta}{L}\left(e^{L|t-\xi|} - 1\right)\ . \tag{3.7}$$

Proof. Existence and uniqueness can be shown using Picard–Lindelöf.
The estimate (3.7) is left as an exercise. □

Remark 3.1. Complex valued systems can be treated similarly. Separating the real and imaginary parts we obtain a real valued system of double the dimension.

Recall: The ODE $\boldsymbol{u}' = \mathbf{A}\boldsymbol{u}$ is called *homogeneous* and $\boldsymbol{u}' = \mathbf{A}\boldsymbol{u} + \boldsymbol{b}$ is called the *inhomogeneous* equation.

3.1 Homogeneous Linear ODE Systems

If not stated otherwise, we assume $I \subset \mathbb{R}$ to be an interval, $\xi \in I$, $\mathbf{A} : I \mapsto \mathbb{R}^{n\times n}$ continuous and $\boldsymbol{\eta} \in \mathbb{R}^n$. We consider the IVP

$$\boldsymbol{u}' = \mathbf{A}(t)\boldsymbol{u}\ , \qquad \boldsymbol{u}(\xi) = \boldsymbol{\eta}\ . \tag{3.8}$$

Thanks to the Picard–Lindelöf Theorem 2.3, there exist for every initial value $\boldsymbol{u}(\xi) = \boldsymbol{\eta} \in \mathbb{R}^n$ a unique solution $\boldsymbol{u}(\cdot;\boldsymbol{\eta})$ of the homogeneous IVP (3.8). Let

$$S_H := \left\{\boldsymbol{u} \in \mathcal{C}^1(I,\mathbb{R}^n)\ :\ \boldsymbol{u}' = \mathbf{A}\boldsymbol{u}\right\}$$

denote the set of all solutions to the homogeneous linear ODE. Now, we can define the mapping $\Phi : \mathbb{R}^n \mapsto S_H$ by $\Phi(\boldsymbol{\eta}) := \boldsymbol{u}(\cdot;\boldsymbol{\eta})$. In other words, Φ which maps an initial value $\boldsymbol{\eta}$ onto the solution $\boldsymbol{u}(\cdot;\boldsymbol{\eta})$ of the homogeneous IVP (3.8). It is immediate to see, that this map is a bijection and hence the set S_H can be viewed as a vector space of dimension n.

Lemma 3.2. *Let $\boldsymbol{u} \in S_H$ and $\boldsymbol{u}(t) = 0$ for* some $t \in I$. *Then $\boldsymbol{u} \equiv 0$.*

Definition 3.3 (Solution matrix and fundamental matrix)**.** A family $\boldsymbol{u}_1(t), \dots, \boldsymbol{u_n}(t)$ of linear independent solutions to the linear homogeneous ODE $\boldsymbol{u}' = \mathbf{A}(t)\boldsymbol{u}$ is called a *fundamental system.* The matrix

$$\mathbf{U}(t) = (\boldsymbol{u}_1(t)|\cdots|\boldsymbol{u}_n(t)) \in \mathbb{R}^{n\times n}$$

composed out of a fundamental system is called solution matrix of the ODE. It is immediate to see, that the solution matrix is *not* unique.

A particular fundamental system is given by the solutions of the initial value problems $\boldsymbol{x}_i(\xi) = \boldsymbol{e_i}$, i.e. $\boldsymbol{x}_i = \Phi(\boldsymbol{e_i})$. The according solution matrix $\mathbf{X}(t)$ is also called *fundamental matrix.* With the help of the fundamental matrix, the unique solution of (3.8) can be written as

$$\boldsymbol{u}(t) = \mathbf{X}(t)\,\boldsymbol{\eta}\ .$$

Definition 3.4 (Wronski–determinant)**.** Let $\mathbf{U}(t) = (\boldsymbol{u}_1(t)|\cdots|\boldsymbol{u}_n(t))$ be a solution matrix. We call its determinant $w(t) = \det \mathbf{U}(t)$ the W<span style="font-variant:small-caps">ronski</span>–*determinant.*

Lemma 3.3. *The* WRONSKI*–determinant satisfies the homogeneous linear ODE*

$$w' = \operatorname{tr}\mathbf{A}(t)\,w\ .$$

It holds that either $w \equiv 0$ or $w \neq 0$ on the entire interval I. If $w \neq 0$, the columns of the solution matrix $\boldsymbol{u}_1(t), \dots, \boldsymbol{u}_n(t)$ define a fundamental system.

Proof. See [Wal98, Ch. IV, §15 III] □

3.2 Inhomogeneous Systems

Given the inhomogeneous problem

$$\boldsymbol{u}' = \mathbf{A}(t)\boldsymbol{u} + \boldsymbol{b}(t)\ , \quad \boldsymbol{u}(\xi) = \boldsymbol{\eta} \tag{3.9}$$

we define the set of *all* solutions to the inhomogeneous ODE $\boldsymbol{u}' = \mathbf{A}(t)\boldsymbol{u} + \boldsymbol{b}(t)$ as

$$S_I := \left\{\boldsymbol{u} \in \mathcal{C}^1(I,\mathbb{R}^n)\ :\ \boldsymbol{u}' = \mathbf{A}\boldsymbol{u} + \boldsymbol{b}\right\}\ .$$

Proposition 3.4. *Let $\overline{\boldsymbol{u}}$ denote an* arbitrary *solution of the inhomogeneous ODE $\boldsymbol{u}' = \mathbf{A}(t)\boldsymbol{u} + \boldsymbol{b}(t)$. Then $S_I = S_H + \overline{\boldsymbol{u}}$, i.e. for every solution $\boldsymbol{u}$ of the inhomogeneous equation there exists a solution $\boldsymbol{z}$ of the homogeneous problem such that $\boldsymbol{u} = \boldsymbol{z} + \overline{\boldsymbol{u}}$.*
(General solution of inhomogeneous problem
= General solution of homogeneous problem + particular solution)

To obtain a *particular* solution of the inhomogeneous problem, we use the variation of constants.

Let $\mathbf{X}(t)$ be the fundamental matrix of the homogeneous equation $\boldsymbol{x}' = \mathbf{A}(t)\boldsymbol{x}$. Then, the general solution of the homogeneous problem equals $\boldsymbol{z}(t) = \mathbf{X}(t)\,\boldsymbol{c}$ for some constant $\boldsymbol{c} \in \mathbb{R}^n$. Analogously to the scalar case (see Section 1.5 on page 18) we seek the solution of the inhomogeneous problem (3.9) using the ansatz $\boldsymbol{u}(t) = \mathbf{X}(t)\,\boldsymbol{c}(t)$, where the constant $\boldsymbol{c}$ is replaced by some function $\boldsymbol{c}(t)$. It holds that

$$\begin{aligned} \boldsymbol{u}'(t) &= \mathbf{X}'(t)\,\boldsymbol{c}(t) + \mathbf{X}(t)\,\boldsymbol{c}'(t) = \mathbf{A}(t)\,\mathbf{X}(t)\,\boldsymbol{c}(t) + \mathbf{X}(t)\,\boldsymbol{c}'(t) \\ &\overset{!}{=} \mathbf{A}(t)\,\boldsymbol{u}(t) + \boldsymbol{b}(t) \end{aligned}$$

and hence

$$\boldsymbol{b}(t) = \mathbf{X}(t)\,\boldsymbol{c}'(t)\ .$$

Since $\mathbf{X}(t)$ is a fundamental matrix, we get $\det \mathbf{X}(t) \neq 0$, and

$$\boldsymbol{c}'(t) = \mathbf{X}^{-1}(t)\,\boldsymbol{b}(t)\ .$$

Integration leads to

$$\boldsymbol{c}(t) = \boldsymbol{c}(\xi) + \int_\xi^t \mathbf{X}^{-1}(s)\,\boldsymbol{b}(s)\,\mathrm{d}s\ .$$

This shows the following

Theorem 3.5. *The unique solution of the inhomogeneous initial value problem (3.9) is given by*

$$\boldsymbol{u}(t) = \mathbf{X}(t)\left(\boldsymbol{\eta} + \int_\xi^t \mathbf{X}^{-1}(s)\,\boldsymbol{b}(s)\,\mathrm{d}s\right)\ .$$

Remark 3.2. We compare the scalar and the vectorial case (both with constant coefficients) :

	scalar	vectorial
homogen.	$x' = ax, \quad x(0) = \eta$	$\boldsymbol{x}' = \mathbf{A}\boldsymbol{x}, \quad \boldsymbol{x}(0) = \boldsymbol{\eta}$
solution	$x = e^{at}\eta$	$\boldsymbol{x} = \mathbf{X}(t)\,\boldsymbol{\eta}$
inhom.	$u' = au + b(t), \quad u(0) = \eta$	$\boldsymbol{u} = \mathbf{A}\boldsymbol{u} + \boldsymbol{b}(t), \quad \boldsymbol{u}(0) = \boldsymbol{\eta}$
solution	$u = e^{at}\left(\eta + \int_0^t e^{-as} b(s)\,\mathrm{d}s\right)$	$\boldsymbol{u} = \mathbf{X}(t)\left(\boldsymbol{\eta} + \int_0^t \mathbf{X}^{-1}(s)\,\boldsymbol{b}(s)\,\mathrm{d}s\right)$

How to obtain the fundamental system $\mathbf{X}(t)$ of a system of homogeneous linear ODEs with constant coefficients $\boldsymbol{x}' = \mathbf{A}\boldsymbol{x}$?
Does there exist an analogy to the scalar case, i.e. the exponential e^{at}?

3.3 The Matrix–Exponential Function

In this section we consider a system of linear ODEs with *constant coefficients*. The fundamental matrix $\mathbf{X}(t)$ is the solution of the matrix valued initial value problem

$$\mathbf{X}'(t) = \mathbf{A}\,\mathbf{X}(t), \qquad \mathbf{X}(0) = \mathbf{E}\,, \tag{3.10}$$

for $\mathbf{A} \in \mathbb{R}^{n\times n}$ *constant*. Wlog we assume $\xi = 0$.

According to Picard–Lindelöf (Thm. 2.3) and Banach (Thm. 2.2) we know:

- ▷ The solution is unique and exists for all $t \in \mathbb{R}$.
- ▷ The fixpoint iteration

  $$\mathbf{X}_{k+1} = \mathbf{E} + \int_0^t \mathbf{A}\,\mathbf{X}_k\,\mathrm{d}s$$

  converges for every initial guess $\mathbf{X}_0 \in \mathbb{R}^{n\times n}$ to the solution $\mathbf{X}(t)$ of (3.10), i.e. to the fundamental matrix.

The k-th fixpoint iterate for the initial guess $\mathbf{X}_0 = \mathbf{E}$ reads as

$$\mathbf{X}_k(t) = \mathbf{E} + \mathbf{A}t + \frac{\mathbf{A}^2 t^2}{2} + \cdots + \frac{\mathbf{A}^k t^k}{k!}\ .$$

Its scalar analogue

$$x_k(t) = 1 + at + \frac{a^2 t^2}{2} + \cdots + \frac{a^k t^k}{k!}$$

converges to $x(t) = e^{at}$. This motivates the following

Definition 3.5 (Matrix–Exponential–Function)**.** The matrix–valued series

$$\sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}$$

converges for all constant matrices $\mathbf{A} \in \mathbb{R}^{n\times n}$ or $\mathbb{C}^{n\times n}$. It defines the so–called *matrix–exponential–function*

$$\exp(\mathbf{A}) = e^{\mathbf{A}} := \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} \in \mathbb{R}^{n\times n} \text{ or } \mathbb{C}^{n\times n} \ .$$

Theorem 3.6. *The fundamental matrix of the homogeneous ODE with constant coefficients is given by*

$$\mathbf{X}(t) = e^{\mathbf{A}t} \ .$$

Remark 3.3. Based on the analogy to the scalar case, one may assume, that the fundamental matrix for a problem with *non–constant* coefficients

$$\mathbf{X}' = \mathbf{A}(t)\,\mathbf{X}$$

is also given in the form

$$\mathbf{X}(t) = \exp\left[\int_0^t \mathbf{A}(s)\,\mathrm{d}s\right] \ .$$

This is in general **not** true!

However, if $\mathbf{A}(t)$ and $\int \mathbf{A}(s)\,\mathrm{d}s$ commute, then the above formula is true.

Example 3.3. The solution of the linear ODE system

$$\boldsymbol{x}' = \mathbf{A}(t)\,\boldsymbol{x} \quad \text{where} \quad \mathbf{A}(t) = \begin{pmatrix} 1 & 2t \\ 0 & 0 \end{pmatrix}$$

is **not** given by

$$\boldsymbol{x}(t) = \exp\left(\int_0^t \mathbf{A}(s)\,\mathrm{d}s\right) \ .$$

Details of this will be worked out in the tutorials.

Theorem 3.7 (Properties of $e^{\mathbf{A}}$)**.** *Analogous to the scalar case it holds that*

(1) $\dfrac{d}{dt} e^{\mathbf{A}t} = \mathbf{A} \cdot e^{\mathbf{A}t}$

(2) Let $\Lambda = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n)$ *be a diagonal matrix. Then*

$$e^{\Lambda} = \mathbf{diag}\left(e^{\lambda_1}, \ldots, e^{\lambda_n}\right) \ .$$

(3) If $\mathbf{B}$ *and* $\mathbf{C}$ *commute, i.e.* $\mathbf{BC} = \mathbf{CB}$*, then*

$$e^{\mathbf{B}+\mathbf{C}} = e^{\mathbf{B}}\, e^{\mathbf{C}} = e^{\mathbf{C}}\, e^{\mathbf{B}} \,.$$

(4) If $\mathbf{C}$ *is invertible, i.e.* $\det \mathbf{C} \neq 0$*, then*

$$e^{\mathbf{CBC}^{-1}} = \mathbf{C}\, e^{\mathbf{B}}\, \mathbf{C}^{-1} \,.$$

(5) $\left(e^{\mathbf{A}}\right)^{-1} = e^{-\mathbf{A}}$

(6) $e^{\mathbf{A}(s+t)} = e^{\mathbf{A}s}\, e^{\mathbf{A}t} = e^{\mathbf{A}t}\, e^{\mathbf{A}s}$

(7) $e^{\mathbf{A}+\lambda\mathbf{E}} = e^{\mathbf{A}} e^{\lambda} = e^{\lambda}\, e^{\mathbf{A}}$

The properties (2), (4) and (7) show an efficient way to compute $e^{\mathbf{A}}$.

Proposition 3.8. *Let* $\mathbf{A}$ *be diagonalizable, i.e. there exists* $\mathbf{Q} \in \mathbb{C}^{n\times n}$*,* $\det \mathbf{Q} \neq 0$ *such that* $\mathbf{A} = \mathbf{Q} \cdot \mathbf{diag}\,(\lambda_1, \dots, \lambda_n) \cdot \mathbf{Q}^{-1}$*. Then*

$$e^{\mathbf{A}} = \mathbf{Q} \cdot \mathbf{diag}\left(e^{\lambda_1}, \dots, e^{\lambda_n}\right) \cdot \mathbf{Q}^{-1} \,.$$

What to do, if $\mathbf{A}$ is not diagonalizable? Use the JORDAN normalform (Thm. A.3)

Proposition 3.9. *Let* $\mathbf{J}_k \in \mathbb{C}^{m_k \times m_k}$ *be a* JORDAN*–block of dimension* m_k *to the eigenvalue* λ_k*. Then*

$$e^{\mathbf{J}_k t} = e^{\lambda t} \begin{pmatrix} 1 & t & t^2/2 & \cdots & t^{m_k-1}/(m_k-1)! \\ & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & \ddots & t^2/2 \\ & & & \ddots & t \\ 0 & & & & 1 \end{pmatrix} . \tag{3.11}$$

Proof. As an exercise. □

Theorem 3.10. *Let* $\mathbf{A} = \mathbf{Q} \cdot \mathbf{diag}\,(\mathbf{J}_1, \dots, \mathbf{J}_k) \cdot \mathbf{Q}^{-1}$ *be the* JORDAN *normalform of* $\mathbf{A}$*. Then*

$$e^{\mathbf{A}t} = \mathbf{Q} \cdot \mathbf{diag}\left(e^{\mathbf{J}_1 t}, \dots, e^{\mathbf{J}_k t}\right) \cdot \mathbf{Q}^{-1} \,.$$

Proof. As an exercise. □

How to do this in a practical example?

Example 3.4. We consider the ODE

$$\boldsymbol{x}' = \mathbf{A}\boldsymbol{x}, \quad \text{where} \quad \mathbf{A} = \begin{pmatrix} 1 & -1 \\ 4 & -3 \end{pmatrix} .$$

The characteristic polynomial of $\mathbf{A}$ is given by $\chi_A(\lambda) = (\lambda+1)^2$. Hence $\lambda = -1$ is the only eigenvalue with algebraic multiplicity 2. If A is diagonalizable, we expect each component of solution to be of the form ce^{-t}. On the other hand, if A is not diagonalizable, the Jordan form (cf. (3.11)) yields each component to be of the form e^{-t} times a polynomial of degree less or equal to one. Hence, in both cases we seek the solution of the ODE of the form

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a+bt \\ c+dt \end{pmatrix} e^{-t}$$

where the coefficients a, b, c and d are still to be determined. Inserting this into the ODE yields

$$\boldsymbol{x}' = \begin{pmatrix} -a+b-bt \\ -c+d-dt \end{pmatrix} e^{-t} \overset{!}{=} \mathbf{A}\boldsymbol{x} = \begin{pmatrix} a-c+(b-d)t \\ 4a+3c+(4b-3d)t \end{pmatrix} e^{-t}$$

Comparing coefficients yields (up to scalar multiples) the following two cases

(1) $b=0$, $d=0$, $a=1$, $c=2$, hence $\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} e^{-t}$.

(2) $b=1$, $d=2$, $a=0$, $c=-1$, hence $\begin{pmatrix} x_2 \\ y_2 \end{pmatrix} = \begin{pmatrix} t \\ -1+2t \end{pmatrix} e^{-t}$.

The according solution matrix $\mathbf{Y}(t)$ reads as

$$\mathbf{Y}(t) = \begin{pmatrix} 1 & t \\ 2 & -1+2t \end{pmatrix} e^{-t}, \quad \text{and} \quad \mathbf{Y}(0) = \begin{pmatrix} 1 & 0 \\ 2 & -1 \end{pmatrix} .$$

Column operations (linear combinations of fundamental solutions) finally yields the fundamental matrix

$$\mathbf{X}(t) = \begin{pmatrix} 1+2t & -t \\ 4t & 1-2t \end{pmatrix} e^{-t}, \quad \text{and} \quad \mathbf{X}(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{E}.$$

Corollary 3.11 (to Theorem 3.5)**.** *The inhomogeneous system*

$$\boldsymbol{y}' = \mathbf{A}\boldsymbol{y} + \boldsymbol{b}(t), \quad \boldsymbol{y}(\xi) = \boldsymbol{\eta}$$

has the unique solution

$$\boldsymbol{y} = e^{\mathbf{A}(t-\xi)}\boldsymbol{\eta} + \int_\xi^t e^{\mathbf{A}(t-s)}\boldsymbol{b}(s)\,\mathrm{d}s .$$

3.4 Higher Order Differential Equations

The linear n–th order ODE

$$u^{(n)} + a_{n-1}(t)u^{(n-1)} + \cdots + a_1(t)u' + a_0(t)u = b(t) \tag{3.12}$$

is equivalent to the system

$$\boldsymbol{u}' = \mathbf{A}(t)\,\boldsymbol{u} + \boldsymbol{b}(t)\ ,$$

setting $\boldsymbol{u} = \big(u, u', \dots, u^{(n-1)}\big)$ and $\boldsymbol{b}(t) = (0, \dots, 0, b(t))$ as well as the matrix

$$\mathbf{A}(t) = \begin{pmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & & & 1 \\ -a_0(t) & -a_1(t) & \cdots & -a_{n-1}(t) \end{pmatrix} .$$

Theorem 3.12. *Let $a_k : I \to \mathbb{R}$, $k = 0, \dots, n-1$ and $b : I \to \mathbb{R}$ be continuous functions. Then the linear ODE (3.12) of order n admits for any initial value*

$$u(\xi) = \eta_0, \quad u'(\xi) = \eta_1, \quad \dots \quad , u^{(n-1)} = \eta_{n-1}$$

a unique *solution.*

The set $S_H := \big\{u \in \mathcal{C}^n(I)\ : u^{(n)} + a_{n-1}(t)u^{(n-1)} + \cdots + a_1(t)u' + a_0(t)u = 0\big\}$ of all solutions to the homogeneous *problem is a vector space of dimension n.*

Let $\overline{u}$ be an arbitrary solution of the inhomogeneous *problem (3.12). Then*

$$S_I := \Big\{u \in \mathcal{C}^n(I)\ : u \textit{ satisfies } (3.12)\Big\} = \overline{u} + S_H$$

is the affine linear space of all solution of the inhomogeneous problem.

Let $x_0, \dots, x_{n-1}$ be a fundamental system to the initial values $x_i^{(k)}(0) = \delta_{ik}$ for $i, k = 0, \dots, n-1$. The resulting fundamental matrix is given by

$$\mathbf{X}(t) = \begin{pmatrix} x_0(t) & x_1(t) & \cdots & x_{n-1}(t) \\ x_0'(t) & x_1'(t) & & x_{n-1}'(t) \\ \vdots & & & \vdots \\ x_0^{(n-1)}(t) & x_1^{(n-1)}(t) & \cdots & x_{n-1}^{(n-1)}(t) \end{pmatrix} .$$

The Wronskian $w(t) = \det \mathbf{X}(t)$ satisfies the ODE

$$w' = -a_{n-1}(t)w\ .$$

In the sequel we restrict to the case of *constant coefficients*, i.e. we assume the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ to be constant.

Let $b : I \to \mathbb{R}$ be continuous and $a_0, \ldots, a_{n-1} \in \mathbb{R}$. We consider the ODE

$$u^{(n)} + a_{n-1}u^{(n-1)} + \cdots + a_1 u' + a_0 u = b(t) \tag{†}$$

or the equivalent system

$$\boldsymbol{u}' = \mathbf{A}\,\boldsymbol{u} + \boldsymbol{b}(t)$$

where

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \\ -a_0 & -a_1 & \cdots & -a_{n-1} \end{pmatrix} .$$

The characteristic polynomial $\chi_{\mathbf{A}}(\lambda) = \det(\mathbf{A} - \lambda\mathbf{E})$ is also called the *characteristic polynomial of the ODE* (†).

Lemma 3.13. *The characteristic polynomial of the ODE* (†) *is given by*

$$\chi_{\mathbf{A}}(\lambda) = (-1)^n \left(\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_1\lambda + a_0\right) .$$

Proof. Left as an exercise. □

Theorem 3.14. *Let $\lambda_1, \ldots, \lambda_m$ be the roots of the characteristic polynomial of the ODE* (†) *with multiplicities $k_1, \ldots, k_m$. Then the functions*

$$\left\{e^{\lambda_1 t}, te^{\lambda_1 t}, \ldots, t^{k_1-1}e^{\lambda_1 t}; \ldots; e^{\lambda_m t}, te^{\lambda_m t}, \ldots, t^{k_m-1}e^{\lambda_m t}\right\}$$

constitute a fundamental system of (†).

Remark 3.4. There might appear complex roots $\lambda = \alpha + i\omega$. Since all the coefficients $a_0, \ldots, a_{n-1}$ are assumed to be real, the roots appear in pairs with their complex conjugate $\overline{\lambda} = \alpha - i\omega$. A real–valued fundamental system can be obtained replacing the pairs of complex conjugate solutions

$$\left\{e^{\lambda t}, te^{\lambda t}, \ldots, t^{k-1}e^{\lambda t};\ e^{\overline{\lambda} t}, te^{\overline{\lambda} t}, \ldots, t^{k-1}e^{\overline{\lambda} t}\right\}$$

by their real counterparts

$$e^{\alpha t} \cdot \left\{\cos(\omega t), t\cos(\omega t), \ldots, t^{k-1}\cos(\omega t);\ \sin(\omega t), t\sin(\omega t), \ldots, t^{k-1}\sin(\omega t)\right\} .$$

Example 3.5 (Linear second order ODE with constant coefficients)**.**

The ODE

$$u'' + 2au' + bu = 0 \tag{3.13}$$

can be used to model a *damped oscillator*.
A mass $m > 0$ is attached to a spring with spring constant $k > 0$, $z(t)$ denotes the deviation from equilibrium

$$mz'' = -kz \ .$$

According to STOKES law, the viscous friction in the fluid equals $F_R = 6\pi\eta r v$ where η is the viscosity of the fluid, r denotes the radius and $v = z'$ the velocity. Hence we obtain

$$mz'' = -\beta z' - kz \ ,$$

using $\beta = 6\pi\eta r > 0$.

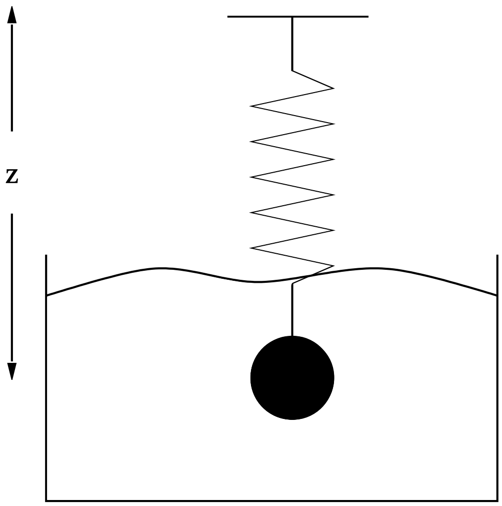


Figure 3.1: Sketch of a damped oscillator.

The characteristic polynomial of ODE (3.13) is given by $\lambda^2 + 2a\lambda + b$. Its roots are

$$\lambda_{1,2} = -a \pm \sqrt{a^2 - b}, \qquad \delta^2 = a^2 - b \ .$$

Based on the discriminant δ we distinguish three cases:

(1) The *overdamped case* $\delta^2 > 0$. Here we have two real roots $\lambda_1 = -a + \delta$ and $\lambda_2 = -a - \delta$. The fundamental system reads as $u_1 = e^{\lambda_1 t}$ and $u_2 = e^{\lambda_2 t}$. The general solution is therefore given by

$$u = \alpha e^{\lambda_1 t} + \beta e^{\lambda_2 t}$$

where $\alpha, \beta \in \mathbb{R}$.

(2) The *critically damped case* $\delta^2 = 0$ with one root $\lambda = -a$ of multiplicity two. The fundamental system equals $u_1 = e^{\lambda_1 t}$ and $u_2 = te^{\lambda_2 t}$ and the general solution is given by

$$u = \alpha e^{\lambda_1 t} + \beta t e^{\lambda_2 t}$$

where $\alpha, \beta \in \mathbb{R}$.

(3) The *oscillatory case* for $\delta^2 < 0$. In this situation we encounter two complex conjugate roots $\lambda, \overline{\lambda} = -a \pm i\,|\delta|$. The complex–valued fundamental system

$$u_1 = e^{\lambda t} = e^{-at+i|\delta|t} \quad \text{and} \quad u_2 = e^{\overline{\lambda}t} = e^{-at-i|\delta|t}$$

leads to the following real–valued fundamental system

$$\begin{aligned} \tilde{u}_1 &= (u_1+u_2)/2 &&= e^{-at}\cos(|\delta|\,t)\ , \\ \tilde{u}_2 &= (u_1+u_2)/(2i) &&= e^{-at}\sin(|\delta|\,t) \end{aligned}$$

and the general solution $(\alpha, \beta \in \mathbb{R})$

$$u = \alpha e^{-at}\cos(|\delta|\,t) + \beta e^{-at}\sin(|\delta|\,t)\ .$$

In the physically relevant case $a > 0$, this solution describes a *damped oscillation*.

The following Figure 3.2 shows the solutions using $b = 1$, $u(0) = 2$, $u'(0) = 0$ and damping parameter $a = \sqrt{2}$ (overdamped), $a = 1$ (critically damped) and $a = 1/4$ (damped oscillation). The left hand side shows the solution $u(t)$; the right hand side depicts the trajectories in the $u\,u'$–phase space.

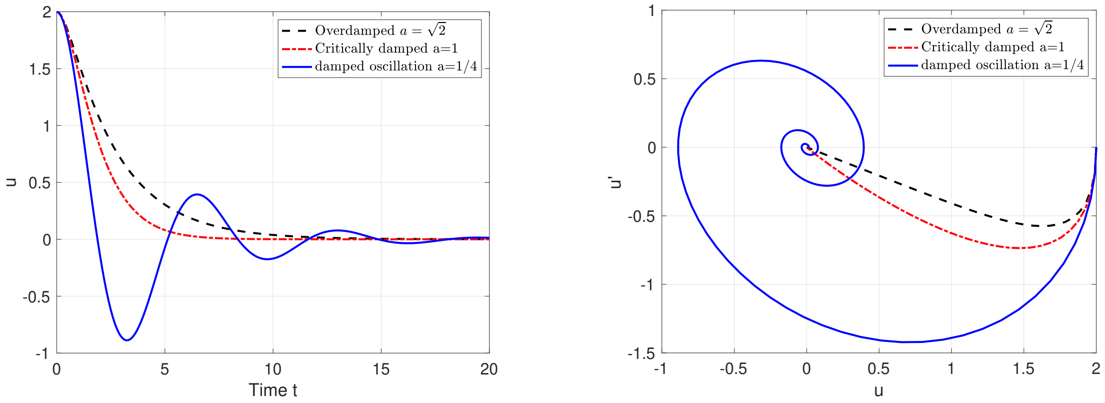


Figure 3.2: Solution trajectories (left) and phase portrait (right) of oscillator (3.13). Overdamped ($a = \sqrt{2}$; dashed), critically damped ($a = 1$; dash–dotted) and oscillating ($a = 1/4$; solid line) case.

Chapter 4

Numerical Methods

In the sequel we consider the following initial value problem for a first order ODE

$$u' = f(t,u), \quad u(t_0) = u_0 \tag{4.1}$$

on an interval $t \in I = [t_0, T]$. The right hand side $f : \Omega \to \mathbb{R}^n$ is assumed to be smooth, i.e. continuously differentiable up to needed orders and hence in particular locally *Lipschitz*–continuous on $\Omega \subset I \times \mathbb{R}^n$.

Example 4.1. For $a \in \mathbb{R}$, the solution of the IVP

$$u' = a\,u, \quad u(0) = 1$$

is given by $u(t) = \exp(at)$. Using PYTHON, the following code–snippet solves this IVP on the interval $t \in [0,3]$.

```
from numpy import *
from scipy.integrate import solve_ivp
def f(t,u, a):
        return a*u

a = 1.5
Sol = solve_ivp(f, [0, 3], [1], args=(a,))
t = Sol.t
u = Sol.y[0,:]
print(t)
print(u)
```

When running this code, we observe, that the t–values are strangely spaced inside the interval $[0,3]$. For different values of the parameter a, the solver solve_ivp even chooses different time points for computing the solution. Why?

Definition 4.1 (Evolution)**.** Consider a differential equation $u' = f(t,u)$ with Lipschitz–continuous right hand side $f : \Omega \to \mathbb{R}^n$. For an initial value $u(t_0) = u_0$ we denote the according *unique* solution $u(t)$ as

$$u(t) = \Phi^{t,t_0} u_0 \ .$$

The operator $\Phi^{t,t_0} : \mathbb{R}^n \to \mathbb{R}^n$ maps the initial value $u_0 \in \mathbb{R}^n$ at time t_0 onto the solution $u(t) \in \mathbb{R}^n$ evaluated at time t. The two–parameter family of maps $\Phi^{\cdot,\cdot}$ is called the *evolution* of the differential equation.

Lemma 4.1. *Let Φ be the evolution of a differential equation. Then for all initial values $(t_0, u_0) \in \Omega$ the following holds*

(1) $\Phi^{t_0,t_0} u_0 = u_0$,

(2) $\Phi^{t,s}\Phi^{s,t_0} u_0 = \Phi^{t,t_0} u_0$ *for all* $t, s \in I$.

Proof. As an exercise. □

Example 4.2 (Euler–Method (Idea 1768))**.** To determine the evolution $\Phi^{t,t_0} u_0$ of a differential equation, we use an *approximation* of the derivative

$$f(t, u(t)) = u'(t) \simeq \frac{u(t+h) - u(t)}{h}$$

for h small. This yields the following linear approximation of $u(t+h)$

$$u(t+h) = u(t) + hf(t, u(t)) \ .$$

Iterating this, we can construct an approximate solution $u_k \simeq u(t_k)$ at *discrete* time points $t_k = t_0 + kh$, $k \in \mathbb{N}$ by

$$u_{k+1} = u_k + hf(t_k, u_k) \ .$$

Graphically, we may interpret this approximate solution as a piecewise linear curve in the directional field of the differential equation. This piecewise linear curve is also known as the Euler–polygon.

Python Example 4.1 (Euler–method)**.**

```
from numpy import *

def euler(f,t,u0):
    # f: f(t,u), t scalar, u vector; rhs of the ODE
    # t: vector of discrete time points
    # u0: vector of initial values
    # returns U: solution as matrix, rows = componentes

    n = len(t)
    m = len(u0)
    U = zeros((m,n))
    U[:,0] = u0

    for k in range(n-1):
        h = t[k+1]-t[k]
        U[:,k+1] = U[:,k] + h*f(t[k], U[:,k])

    return U
```

The above introduced approximation is an example of an *explicit one–step–method*, since the new value for the approximate u_{k+1} can be explicitly computed knowing the previous value u_k; no system of equations needs to be solved.

Alternatively, we may write the ODE as an integral equation

$$u(t) = u(t_0) + \int_{t_0}^{t} f(s, u(s))\, ds$$

and try to evaluate the integral over an interval $[t_0, t_0 + h]$ by a suitable numerical quadrature.

The *left–sided rectangle rule*

$$\int_{t_0}^{t_0+h} f(s, u(s))\, ds \simeq h \cdot f(t_0, u(t_0))$$

leads to

$$u(t_0 + h) \simeq u(t_0) + hf(t_0, u(t_0))\ .$$

We recover the previous (explicit) Euler–method.

The *right–sided rectangle rule*

$$\int_{t_0}^{t_0+h} f(s, u(s))\, ds \simeq h \cdot f(t_0 + h, u(t_0 + h))$$

leads to the so–called *implicit* Euler–method

$$u(t_0 + h) \simeq u(t_0) + hf(t_0 + h, u(t_0 + h))\ .$$

Here the new value $u(t_0+h)$ appears on both sides of the equations. Hence it is only given implicitly and in general we need to solve a non–linear system for $u(t_0+h)$.

The *midpoint–rule*

$$\int_{t_0}^{t_0+h} f(s,u(s))\,ds \simeq h\cdot f(t_0+h/2, u(t_0+h/2))$$

leads to

$$u(t_0+h) \simeq u(t_0) + h\cdot f(t_0+h/2, u(t_0+h/2))\ .$$

Replacing the yet unknown value $u(t_0+h/2)$ by the Euler–approximation $u_0 + h/2\cdot f(t_0,u_0)$, we obtain the Runge– or improved Euler–method

$$u(t_0+h) \simeq u(t_0) + h\cdot f\left(t_0+\frac{h}{2}, u(t_0)+\frac{h}{2}f(t_0,u(t_0))\right)\ .$$

4.1 Convergence Theory

For the sake of simplicity, we are just considering scalar ODEs in this section. All the results also apply for systems of ODEs, replacing the scalar values of u by appropriate vectors.

Definition 4.2 (Grid)**.** Consider an interval $[t_0,T]\subset\mathbb{R}$ and a subdivision called *grid* $\mathcal{T}_h = \{t_i :\ i=0,\dots,n\}\subset[t_0,T]$ where $t_0 < t_1 < \cdots < t_n = T$. We define the *step size* $h_i = t_{i+1}-t_i$ for $i=0,\dots,n-1$ and the *step size vector* $\boldsymbol{h} = (h_0,\dots,h_{n-1})$. The maximal step size is denoted by $h_{\max} = |h| = \max_i h_i$. If the step sizes are constant, i.e. $h_i = h$ for all i, we call the grid *equidistant.*
A second grid $\mathcal{S} = \{s_j :\ j = 0\dots m\}$ is called *refinement* of $\mathcal{T}$, if $\mathcal{T}\subset\mathcal{S}$ and $h_{\max}(\mathcal{S}) < h_{\max}(\mathcal{T})$.
Let $u : [t_0,T]\to\mathbb{R}$ be a function. We call the vector

$$u_{h,i} = u_i = u(t_i)\in\mathbb{R}^{n+1}$$

the *grid function associated* to u. To compare a function $y : [t_0,T]\to\mathbb{R}$ to a grid function $u_h\in\mathbb{R}^{n+1}$ we use

$$\|y-u_h\| := \max_{i=0\dots n}|y(t_i)-u_{h,i}|\ .$$

For the sake of shorter notation, we often denote the grid function u_h also by u, if it is clear, that we are speaking about a grid function and if confusion with the function itself can be excluded.

Definition 4.3 (Increment, numerical evolution)**.** A numerical method to solve the IVP (4.1) is called *one–step method*, if there exists an *increment* (function) $\psi : [t_0, T] \times \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ such that

$$u_{j+1} = u_j + h_j \psi(t_j, u_j; h_j)$$

for $j = 0, \ldots, n-1$ and $u_0 = u(t_0)$.
We call $\Psi^{t_{j+1}, t_j} u_j := u_j + h_j \psi(t_j, u_j; h_j)$ the *numerical evolution* of the method.

Example 4.3. The Euler–method is a one–step method with increment

$$\psi(t, u, h) = f(t, u) \ .$$

The Runge–method uses the increment

$$\psi(t, u, h) = f\left(t + \frac{h}{2},\, u + \frac{h}{2} f(t, u)\right) \ .$$

For a given IVP (4.1) with (exact) evolution Φ we will try to construct a numerical evolution Ψ which is as close as possible to the exact evolution Φ. This leads us to the following

Definition 4.4 (Consistency)**.** Let $u(t)$ denotes the exact solution of the IVP. We call

$$\delta_h(t, u) = \frac{1}{h}\left(\Phi^{t+h,t} u(t) - \Psi^{t+h,t} u(t)\right) = \frac{u(t+h) - u(t)}{h} - \psi(t, u(t), h)$$

the *local discretization error* or *consistency error* of the numerical evolution Ψ.
We call the numerical evolution Ψ or its defining method *consistent*, if

$$\max_{t \in [t_0, T]} |\delta_h(t, u)| \longrightarrow 0$$

for step size $h \to 0$.
We call the numerical evolution *consistent of order* $p \in \mathbb{N}$, if

$$\delta_h(t, u) = \mathcal{O}(h^p) \ .$$

Remark 4.1. Some authors, cf. Deuflhard [DB02], define the local discretization error alternatively by $\delta_h(t, u) = \Phi^{t+h,t} u(t) - \Psi^{t+h,t} u(t)$. However, then the condition for consistency order p reads as $\delta_h(t, u) = \mathcal{O}(h^{p+1})$.

Lemma 4.2. *The following are equivalent:*

(1) The numerical evolution Ψ is consistent*.*

(2) The increment satisfies $\psi(t,u,0) = f(t,u)$.

Proof. As an exercise. □

Theorem 4.3. *The explicit* Euler*–method is consistent of order $p = 1$.*

Proof. We start from the definition of the local discretization error

$$\delta_h = \frac{1}{h}\left(u(t+h) - u(t)\right) - \psi(t,u,h)\ .$$

Taylor expansion yields

$$\delta_h = \left(u'(t) + \frac{h}{2}u''(t) + \mathcal{O}(h^2)\right) - \psi(t,u,h)\ .$$

Plugging in the increment $\psi(t,u,h) = f(t,u)$ of the Euler–method and using the ODE $u' = f(t,u)$, we obtain

$$\delta_h = \frac{h}{2}u''(t) + \mathcal{O}(h^2) = \mathcal{O}(h)\ .$$

□

Definition 4.5 (Convergence)**.** Let $v(t) = \Phi^{t,t_0}u_0$ denote the exact solution of the IVP (4.1) and let u_h be a numerical solution. We call

$$e_h := \left\|v - u_h\right\|_\infty = \max_{k=1\ldots n}|v(t_k) - u_{h,k}|$$

the *global error* of the numerical method on the grid $\mathcal{T}_h$.
We call the method *convergent*, if $e_h \longrightarrow 0$ for step size $h \to 0$. We call the method *convergent of order* $p \in \mathbb{N}$, if

$$e_h = \mathcal{O}(h^p)\ .$$

The following convergence theorem requires a discrete version of the Gronwall Lemma as a tool.

Lemma 4.4 (Gronwall, discrete version)**.** *Let $(e_n)_{n\in\mathbb{N}}$, $(p_n)_{n\in\mathbb{N}}$ and $(q_n)_{n\in\mathbb{N}}$ be non–negative sequences. Assume that*

$$e_{n+1} \le (1+q_n)e_n + p_n$$

holds for all n, then the elements of the sequence (e_n) can be estimated by

$$e_n \le \left(e_0 + \sum_{j=0}^{n-1} p_j\right)\cdot\exp\left[\sum_{j=0}^{n-1} q_j\right]\ .$$

Proof. As an exercise. □

Theorem 4.5 (Convergence of one–step methods)**.** *Consider the IVP* (4.1) *and its solution* $u(t) = \Phi^{t,t_0} u_0$*. Furthermore let* Ψ *be the numerical evolution of a one–step method with increment* ψ*. Assume that*

(1) The increment ψ *is continuous on the set*

$$G = \{(t, u, h) :\ t_0 \le t \le T,\ |u - u(t)| \le \gamma \textit{ and } 0 \le h \le h_0\}$$

for some $\gamma, h_0 > 0$*. This rather technical condition ensures, that the increment is well defined at least in a neighborhood of the solution trajectory.*

(2) There exists a constant $M > 0$*, such that*

$$|\psi(t, u, h) - \psi(t, v, h)| \le M\,|u - v|$$

for all $(t, u, h), (t, v, h) \in G$*, i.e. the increment is* Lipschitz*–continuous w.r.t. u. This condition is also called* stability *of the method.*

(3) The method is consistent *or even* consistent of order *p, i.e.*

$$|\delta_h(t, u)| \longrightarrow 0 \quad \textit{or} \quad \delta_h(t, u) = \mathcal{O}(h^p)\ .$$

Then the method is convergent *or even* convergent of order *p*

$$e_h \longrightarrow 0 \quad \textit{or} \quad e_h = \mathcal{O}(h^p)\ .$$

In short: Consistency + Stability $\Longrightarrow$ Convergence.

Proof. We proceed in several steps.

(1) First (a rather technical step) we define an extension of the increment

$$\tilde{\psi}(t, u, h) = \begin{cases} \psi(t, u, h) & \text{for } (t, u, h) \in G \\ \psi(t, u(t) \pm \gamma, h) & \text{for } |u - u(t)| > \gamma\ . \end{cases}$$

Then the extended increment function $\tilde{\psi}$ is continuous not only on G but even on the larger set $\tilde{G} = \{(t, u, h) :\ t_0 \le t \le T,\ u \in \mathbb{R} \text{ and } 0 \le h \le h_0\}$ and it holds that

$$\left|\tilde{\psi}(t, u, h) - \tilde{\psi}(t, v, h)\right| \le M\,|u - v|\ . \tag{$*$}$$

Using $\tilde{\psi}(t, u(t), h) = \psi(t, u(t), h)$ we obtain from *consistency*

$$\delta_h(t, u) = \left|\frac{u(t + h) - u(t)}{h} - \tilde{\psi}(t, u(t), h)\right| \longrightarrow 0$$

or, if the method is even consistent of order p, then there exists K_1, such that

$$\delta_h(t, u) \le K_1\,|h|^p\ .$$

(2) Next, we consider the numerical solution $\tilde{u}_h$ generated by the extended increment $\tilde{\psi}$ on a grid $\mathcal{T}_h$ with step size $|h| < h_0$

$$\tilde{u}_0 := u(t_0) \quad \text{and} \quad \tilde{u}_{i+1} = \tilde{u}_i + h_i \tilde{\psi}(t_i, \tilde{u}_i, h) \ .$$

The *exact* solution u satisfies on this grid the equation

$$u(t_{i+1}) = u(t_i) + h_i \frac{u(t_{i+1}) - u(t_i)}{h_i} \ .$$

Now, the *error* $\tilde{e}_i = \tilde{u}_i - u(t_i)$ satisfies

$$\begin{aligned} \tilde{e}_{i+1} &= \tilde{e}_i + h_i \left[\tilde{\psi}(t_i, \tilde{u}_i, h_i) - \frac{u(t_{i+1}) - u(t_i)}{h_i} \right] \\ &= \tilde{e}_i + h_i \left[\tilde{\psi}(t_i, \tilde{u}_i, h) - \tilde{\psi}(t, u(t_i), h) \right] \\ &\quad + h_i \left[\tilde{\psi}(t_i, u(t_i), h) - \frac{u(t_{i+1}) - u(t_i)}{h_i} \right] \end{aligned}$$

Thanks to the previous result (∗) it holds that

$$\left| \tilde{\psi}(t_i, \tilde{u}_i, h) - \psi(t, u(t_i), h) \right| \leq M \left| \tilde{u}_i - u(t_i) \right| = M \left| \tilde{e}_i \right|$$

and consistency yields

$$\left| \tilde{\psi}(t_i, u(t_i), h) - \frac{u(t_{i+1}) - u(t_i)}{h_i} \right| = \delta_{h_i}(t_i, u(t_i)) \ .$$

Hence we obtain the following recursion for the error

$$|\tilde{e}_{i+1}| \leq (1 + h_i M) \, |\tilde{e}_i| + h_i \delta_{h_i}(t_i, u(t_i))$$

as well as $\tilde{e}_0 = 0$. The discrete Gronwall lemma 4.4 shows that

$$\begin{aligned} |\tilde{e}|_i &\leq \sum_{j=0}^{i-1} h_j \cdot \delta_{h_j}(t_j, u(t_j)) \cdot \exp\left[M \sum h_j \right] \\ &\leq |t_i - t_0| \cdot \max_{j=0 \ldots i-1} \delta_{h_j}(t_j, u(t_j)) \cdot e^{M(t_i - t_0)} \\ &\leq |T - t_0| \cdot \max_{j=0 \ldots i-1} \delta_{h_j}(t_j, u(t_j)) \cdot e^{M(T - t_0)} \ . \end{aligned}$$

(3) Since we assume the method to be consistent, i.e. $\max \delta_{h_j} \longrightarrow 0$ for $h \to 0$, there exists h_1, $0 < h_1 \leq h_0$, such that $|\tilde{e}_i| \leq \gamma$ for all i and all step sizes $|h| < h_1$.
Now we can choose a grid $\mathcal{T}_h$ with step size $|h| < h_1$ and consider the

numerical solution on that grid. Then we can skip the extension of the increment, i.e. $\psi = \tilde{\psi}$ and we finally obtain

$$e_h \leq |T - t_0| \cdot \max_{t \in [t_0, T]} \delta_h(t, u) \cdot e^{M|T-t_0|} \ .$$

If the method is just consistent, i.e. $\delta_h \to 0$, or if we even have consistency of order p, i.e. $\delta_h = \mathcal{O}(h^p)$, we get

$$e_h \longrightarrow 0 \quad \text{or} \quad e_h = \mathcal{O}(h^p) \ .$$

□

Since we know that EULER's method is consistent of order $p = 1$ we can even prove that it is convergent of order $p = 1$. Is this the end of the story about numerical methods for ordinary differential equations? Or why should one strive for higher order methods?

Let us consider the influence of numerical round–off errors:
Consider an equidistant grid $\mathcal{T}_h$ and a method with increment ψ. Let u denote the numerical solution *without* round–off errors. Besides that, let U denote the numerical solution *including* round–off errors. Then it holds, that

$$U_0 = u(t_0) + \rho_o \quad \text{and} \quad U_{j+1} = U_j + h\psi(t_j, U_j, h) + \rho_j \ ,$$

where ρ_j denotes the round–off error in the j–th step. The global error $E_j = U_j - u(t_j)$ satisfies the following recursive estimate which can be obtained analogously to the above proof

$$\begin{aligned} |E_{j+1}| &\leq |E_j| + h\left(M\,|E_j| + \delta_h\right) + \rho_j \\ &\leq |E_j| + h\left(M\,|E_j| + \max \delta_h\right) + \rho \end{aligned}$$

where $\rho = \max_j \rho_j$. The discrete GRONWALL lemma now shows, that

$$\begin{aligned} |E_j| &\leq \left[\rho_0 + \sum_{i=0}^{j-1} (h \cdot \max \delta_h + \rho)\right] e^{M(T-t_0)} \\ &\leq \left[\rho + (T - t_0) \max \delta_h + \frac{T - t_0}{h}\rho\right] e^{M(T-t_0)} \\ &\simeq \mathcal{O}\left(\delta + \frac{\rho}{h}\right) \ . \end{aligned}$$

Hence there exists an *optimal step size* h_{opt} for which the influence of the local discretization error balances with the round–off errors. For smaller step sizes $h < h_{\text{opt}}$ the local discretization error gets smaller, but the round–off errors start to accumulate, since we have to do many steps. For larger step sizes $h > h_{\text{opt}}$, the round–off error is less important, but the local discretization error gets dominant. If the method is consistent of order p, the the overall error will be of order $h^p + \frac{\rho}{h}$. Minimizing this overall error, we obtain the optimal step size to be of order

$$h_{\text{opt}} \simeq \sqrt[p+1]{\rho/p}$$

The larger the order p of the method, the larger the optimal step size will be. Hence we aim to construct higher order methods.

4.2 Runge–Kutta Methods

We wish to construct a *second order consistent* method. We start from the integral formulation

$$u(t+h) = u(t) + \int_t^{t+h} f(s, u(s))\, ds$$

and try to construct the increment ψ such that

$$u(t+h) = u(t) + h\psi(t, u, h) + \mathcal{O}(h^3)\ .$$

To approximate the integral up to second order, we use the midpoint rule

$$\int_t^{t+h} f(s, u(s))\, ds = h\, f\left(t + \tfrac{h}{2}, u(t + \tfrac{h}{2})\right) + \mathcal{O}(h^3)\ .$$

Replacing the unknown function value $u(t + \frac{h}{2})$ again by its integral form, we obtain

$$\int_t^{t+h} f(s, u(s))\, ds = h\, f\left(t + \tfrac{h}{2},\ u(t) + \int_t^{t+h/2} f(s, u(s))\, ds\right) + \mathcal{O}(h^3)\,.$$

Now, we have to approximate the *inner* integral maintaining the overall approximation order. It is sufficient to approximate the inner integral up to *first order*, i.e. the rectangle rule is sufficient. This yields

$$\begin{aligned}\int_t^{t+h} f(s, u(s))\, ds &= hf\left(t + \tfrac{h}{2},\ u(t) + \tfrac{h}{2} f(t, u(t)) + \mathcal{O}(h^2)\right) + \mathcal{O}(h^3)\\ &= hf\left(t + \tfrac{h}{2},\ u(t) + \tfrac{h}{2} f(t, u(t))\right) + \mathcal{O}(h^3)\ .\end{aligned}$$

This is the Runge method (1895), which can be written as a *two–stage method*

$$\begin{array}{rll} k_1(t,u,h) = f(t,u) & \text{stage 1} & \\ k_2(t,u,h) = f\left(t+\frac{h}{2}, u+\frac{h}{2}k_1\right) & \text{stage 2} & \\ \psi(t,u,h) = k_2 & \text{update} & u_{j+1} = u_j + h\,(0\cdot k_1 + 1\cdot k_2) \end{array}$$

The Runge method is an example of an *explicit two–stage* Runge–Kutta *method*.

As a generalization, we consider an s–stage method

$$\psi = \sum_{i=1}^{s} b_i k_i$$

where the *stage functions* k_i, $i = 1,\dots,s$ are iteratively defined as

$$k_i = f\left(t + c_i h,\, u + h\sum_{j=1}^{i-1} a_{ij}k_j\right), \quad i = 1\dots s\,.$$

The coefficients $c = (c_1,\dots,c_s) \in \mathbb{R}^s$, $b = (b_1,\dots,b_s) \in \mathbb{R}^s$ and the matrix $A = (a_{ij}) \in \mathbb{R}^{s\times s}$ are collected in the Butcher–Array (b, c, A)

$$\begin{array}{c|c} c & A \\ \hline & b \end{array}$$

Due to the construction, the matrix A is a *strict lower triangular* matrix, i.e. $a_{ij} = 0$ for $j \geq i$.

Example 4.4 (Euler–method as a *one–stage* Runge–Kutta method)**.**

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

Number of stages $s = 1$.
Order of consistency $p = 1$.

Example 4.5 (Runge–method; improved Euler)**.**

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

Number of stages $s = 2$.
Order of consistency $p = 2$.

Python Example 4.2 (2–stage Runge–method).

```python
def runge(f, t, u0):
    # 2-stage Runge method
    # f  : rhs function: R x R^m -> R^m
    # t  : vector of discrete time points
    # u0 : R^m vector of initial values
    # returns
    # U  : U[:,j] solution at time point t[j]

    n = len(t)
    m = len(u0)
    U = zeros((m,n))
    U[:,0] = u0

    for j in range(n-1):
        h  = t[j+1]-t[j]
        k1 = f(t[j], U[:,j])
        k2 = f(t[j]+h/2, U[:,j]+h/2*k1)
        U[:,j+1] = U[:,j] + h*k2

    return U
```

Example 4.6. Two methods of consistency order $p = 3$:

Heun method with $s = 3$ stages

$$\begin{array}{c|ccc} 0 & 0 & & \\ 1/3 & 1/3 & 0 & \\ 2/3 & 0 & 2/3 & 0 \\ \hline & 1/4 & 0 & 3/4 \end{array}$$

Runge–method with $s = 4$ stages

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

Example 4.7 (Classical Runge–Kutta method).

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

Number of stages $s = 4$.
Order of consistency $p = 4$.

Besides these methods, there exist may more, see e.g. Dormand–Prince.

How to construct these Runge–Kutta methods systematically?
Here are some ideas and remarks.

Lemma 4.6 (Consistency). *The* Runge–Kutta *method* (b, c, A) *is consistent, if and only if*

$$\sum_{i=1}^{s} b_i = 1 \ .$$

Idea of the proof. Considering the numerical evolution $u(t+h) = u(t) + h\psi$ with increment function $\psi = \sum b_i k_i$, we obtain in the limit $h \to 0$, that for each stage function $k_i \to f$ holds. Hence $\psi = (\sum b_i)\, f = f$, if and only if $\sum b_i = 1$. □

Lemma 4.7 (Stage number s vs. consistency order p). *The consistency order p of an s–stage* Runge–Kutta *method satisfies*

$$p \leq s \ .$$

Idea of the proof. To see this, we consider the IVP $u' = u$, $u(0) = 1$. The solution after one step of size h satisfies

$$\Phi^{h,0} 1 = e^h = 1 + h + \frac{h^2}{2} + \cdots + \frac{h^p}{p!} + \mathcal{O}(h^{p+1}) \ .$$

The ith–stage k_i of the RK–method is a polynomial of degree $\deg k_i \leq i - 1$ in h. Hence the increment function ψ is also a polynomial of $\deg \psi \leq s$ in h. If ψ is consistent of order p in the above IVP, the ψ approximates the exponential up to an error $\mathcal{O}(h^p)$; i.e. $p \leq s$. □

The following table due to Butcher (1963,1965,1985) contains the minimal number of stages s necessary for achieving consistency order p

p	1	2	3	4	5	6	7	8	≥ 9
$s_{\min}$	1	2	3	4	6	7	9	11	$\geq p+3$

Remark 4.2 (Autonomization). Given a non–autonomous IVP $u' = f(t, u)$ and $u(t_0) = u_0$ for $u \in \mathbb{R}^d$. Introducing the extended variable $y = (t, u) \in \mathbb{R}^{d+1}$, we can re–write the problem as an autonomous one

$$y' = g(y) := \begin{pmatrix} 1 \\ f(t, u) \end{pmatrix}, \quad y(0) = y_0 := \begin{pmatrix} t_0 \\ u_0 \end{pmatrix} \ .$$

This trick is called *autonomization.*

Does the same trick hold for a Runge–Kutta–method?

Consider the stage–functions of a RK–method for the original problem

$$k_i = f\left(t + c_i h, u + h \sum a_{ij} k_j\right)$$

and the numerical evolution after autonomization

$$\begin{pmatrix} t+h \\ u(t+h) \end{pmatrix} = \begin{pmatrix} t \\ u \end{pmatrix} + h\hat{\psi} \;, \quad \text{where} \quad \hat{\psi} = \sum_{i=1}^{s} b_i \begin{pmatrix} \theta_i \\ k_i \end{pmatrix}$$

introducing the extended stage functions

$$\begin{aligned} \hat{k}_i &= f\left(t + \sum_{j=1}^{i-1} a_{ij}\theta_j,\; u + h\sum_{j=1}^{i+1} a_{ij}\hat{k}_j\right) \\ \theta_i &= 1 \;. \end{aligned}$$

Invariance under autonomization requires $k_i = \hat{k}_i$, hence

$$\sum_{j=1}^{i-1} a_{ij} = \sum_{j=1}^{s} a_{ij} \overset{!}{=} c_i \;.$$

Lemma 4.8. *Consider a* Runge–Kutta*–method* (b, c, A).
If $\sum b_i = 1$*, then the method is* consistent*.*
If $\sum_j a_{ij} = c_i$*, then the method is* invariant under autonomization*.*

Example 4.8 (Constructing a two–stage method of second order)**.** We consider a general two–stage Runge–Kutta method with increment

$$\psi = b_1 k_1 + b_2 k_2$$

and the the two stages

$$\begin{aligned} k_1 &= f(t, u) \\ k_2 &= f\left(t + c_2 h,\; u + h a_{21} k_1\right) = f\left(t + c_2 h,\; u + h a_{21} f(t, u)\right) \;. \end{aligned}$$

A Taylor–expansion of k_2 yields

$$k_2 = f + h\left[c_2 \partial_t f + a_{21} f \cdot \partial_u f\right] + h^2\left[\frac{c_2^2}{2}\partial_{tt} f + \frac{a_{21}^2}{2} f^2 \cdot \partial_{uu} f + c_2 a_{21} f \cdot \partial_{tu} f\right] + \mathcal{O}(h^3)$$

and hence for the increment

$$\begin{aligned} \psi = (b_1 + b_2)\, f + h\left[b_2 c_2 \partial_t f + b_2 a_{21} f \cdot \partial_u f\right] \\ + \frac{h^2}{2}\left[b_2 c_2^2 \partial_{tt} f + b_2 a_{21}^2 f^2 \cdot \partial_{uu} f + 2 b_2 c_2 a_{21} f \cdot \partial_{tu} f\right] + \mathcal{O}(h^3) \;. \quad (\dagger) \end{aligned}$$

To compute the consistency error, we still need an expansion of

$$\begin{aligned}\frac{u(t+h)-u(t)}{h} &= f + \frac{h}{2}\left[\partial_t f + f\cdot\partial_u f\right] \\ &+ \frac{h^2}{6}\left[f^2\cdot\partial_{uu}f + 2f\cdot\partial_{tu}f + (\partial_u f)^2\cdot f + f\cdot\partial_u f\cdot\partial_t f + \partial_{tt}f\right] + \mathcal{O}(h^3) \qquad (*)\end{aligned}$$

The consistency error

$$\delta = \frac{u(t+h)-u(t)}{h} - \psi$$

can be obtained by comparing the two expansions (†) and (∗). This yields the conditions

- ▷ To cancel h^0: $b_1 + b_2 = 1$.
- ▷ To cancel h^1: $b_2c_2 = \frac{1}{2}$ and $b_2a_{21} = \frac{1}{2}$.
- ▷ Canceling h^2: not possible!

Invariance under autonomization yields the additional constraint $c_2 = a_{21}$. Hence all RK–methods with $s = 2$ stages and of consistency order $p = 2$ can be written in the form

$$(b_1, b_2, c_2, a_{21}) = \left(\frac{2\lambda-1}{2\lambda}, \frac{1}{2\lambda}, \lambda, \lambda\right) .$$

For $\lambda = 1$ we obtain the method

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

and for $\lambda = 1/2$ we recover the original Runge–method

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

Remark 4.3. The following table (also due to Butcher) contains the number of conditions that have to be satisfied constructing a RK–method of order p.

p	1	2	3	4	5	6	7	8	9	10
# conditions	1	2	4	8	17	37	85	200	486	1205

This illustrates, that constructing high–order RK–methods is a *non–trivial task.*

4.3 Extrapolation Methods

Does there exist an alternative and even simpler approach to construct high–order methods?

The idea of Richardson *extrapolation* provides a principle to construct methods with both *variable* and *arbitrary* order.

Consider a function $d : \mathbb{R}_+ \to \mathbb{R}$. Assume, we can compute values $d(h_n)$ for a zero sequence $h_n \to 0$. Based on these values, we *extrapolate* the limit $\lim\limits_{h\to 0} d(h) =: d(0)$.

Example 4.9 (Richardson–Extrapolation)**.** Consider a smooth function f and we want to compute its derivative $f'(x_0)$ at some point x_0 with high accuracy.

(1) Choose a step size h and consider the centered difference approximation

$$d_f(h) := \frac{f(x_0+h) - f(x_0-h)}{2h} .$$

Using Taylor, we obtain the following expansion of the error

$$e_f(h) := |f'(x_0) - d_f(h)| \le \left|\frac{h^2}{3!} f^{(3)}(x_0)\right| + \left|\frac{h^4}{5!} f^{(5)}(x_0)\right| + \dots$$

The error admits a power series expansion in even powers of h and the original approximation $d_f(h)$ is of *second* order. How to do better?

(2) Consider a step size sequence $h_0 > h_1 > \dots h_k > 0$.

(3) Construct the interpolating polynomial D for the data points $(h_i, d_f(h_i))$.

(4) Evaluate the interpolating polynomial at $h = 0$, i.e. $D(0)$.

Remark 4.4. The last two steps can be done effectively using the Neville–Aitken–scheme

$$\begin{tikzcd}
d(h_0) & := & D_{00} \arrow[dr] & & & \\
d(h_1) & := & D_{11} \arrow[r] & D_{11} & & \\
\vdots & & & & \ddots & \\
d(h_{k-1}) & := & D_{k-1,1} \arrow[r] \arrow[dr] & \cdots \arrow[r] \arrow[dr] & D_{k-1,k-1} \arrow[dr] & \\
d(h_k) & := & D_{k1} \arrow[r] & \cdots \arrow[r] & D_{k,k-1} \arrow[r] & D_{kk} \approx D(0)
\end{tikzcd}$$

Here

$$D_{jl} = D_{j,l-1} + \frac{1}{\mu_{jl}} \left(D_{j,l-1} - D_{j-1,l-1}\right), \quad \mu_{j,l} = \left(\frac{h_{j-l}}{h_j}\right)^r - 1$$

where $r = 2$ is due to the expansion of the error in even powers of h.

Theorem 4.9 (Richardson–Extrapolation)**.** *Consider a function $d : \mathbb{R}_+ \to \mathbb{R}$ and assume the error term $e_f(h) = |d(h) - \lim_{h\to 0} d(h)|$ admits an expansion in powers of h^r*

$$e_f(h) = a_1 h^r + a_2 h^{2r} + \dots a_m h^{mr} + \mathcal{O}(h^{(m+1)r}) \ .$$

Then the coefficients of the Neville–Aitken*–scheme satisfy*

$$e_{jl} := \left| D_{jl} - \lim_{h\to 0} d(h) \right| = b_{jl} h^r_{j-l} \cdots h^r_j = \mathcal{O}(h^{lr}) \ .$$

Proof. See lecture on Numerics. □

Now, we transfer this idea to the numerical solution of ODEs. Our goal is to compute the solution $u(t+h) = \Phi^{t+h,t} u(t)$ after one base step of step size h up to a desired order q.

We choose a *base method* ψ of order p and a sequence $\mathcal{F} = (n_0, n_1, \dots)$ with $n_k \to \infty$, e.g. the Romberg–sequence $n_k = 2^k$ or the harmonic sequence $n_k = k+1$. We define the step sizes $h_k = h/n_k$ and compute $u_{h_k}(t+h)$ using the base method with local step size h_k. Next, we apply the Neville–Aitken–scheme to obtain an extrapolation of the $u(t+h)$ up to the order h^{p+k}.

What is a suitable base–scheme?

The following table compares the number of function evaluation necessary for order q of the extrapolation scheme using a base scheme of order p.

$_p\diagdown^q$	1	2	3	4	5	6	7	8
1	1	2	4	7	11	16	22	29
2		2	5	10	17	26	37	
3			3	8	16	27	41	
4				4	11	22	37	
5					6	17	34	
6						7	20	
7							9	26

This table shows, that the explicit Euler–method with $p = 1$ is sub–optimal, but for sure the most simple choice.

Algorithm 4.1 (Extrapolation–Method)**.**
Input $u(t)$, a base step size h, sequence $\mathcal{F} = (n_0, n_1, \dots)$ and the number K of extrapolation steps.
Output $u(t+h)$ up to order $1 + K$.

For $j = 0 \dots K$ do

$$\begin{aligned} h_j &:= h/n_j \\ u_0^{(j)} &:= u(t) \\ u_{l+1}^{(j)} &:= u_l^{(j)} + h_j f\left(t + lh_j, u_l^{(j)}\right), \quad l = 0 \dots n_j - 1 \\ U_{j0} &:= u_{n_j}^{(j)} \end{aligned}$$

Next, apply Neville–Aitken to extrapolate.
For $l = 1 \dots K$ do

$$U_{jl} := U_{j,l-1} + \frac{U_{j,l-1} - U_{j-1,l-1}}{n_j/n_{j-l} - 1}, \quad \text{for } j = K, \dots l$$

Set $u(t + h) := U_{KK}$.

Remark 4.5. Note, that we have to use $r = 1$ in the Neville–Aitken scheme since the Euler method is first oder consistent, i.e. the error has an expansion in powers of h^1.

How to choose the number K of extrapolation steps?

Assume, we wish to achieve an overall accuracy δ of the numerical solution, i.e. the numerical solution U_{kk} of the above extrapolation method shall approximate the exact solution $u(t + h) = \Phi^{t+h,t}u(t)$ up to this error

$$|u(t + h) - U_{kk}| \leq \delta \; .$$

The following idea is due to Deuflhard and Hohmann [DH03, Chapter 9.5]: We use

$$\varepsilon_{k,k-1} := |U_{k,k-1} - U_{kk}|$$

as an *estimator* for the error

$$e_{k,k-1} := |U_{kk} - u(t + h)|$$

Why? It holds that $e_{jl} \ll e_{j,l-1}$, since U_{jl} is of higher order. Due to

$$\varepsilon_{k,k-1} = |(U_{k,k-1} - u(t + h)) - (u(t + h) - U_{kk})| \leq e_{k,k-1} + e_{kk}$$

and thanks to the inverse triangle inequality $|x - y| \geq ||x| - |y||$

$$\varepsilon_{k,k-1} \geq e_{k,k-1} - e_{kk}$$

we obtain

$$\begin{aligned} e_{k,k-1} - e_{kk} \le \varepsilon_{k,k-1} \le e_{k,k-1} + e_{kk} \\ \Big(1 - \underbrace{\frac{e_{kk}}{e_{k,k-1}}}_{\ll 1}\Big) e_{k,k-1} \le \varepsilon_{k,k-1} \le \Big(1 + \underbrace{\frac{e_{kk}}{e_{k,k-1}}}_{\ll 1}\Big) e_{k,k-1} \end{aligned}$$

and therefore

$$\varepsilon_{k,k-1} \approx e_{k,k-1} \ .$$

If the estimator $\varepsilon_{k,k-1}$ is small enough, i.e. $\varepsilon_{k,k-1} \le \rho \cdot \delta$ for some *safety factor* $\rho \sim {}^1\!/\!_4, {}^1\!/\!_5$, then we *accept* the numerical solution U_{kk} as approximation of the exact solution $u(t+h)$. Otherwise, we add an additional row to the NEVILLE–AITKEN–scheme, i.e. we use an additional extrapolation step to increase the order and re–check the accuracy afterwards.

4.4 Adaptive Step Size Selection

In general we cannot choose the step size of a numerical method *a priori* but we need an *adaptive* choice of the step size. The step size shall be small to allow *accurate* computation, if the solution undergoes large and rapid changes. On the other hand the step size can be large to allow *fast* computation, if the solution does not change much. However, we do not know the solution in advance, hence we cannot compute the needed step sizes in advance; we need adaptivity. To construct an adaptive step size method, we need the following ingredients

(1) An *error estimator* indicating, **where** to change the step size.

(2) A rule **how** to change the step size.

(3) An **efficient** method to do this.

As an *error estimator* ε we use the difference between two numerical solutions $\overline{u}$ und $\hat{u}$ of different order. Let $\overline{u}$ be the approximation using a method $\overline{\psi}$ of order p, i.e.

$$\overline{\delta} = \mathcal{O}(h^p) = c(t_j) \cdot h^p + \mathcal{O}(h^{p+1})$$

and $\hat{u}$ be the approximation using a second method $\hat{\psi}$ of order $p+1$

$$\hat{\delta} = \mathcal{O}(h^{p+1}) \ .$$

Since

$$\Phi^{t+h,t}u(t) = u(t) + h\psi + h\delta = \overline{u} + h\overline{\delta} = \hat{u} + h\hat{\delta}\ ,$$

it holds that

$$\varepsilon = \hat{u} - \overline{u} = h(\overline{\delta} - \hat{\delta}) = ch^{p+1} + \mathcal{O}(h^{p+2}) \approx ch^{p+1}\ .$$

The constant c is *unknown*.

How to determine a suitable the step size? Our goal is to determine an *optimal step size* h^* such that the error ε equals a desired tolerance τ

$$\tau = c(h^*)^{p+1}\ .$$

For a given step size h, the error estimator yields

$$\varepsilon = ch^{p+1}\ .$$

Combining the above two equations, we can get rid of the unknown constant c and obtain for the optimal step size

$$h^* = h\ \sqrt[p+1]{\tau/\varepsilon}\ .$$

To be "on the safe side", we use instead

$$h^* = \min\left(h_{\max},\ qh,\ h\ \sqrt[p+1]{\rho\tau/\varepsilon}\right)\ . \tag{$*$}$$

Here $h_{\max} > 0$ denotes an upper bound for the step size, $q > 1$ is a factor limiting the maximal increase of the step size ($q \sim 5$) and $0 < \rho < 1$ denotes a safety factor ($\rho \sim 1/2, 3/4$).

If, for a given step size h, the error estimator ε is smaller than the tolerance τ, then ($*$) shows how much can we can to *increase* the step size, without violating the accuracy constraint. If $\varepsilon > \tau$, then ($*$) shows how much we have to *decrease* the step size to fulfill the accuracy constraint.

In order to get an *efficient computation*, we must be able to obtain the lower order auxiliary solution $\overline{u}$ with almost no additional effort compared to computing the high–order solution $\hat{u}$. The *embedded* Runge–Kutta *methods* due to Fehlberg and Dormand/Prince are designed for this purpose. They use two Runge–Kutta methods with s and $s+1$ stages that share the same intermediate stages, but differ just in the weighting coefficients b for the increment.

Example 4.10 (Embedded RK–method of order 4 and 3: RK4(3))**.**

$$
\begin{array}{c|ccccc}
0 & 0 & & & & \\
1/2 & 1/2 & 0 & & & \\
1/2 & 0 & 1/2 & 0 & & \\
1 & 0 & 0 & 1 & 0 & \\
1 & 1/6 & 1/3 & 1/3 & 1/6 & 0 \\
\hline
 & 1/6 & 1/3 & 1/3 & 1/6 & 0 \\
 & 1/6 & 1/3 & 1/3 & 0 & 1/6
\end{array}
\qquad
\begin{array}{l}
\left.\vphantom{\begin{array}{c}0\\0\\0\\0\end{array}}\right\} \text{classical RK 4} \\
\\
\rightsquigarrow \text{ order } p = 4 \\
\rightsquigarrow \text{ order } p = 3
\end{array}
$$

The *standard* method for solving ODEs is a $s = 6$–stage RK4(5)–method due to Dormand and Prince; its coefficients can be found in [DB02, Chapter 5.4]. This method is also pre–implemented as the default solver in the function solve_ivp in the python module scipy.integrate. If higher accuracy is required, there exists also a Runge–Kutta method of order 8 by Dormand and Prince. To use this one, we refer to the manual pages of scipy.integrate.solve_ivp.

Chapter 5

Stability

Differential equations are often used to model technological, physical, chemical or biological processes. Parameters appearing in those equations are often obtained by measurements and subject to inaccuracies and errors. It is desirable, that the mathematical model, i.e. the differential equations, react *insensitive* to parameter perturbations. To be more precise, we expect, that a *well–posed* problem shares the following features

existence The problem shall have *at least one* solution.

uniqueness The problem shall have *at most one* solution.

continuous dependence The solution shall *depend continuously* on the data. A small change in the data results in a small change of the solution.

We have touched these issues already in chapter 2. The upper– and lower functions introduced in chapter 2.5 serve as an important tool.

In general we distinguish between the local behavior (e.g. on a compact interval I) of the solution and the long–term behavior on unbounded intervals. First, we will analyze the situation on compact (i.e. bounded) intervals.

5.1 Continuous Dependence on Initial Data

We consider an IVP of the form

$$u' = f(t, u), \qquad u(\xi) = \eta \; , \tag{5.1}$$

on a compact interval $I \subset \mathbb{R}$.

Theorem 5.1. *Let f be defined on $D \subset I \times \mathbb{R}^n$ and assume it to be L–continuous*

$$|f(t,u_1) - f(t,u_2)| \leq L\,|u_1 - u_2| \ , \tag{L}$$

for some $L \geq 0$. Let u be the solution of (5.1) where graph $u \subset D$*. Let z be an* approximate solution *of (5.1) such that*

$$|z(\xi) - u(\xi)| \leq \gamma, \qquad |z' - f(t,z)| \leq \delta \ , \tag{5.2}$$

for some constants γ, $\delta \geq 0$. Then the estimate

$$|u(t) - z(t)| \leq \gamma e^{L|t-\xi|} + \frac{\delta}{L}\left(e^{L|t-\xi|} - 1\right) \tag{5.3}$$

holds for all $t \in I$.

Proof; Sketch. We consider $w := u - z$, where $|w(\xi)| = |u(\xi) - z(\xi)| \leq \gamma$ and

$$\begin{aligned} w' &= u' - z' = f(t,u) - f(t,z) + f(t,z) - z' \\ |w'| &\leq L\,|w| + \delta \\ \rightsquigarrow \quad w(t) &= \leq \gamma e^{L|t-\xi|} + \frac{\delta}{L}\left(e^{L|t-\xi|} - 1\right) \ . \end{aligned}$$

□

The above theorem contains even more information summarized by the following

Theorem 5.2 (Continuous dependence)**.** *Let $u = u(t)$ be the solution of (5.1). For $\alpha > 0$ let*

$$S_\alpha := \{(t,y) : \ t \in I, \ |y - u(t)| \leq \alpha\}$$

denote the α–neighborhood of the solution trajectory. Assume, that there exists some $\alpha_0 > 0$, such that f satisfies the Lipschitz–condition (L) on S_{α_0}. In particular, f is assumed to be continuous on S_{α_0}.
Then the solution u depends continuously on both the initial value η and the right hand side f. In mathematical terms: For all $\varepsilon > 0$, there exists some $\delta > 0$, such that all solutions z of the perturbed *IVP*

$$z' = g(t,z), \qquad z(\xi) = \tilde{\eta} \tag{5.4}$$

where the right hand side g is continuous on S_{α_0} and

$$|g(t,y) - f(t,y)| \leq \delta \quad \text{in } S_{\alpha_0} \ , \qquad |\tilde{\eta} - \eta| \leq \delta \tag{5.5}$$

exist on entire I and the estimate

$$|z(t) - u(t)| \leq \varepsilon \tag{5.6}$$

holds for all $t \in I$.

Proof; Sketch. Let z satisfy (5.4) and (5.5). As long as z is contained in S_{α_0}, the estimate (5.2) holds for $\gamma = \delta$, hence (5.6) also holds setting $\gamma = \delta$. Choosing $\gamma = \delta$ small enough, we can guarantee, that the right hand side of (5.3) is bounded by $\alpha_0/2$. As long as z is inside S_{α_0} the estimate (5.6) holds and therefore $|u_0 - z| \leq \alpha_0/2$. Hence the solution trajectory z cannot leave S_{α_0}.
Let $\varepsilon > 0$. Choosing $\gamma = \delta$ in (5.2) and (5.3) small enough, we can bound the right hand side of (5.3) by ε. □

In applications one frequently encounters the situation, that right hand side f and/or the initial value u_0 of an IVP

$$u' = f(t, u; \lambda), \quad u(t_0) = u_0(\lambda)$$

depend on a parameter $\lambda \in \mathbb{R}^p$. The dependence of the solution $u(t; \lambda)$ of this parameter is called the *sensitivity* of the solution w.r.t. λ.

Definition 5.1 (Sensitivity)**.** Let $I \subset \mathbb{R}$ be a compact interval and $t_0 \in I$. Let $f : I \times \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$ and $u_0 : \mathbb{R}^p \to \mathbb{R}^n$ be twice continuously differentiable. We consider the IVP

$$u' = f(t, u; \lambda), \quad u(t_0) = u_0(\lambda)$$

depending on the parameter $\lambda = (\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^p$ with the unique solution $u(t; \lambda)$. We call

$$S_k(t) := \frac{d}{d\lambda_k} u(t; \lambda), \qquad k = 1, \ldots, p$$

the *sensitivity* of the solution $u(t; \lambda)$ with respect to the parameter λ_k.

How to compute those sensitivities?

Theorem 5.3 (Sensitivity)**.** *The sensitivity $S_k(t)$ satisfies the linear IVP*

$$S_k'(t) = \frac{\partial f}{\partial u} \cdot S_k + \frac{\partial f}{\partial \lambda_k}, \qquad S_k(t_0) = \frac{\partial u_0}{\partial \lambda_k} \,.$$

Proof; Sketch. Differentiating $S_k(t) = \frac{d}{d\lambda_k} u(t; \lambda)$ with respect to t and applying Schwarz theorem to interchange the order of differentiation, we obtain

$$\begin{aligned} S_k'(t) &= \frac{d^2}{dt\, d\lambda_k} u(t; \lambda) = \frac{d}{d\lambda_k} \frac{d}{dt} u(t; \lambda) = \frac{d}{d\lambda_k} f(t, u(t; \lambda); \lambda) = \frac{\partial f}{\partial u} \cdot \frac{du}{d\lambda_k} + \frac{\partial f}{\partial \lambda_k} \\ &= \frac{\partial f}{\partial u} \cdot S_k(t) + \frac{\partial f}{\partial \lambda_k} \end{aligned}$$

together with the initial condition

$$S_k(t_0) = \frac{\partial}{\partial \lambda_k} u(t_0) = \frac{\partial}{\partial \lambda_k} u_0(\lambda) \,.$$

□

Example 5.1. We consider the initial value problem

$$u' = \alpha u =: f(t,u;\alpha), \qquad u(0) = u_0(\beta) = \beta$$

depending on the two–dimensional parameter $\lambda = (\alpha, \beta) \in \mathbb{R}^2$. The solution is given by

$$u(t;\lambda) = \beta e^{\alpha t}$$

and the two sensitivities are easily computed

$$\begin{aligned} S_\alpha(t) &= \frac{d}{d\alpha} u(t;\lambda) = \beta t\, e^{\alpha t} \\ S_\beta(t) &= \frac{d}{d\beta} u(t;\lambda) = e^{\alpha t}\ . \end{aligned}$$

The sensitivities satisfy the IVPs

$$\begin{aligned} S'_\alpha(t) &= \frac{\partial f}{\partial u} \cdot S_\alpha + \frac{\partial f}{\partial \alpha} = \alpha S_\alpha + u = \alpha S_\alpha + \beta e^{\alpha t}, & S_\alpha(0) &= \frac{\partial u_0}{\partial \alpha} = 0\ , \\ S'_\beta(t) &= \frac{\partial f}{\partial u} \cdot S_\beta + \frac{\partial f}{\partial \beta} = \alpha S_\beta + u, & S_\beta(0) &= \frac{\partial u_0}{\partial \beta} = 1\ . \end{aligned}$$

Solving those two IVPs we recover the analytically computed sensitivities.

Example 5.2. We consider the logistic model

$$u' = u(\alpha - u), \qquad u(0) = \beta \tag{5.7a}$$

depending again on two parameters α, β. We expect, that the sensitivity of $u(t;\alpha,\beta)$ for *small* t is large w.r.t. β but small w.r.t. α and for large times t this will be the other way round. The sensitivities themselves satisfy the IVP

$$S'_\alpha = (\alpha - 2u) \cdot S_\alpha + u, \qquad S_\alpha(0) = 0 \tag{5.7b}$$

$$S'_\beta = (\alpha - 2u) \cdot S_\beta \qquad\qquad S_\beta(0) = 1 \tag{5.7c}$$

Now, we can solve the coupled three–dimensional system (5.7) numerically using any method we like to obtain simultaneously the solution u as well as the two sensitivities S_α and S_β. The following Fig. 5.1 shows the solution of the logistic equation together with the two sensitivities. The graph shows, that the sensitivity S_β for the initial vlaue β is large for small times t and then decreases. The parameter α, being the terminal value of the solution u, has a high sensitivity only for large times t.

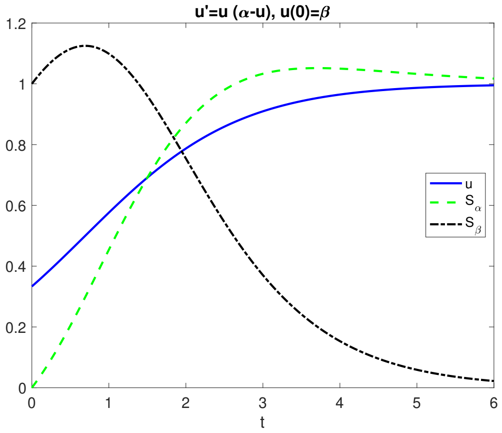


Figure 5.1: Logistic model and sensitivities.

In the next section we we consider stability issues on *unbounded* intervals. Without loss of generality we consider $I = [0, \infty)$.

5.2 Asymptotic Behavior of Solutions

Example 5.3. We consider the linear ODE

$$u' = \lambda u, \qquad u(0) = 1 \ .$$

with the solution $u = e^{\lambda t}$ on entire $\mathbb{R}$.

Perturbing the initial value, i.e.

$$z(0) = 1 + \varepsilon \ ,$$

the difference of the according solutions is given by

$$(z - u)(t) = (1 + \varepsilon)e^{\lambda t} - e^{\lambda t} = \varepsilon e^{\lambda t} \ .$$

- ▷ For $\lambda > 0$ this difference *grows* for arbitrary small $\varepsilon > 0$ exponentially without bounds, $\lim\limits_{t\to\infty}(z - u)(t) = \infty$.
- ▷ For $\lambda = 0$ the difference remains *constant* $= \varepsilon$.
- ▷ For $\lambda < 0$ the difference *decreases* exponentially to zero, $\lim\limits_{t\to\infty}(z - u)(t) = 0$.

Definition 5.2 (Lyapunov Stability)**.** Let u be a solution of $u' = f(t, u)$ on $0 \le t < \infty$. Let f be continuous on $S_\alpha := \{(t, y) : \ 0 \le t < \infty, \ |y - u(t)| < \alpha\}$ for some $\alpha > 0$. The solution u is called

stable (in the sense of Lyapunov), if for all $\varepsilon > 0$ there exists $\delta > 0$, such that all solutions z with $|z(0) - u(0)| < \delta$ exist for all $t \ge 0$ and satisfy

$$|z(t) - u(t)| < \varepsilon \qquad \text{for all } t > 0 \ .$$

asymptotically stable , if it is stable and there exists $\delta > 0$ such that for all solutions z with $|z(0) - u(0)| < \delta$ it holds that

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

unstable , if it is not stable.

First, we consider linear systems

$$u' = A(t)u + b(t) \ , \tag{5.8}$$

where $A(t)$ and $b(t)$ are assumed to be continuous on $I = [0, \infty)$. For any initial value $u(0)$, the solution of (5.8) exists on entire I.

Proposition 5.4. *If the zero–solution of the homogeneous problem $u' = A(t)u$ is stable, asymptotically stable or unstable, then any solution of the inhomogeneous problem (5.8) has the same property.*

A further characterization of stability of arbitrary linear systems is —up to my knowledge— not available analogous to the question of constructing a fundamental system. Hence we restrict to systems with *constant coefficients*

$$u' = Au \ . \tag{5.9}$$

Then the following theorem holds

Theorem 5.5 (Stability of linear systems)**.** *Let $\gamma = \max\{\operatorname{Re}\lambda : \ \lambda \in \sigma(A)\}$ be the largest real part in the spectrum of A. For the trivial solution $u \equiv 0$ of the homogeneous ODE (5.9) it holds that*

(1) If $\gamma < 0$, it is asymptotically stable,

(2) If $\gamma > 0$, it is unstable,

(3) If $\gamma = 0$, it is not asymptotically stable, but stable, if and only if all eigenvalues λ with $\operatorname{Re}\lambda = 0$ have the same algebraic and geometric multiplicity.

Recall: The algebraic multiplicity of an eigenvalue equals to its multiplicity as a root of the characteristic polynomial. The geometric multiplicity equals to the number of linear independent eigenvectors to the eigenvalue.

Proof; Sketch. The cases $\gamma > 0$ and $\gamma < 0$ are easy to see when recalling the fundamental system in its exponential form e^{At}.
Let $\gamma = 0$ and assume there exists an eigenvalue λ whose geometric multiplicity is strictly less then its algebraic multiplicity. Then, the Jordan normalform contains a Jordan block of dimension $m > 1$. This Jordan block corresponds in the fundamental system to an unstable solution of the form

$$e^{\lambda t}\left(1 + t + \frac{1}{2}t^2 + \cdots + \frac{1}{(m-1)!}t^{m-1}\right) \ .$$

If all eigenvalues with $\operatorname{Re}\lambda = 0$, i.e. $\lambda = i\omega$ have equal algebraic and geometric multiplicities, the according Jordan blocks are all of dimension 1 and the corresponding fundamental solutions are given by $e^{\lambda t} = e^{i\omega t} = \cos\omega t + i\sin\omega t$. Those solutions are stable but not asymptotically stable. □

Moreover, the following estimate holds

Theorem 5.6. *Assume the eigenvalues λ of a matrix $A \in \mathbb{R}^{n\times n}$ satisfy*

$$\beta < \operatorname{Re}\lambda < \alpha \,.$$

Then there exists a norm $\|\cdot\| \in \mathbb{C}^n$ such that

$$e^{\beta t}\,\|x\| \le \left\|e^{At}x\right\| \le e^{\alpha t}\,\|x\|$$

for all $t \ge 0$ and all $x \in \mathbb{C}^n$. Hence

$$e^{\beta t} \le \left\|e^{At}\right\| \le e^{\alpha t}\,.$$

Proof. See [Wal98]. □

Example 5.4 (Two–dimensional systems)**.** Consider

$$u' = \begin{pmatrix} x' \\ y' \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix} = Au\,.$$

The matrix A has one of the following *real normal forms*

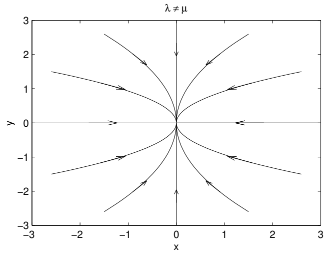
▷ $R(\lambda,\mu) = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}$ with two real eigenvalues.

▷ $R_a(\lambda) = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$ with one eigenvalue with simple geometric multiplicity.

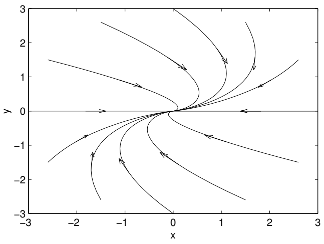
▷ $K(\alpha,\omega) = \begin{pmatrix} \alpha & \omega \\ -\omega & \alpha \end{pmatrix}$ with a pair of conjugate eigenvalues $\lambda = \alpha \pm i\omega$.

The following graphs depict the phase portraits of *asymptotically stable*

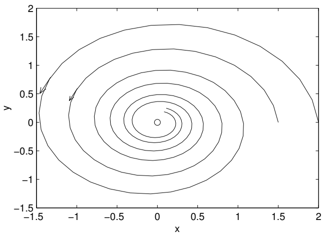
$R(\lambda,\mu)$ for $\lambda,\mu < 0$



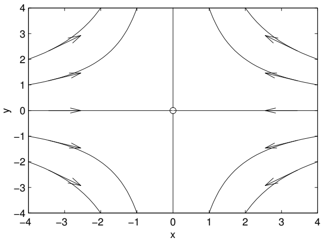
$R_a(\lambda)$ for $\lambda < 0$

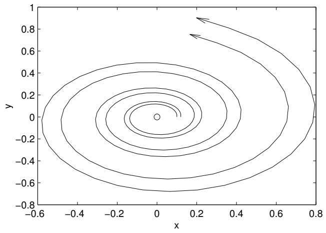


$K(\alpha,\omega)$ for $\alpha < 0$



and *unstable* systems


$R(\lambda, \mu)$ for $\lambda < 0 < \mu$


$K(\alpha, \omega)$ for $\alpha > 0$

In the nonlinear case we need an additional tool; the following

Theorem 5.7 (Gronwall lemma)**.** *Let $I = [0, a[\subset \mathbb{R}$ and $\Phi : I \to \mathbb{R}$ be continuous. Assume, that there exist $\alpha, \beta > 0$ such that*

$$\Phi(t) \leq \alpha + \beta \int_0^t \Phi(\tau)\,\mathrm{d}\tau \tag{5.10}$$

in I. Then

$$\Phi(t) \leq \alpha e^{\beta t}$$

holds for all $t \in I$.

Proof. We define $\psi = \alpha + \beta \int_0^t \Phi(\tau)\,\mathrm{d}\tau$. Then $\psi' = \beta\Phi$ and due to $\Phi \leq \psi$ we get $\psi' \leq \beta\psi$ and $\psi(0) = \alpha$. Let w be the solution of $w' = \beta w$, $w(0) = \alpha$, the w is an upper function to ψ and hence $\Phi \leq \psi \leq w = \alpha e^{\beta t}$. □

The next theorem renders information on the stability of ODEs with *linear principle part*

$$u' = Au + g(t, u)\ . \tag{5.11}$$

Here we assume, that for small u the function $g(t, u)$ is *small* compared to u.

Theorem 5.8 (Stability)**.** *Let $\alpha > 0$. Let $g : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, $(t, z) \mapsto g(t, z)$ be continuous for $t \geq 0$, $|z| \leq \alpha$ and assume*

$$\lim_{|z| \to 0} \frac{|g(t, z)|}{|z|} = 0$$

holds uniformly in t, in particular $g(t, 0) = 0$. Assume, that the matrix $A \in \mathbb{R}^{n \times n}$ is constant and its spectrum is contained in the left half plane, i.e.

$$\operatorname{Re} \lambda < 0 \quad \forall \lambda \in \sigma(A)\ .$$

Then $u \equiv 0$ is an asymptotically stable solution of (5.11).

Proof. There exist $\beta > 0$ und $c < 1$, such that

$$\left|e^{At}\right| \le ce^{-\beta t} \quad \text{for all } t \ge 0 \ . \tag{†}$$

Moreover, there exists $\delta < \alpha$, such that

$$g(t,z) \le \frac{\beta}{2c}\,|z| \quad \text{for all } |z| < \delta \tag{‡}$$

We will show: $|u(0)| \le \varepsilon \le \delta/c$ implies $|u| \le c\varepsilon e^{-\beta t/2}$.

Side remark: The solution of the inhomogeneous equation $u' = Au + b(t)$ is given by $u(t) = e^{At}u_0 + \int_0^t e^{A(t-s)}b(s)\,\mathrm{d}s$. Analogously, the solution of the equation $u' = Au + g(t,u)$ can be written implicitly as

$$u(t) = e^{At}u_0 + \int_0^t e^{A(t-s)}\, g(s,u(s))\,\mathrm{d}s \ .$$

Using (†) and (‡) we obtain

$$|u(t)| \le |u_0|\, ce^{-\beta t} + \int_0^t ce^{-\beta(t-s)}\frac{\beta}{2c}\,|u(s)|\ \mathrm{d}s$$

for $|u| \le \delta$. Let u be the solution of (5.11) for $|u_0| < \varepsilon$ and define $\Phi = |u|\, e^{\beta t}$. Then $|u| \le \delta$ leads to

$$\Phi(t) \le c\varepsilon + \frac{\beta}{2}\int_0^t \Phi(s)\,\mathrm{d}s \ .$$

Applying Gronwall yields $\Phi(t) \le c\varepsilon e^{\beta t/2}$ and hence $|u(t)| \le c\varepsilon e^{-\beta t/2} \le \delta$. □

The above theorem is typically applied when analyzing the stability of autonomous systems

$$u' = f(u);.$$

Definition 5.3 (Stationary point)**.** Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuously diff'able and consider the autonomous system $u' = f(u)$. We call $u^* \in \mathbb{R}^n$ a singular or *stationary point* or *equilibrium* of the system, if $f(u^*) = 0$.

Remark 5.1. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuously diff'able and let u^* be a stationary point of $u' = f(u)$. Then $u(t) \equiv u^*$ is the *unique* solution of the IVP

$$u' = f(u), \quad u(0) = u^* \ .$$

To analyze the stability of u^*, we expand f in its Taylor series around u^* and obtain

$$f(u) = \underbrace{f(u^*)}_{=0} + D_f(u^*)(u-u^*) + g(u-u^*) \ ,$$

where $g(u) = \mathcal{O}(|u - u^*|^2)$. This *linearization* leads to the following ODE with linear principle part

$$(u - u^*)' = D_f(u^*)(u - u^*) + g(u - u^*) \ .$$

Applying Thm. 5.8 leads to

Theorem 5.9 (Nonlinear stability)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuously diff'able and let u^* be a stationary point of $u' = f(u)$. Let $A = D_f(u^*) \in \mathbb{R}^{n \times n}$ denote the Jacobian of f at u^*. Then the stationary solution u^* is*

unstable *, if there exists $\lambda \in \sigma(A)$, such that* $\operatorname{Re} \lambda > 0$.

asymptotically stable *, if for all $\lambda \in \sigma(A)$, it holds that* $\operatorname{Re} \lambda < 0$.

Example 5.5 (Logistic growth)**.** We consider the logistic equation

$$u' = u\,(1 - u) = f(u)$$

with the two equilibria $u_0 = 0$ and $u_1 = 1$. The derivative of the right hand side is given by $D_f(u) = 1 - 2u$. For the stability of these two we get

(1) The trivial equilibrium $u = 0$ is unstable, since $D_f(u_0) = 1 > 0$.

(2) The equilibrium $u_1 = 1$ is asymptotically stable, since $D_f(u_1) = -1 < 0$.

Remark 5.2 (Remark to Thm. 5.8). If $\operatorname{Re} \lambda \leq 0$ for *all* eigenvalues $\lambda \in \sigma(A)$ and $\operatorname{Re} \lambda_k = 0$ for *at least one* eigenvalue $\lambda_k \in \sigma(A)$, then the *stability* of the equilibrium of the *linear* problem $u' = Au$ ***does not*** imply the *stability* of the equilibrium of the *nonlinear* problem $u' = Au + g(t, u)$. This is illustrated by the next

Example 5.6. We analyze the stability of the stationary solution $u \equiv 0$ for the problem

$$u' = \beta u^3 := f(u) \ ,$$

where $\beta \in \mathbb{R}$ is an arbitrary parameter. Linearizing around $u = 0$ leads to the trivial ODE $u' = 0$ with the constant and *stable* solution $u(t) = u_0$. This may lead us to assume, that $u = 0$ is also a stable equilibrium for the nonlinear problem. But: Consider the exact solution of the ODE obtained by separation of variables

$$u(t) = \frac{u_0}{\sqrt{1 - 2u_0^2 \beta t}} \ .$$

This shows, that the analogy to the linearization is not valid. It holds that for

$$\beta \begin{cases} > 0 \ : & \lim_{t\to 1/(2u_0^2\beta)} |u(t)| = +\infty & \textbf{unstable} \\ = 0 \ : & u(t) \equiv u_0 & \textbf{stable} \\ < 0 \ : & \lim_{t\to\infty} u(t) = 0 & \textbf{asymptotically stable}\ . \end{cases}$$

The stability (but not asymptotic stability) of the linearization *does not* imply the stability of the nonlinear problem.

Example 5.7 (Damped mathematical pendulum)**.** We consider the second order ODE

$$u'' + ku' + \sin u = 0\ ,$$

for $k > 0$. Re–writing this as a system of first order

$$\begin{pmatrix} x \\ y \end{pmatrix}' = f(x,y) = \begin{pmatrix} y \\ -ky - \sin x \end{pmatrix}$$

we obtain two stationary points $(0 + 2n\pi, 0)$ and $(\pi + 2n\pi, 0)$ for $n \in \mathbb{Z}$.

(1) At $(0 + 2n\pi, 0)$ the Jacobian is given by

$$D_f(0 + 2n\pi, 0) = \begin{pmatrix} 0 & 1 \\ -1 & -k \end{pmatrix}\ ,$$

the eigenvalues equal to $\lambda_{1,2} = -k/2 \pm \sqrt{k^2/4 - 1}$. Here $\operatorname{Re}\lambda_{1,2} < 0$ and hence the stationary solution $u \equiv 0 + 2n\pi$ is *asymptotically stable.*

(2) At $(\pi + 2n\pi, 0)$ the Jacobian is given by

$$D_f(\pi + 2n\pi, 0) = \begin{pmatrix} 0 & 1 \\ 1 & -k \end{pmatrix}\ ,$$

the eigenvalues equal to $\lambda_{1,2} = -k/2 \pm \sqrt{k^2/4 + 1}$. Here $\operatorname{Re}\lambda_1 > 0$ and $\operatorname{Re}\lambda_2 < 0$. Hence, the stationary solution $u \equiv \pi + 2n\pi$ is *unstable.*

Example 5.8 (Predator–Prey model with Allee effect)**.** We consider the modified predator–prey model

$$\begin{aligned} x' &= x(x - \xi)(1 - x) - \alpha xy \\ y' &= y(x - \eta) \end{aligned}$$

where $0 < \xi, \eta < 1$ and $\alpha > 0$. This model shows four equilibria

$$(x,y) = (0,0), \quad (x,y) = (\xi, 0) \quad (x,y) = (1,0), \quad (x,y) = (\eta, y^*)$$

where $y^* = \frac{1}{\alpha}(x-\xi)(1-x)$. To analyze their stability, we derive the Jacobian

$$D_f(x,y) = \begin{pmatrix} (x-\xi)(1-x) + x(1-x) - x(x-\xi) - \alpha y & -\alpha x \\ y & x - \eta \end{pmatrix}$$

and hence

$$D_f(0,0) = \begin{pmatrix} -\xi & 0 \\ 0 & -\xi \end{pmatrix} \qquad \text{asymptotically stable}$$

$$D_f(\xi,0) = \begin{pmatrix} \xi(1-\xi) & -\alpha\xi \\ 0 & \xi - \eta \end{pmatrix} \qquad \text{unstable, since } \xi(1-\xi) > 0$$

$$D_f(1,0) = \begin{pmatrix} \xi - 1 & -\alpha \\ 0 & 1-\eta \end{pmatrix} \qquad \text{unstable, since } 1-\eta > 0$$

$$D_f(\eta,y^*) = \begin{pmatrix} \eta(1+\xi-2\eta) & -\alpha\eta \\ y^* & 0 \end{pmatrix}$$

The analysis of the stability of the non–trivial (co–existence) equilibrium (η, y^*) is left as an exercise.

5.3 Lyapunov Functions

Example 5.9 (Damped nonlinear oscillator)**.** We consider the equation

$$u'' + ku' + p(u) = 0, \qquad \text{where} \quad k \geq 0\ ,$$

describing a non–linear oscillator. In case of $k = 0$ we neglect damping (friction). in the sequel, we assume that $p(u) > 0$ for $u > 0$.

The *energy* of the system is given by

$$E(u,u') = \frac{u'^2}{2} + P(u), \qquad \text{where} \quad P(u) = \int_0^{u(t)} p(v)\,\mathrm{d}v \geq 0\ .$$

This energy functional has the following properties

- ▷ $E(u,u') \geq 0$
- ▷ $E(u,u') = 0$ iff $u \equiv 0$

These properties are somehow similar to what we expect from a norm.
But what happens to the energy, when we follow it along a solution trajectory?

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} E(u,u') &= u'\,(u'' + p(u)) = u'\,(-ku') \\ &= -ku'^2 \leq 0 \end{aligned}$$

Hence, in case of

no damping , i.e. $k = 0$ we obtain $\frac{\mathrm{d}}{dt}E(u, u') = 0$ along trajectories. The energy is constant along solution trajectories.

damping , i.e. $k > 0$, then $\frac{\mathrm{d}}{dt}E(u, u') < 0$ along trajectories. Hence, the energy is decreasing along solutions.

Can this energy help to decide, whether a solution is stable or not? To answer this question, we consider a real autonomous system of the form

$$u' = f(u) \;, \tag{5.12}$$

where $D \subset \mathbb{R}^n$ is open, $0 \in D$ and $f \in C(D)$ is assumed to be Lipschitz–continuous. Moreover let $f(0) = 0$. The $u \equiv 0$ is an equilibrium of (5.12).

Definition 5.4. The stationary solution $u \equiv 0$ is called *exponentially stable*, if there exist constants $\beta, \gamma, c > 0$ such that for all initial values $|u(0)| < \beta$ the unique solution of (5.12) satisfies the estimate $|u(t)| \leq ce^{-\gamma t}$ for all times $t \geq 0$.

Remark 5.3. Exponential stability implies asymptotic stability; the converse is i.g. *not* true.

Definition 5.5 (Lyapunov–function)**.** A function $V \in C^1(D)$ is called a *Lyapunov–function* to the ODE (5.12), if

(1) $V(x) \geq 0$ for all $x \in D$,

(2) $V(x) = 0$, if and only if $x = 0$,

(3) the derivative of V in direction of f is non–positive, i.e.

$$\begin{aligned} V' &:= (\mathbf{grad}\, V,\, f) = f_1 \cdot \partial_1 V + \cdots + f_n \cdot \partial_n V \\ &\leq 0 \qquad \text{in} \quad D\,. \end{aligned}$$

Remark 5.4. The energy $E(u, u') = u'^2/2 + P(u)$ is a Lyapunov–function for the nonlinear oscillator $u'' + ku' + p(u) = 0$.

Often, for ODEs describing physical systems, Lyapunov functions can be constructed when looking for energies of the system. However, there is no general rule or recipe —at least to my knowledge— on how to construct a Lyapunov function for a given ODE.

Why are Lyapunov–functions important for stability analysis? Due to the following

Theorem 5.10 (Stability Theorem)**.** *Let $D \subset \mathbb{R}^n$ be open and $0 \in D$, let f be Lipschitz–continuous and $f(0) = 0$ and let V be a Lyapunov–function for $u' = f(u)$. If*

(1) $V' \leq 0$ in D, then $u \equiv 0$ is stable*,*

(2) $V' < 0$ in $D \setminus \{0\}$, then $u \equiv 0$ is asymptotically stable*,*

(3) $V' \leq -\alpha V$ and $V(u) \geq b\,|u|^\beta$ in D for some $\alpha, \beta, b > 0$, then $u \equiv 0$ is exponentially stable*.*

Proof. Ad (1). Let $\varepsilon > 0$ and $\overline{B_\varepsilon(0)} := \{x : \ |x| \leq \varepsilon\} \subset D$. We choose $\gamma > 0$ such that $V(x) \geq \gamma$ for all $|x| = \varepsilon$. Next, we choose δ, $0 < \delta < \varepsilon$, such that $V(x) < \gamma$ for all $|x| < \delta$. Now, let u be the solution of (5.12) with an initial value $|u(0)| < \delta$. We call $\Phi(t) := V(u(t))$. Then $\Phi'(t) \leq 0$ by assumption, hence $\Phi(t) \leq \Phi(0) \leq \gamma$. Since $V(u) \geq \gamma$ for $|u| = \varepsilon$, we can conclude, that $|u(t)| < \varepsilon$ for all times t.

Ad (2). Let $u(t)$ be the solution of (5.12) as in part (1) of the proof. Then, there exists $\lim_{t\to\infty} |u(t)| = \beta \leq \gamma$. We will show, that $\beta = 0$.
Assume the contradiction $\beta \neq 0$. Let $M = \left\{x \in \overline{B_\varepsilon(0)} : \beta \leq V(x) \leq \gamma\right\}$. The set M is a compact (closed and bounded) subset of $\overline{B_\varepsilon(0)} \setminus \{0\}$ and hence the maximum of V' exists on M, i.e. $\max_{x\in M} V'(x) = -\alpha < 0$. The solution trajectory u is contained in M and $\Phi'(t) = \frac{\mathrm{d}}{\mathrm{d}t} V(u(t)) < -\alpha$. However, this is a contradiction to $\beta \leq V(x)$ in M.
Therefore $\lim_{t\to\infty} \Phi(t) = 0$, and hence $\beta = \lim_{t\to\infty} |u(t)| = 0$ as well. Now, let $0 < \varepsilon' < \varepsilon$. Then $\min_{\varepsilon' \leq |x| \leq \varepsilon} V(x) = \delta$ and $|u(t)| < \varepsilon'$ for $\Phi(t) < \delta$.

Ad (3). we have $b\,|u(t)|^\beta \leq V(u(t)) =: \Phi(t)$ and $\Phi'(t) \leq -\alpha\Phi(t)$. Hence we arrive at $\Phi(t) \leq \Phi(0)e^{-\alpha t}$ and

$$|u(t)| \leq \left(\frac{\Phi(0)}{b}\right)^{1/\beta} \left(e^{-\alpha t}\right)^{1/\beta} \leq ce^{-\gamma t} .$$

□

Example 5.10. For the nonlinear oscillator, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} E = -ku'^2(t) \leq 0 .$$

Hence, the zero–solution is *stable*.
In case of $k > 0$ it is even asymptotically stable; this does *not* directly follow from the previous theorem.

Theorem 5.11 (Instability theorem). *Let $V \in C^1(D)$ with $V(0) = 0$ and $V(x_k) > 0$ for some zero–sequence $(x_k)_{k\in\mathbb{N}} \subset D$. The zero–solution $u \equiv 0$ is* unstable*, if one of the following two conditions is satisfied*

(1) $V' > 0$ *for* $x \neq 0$ *or*

(2) $V' \geq \lambda V$ *in* D *for some* $\lambda > 0$.

Proof. Let $u(t)$ by the solution of (5.12) for $u(0) = x_k$, hence $\Phi(0) := V(u(0)) = \alpha > 0$.

Ad (1). Let $\varepsilon > 0$ such that $V < \alpha$ in $\overline{B_\varepsilon(0)}$. Since $\Phi' \geq 0$ we obtain $\alpha = \Phi(0) \leq \Phi(t)$ and therefore $|u(t)| > \varepsilon$. Let $r > \varepsilon$, such that $\overline{B_r(0)} \subset D$. For $\varepsilon \leq |x| < r$ we have that $V'(x) \geq \beta > 0$, hence $\Phi' \geq \beta$ and $\Phi \geq \alpha + \beta t$ as long as $u(t) \in B_r(0)$. Since V is bounded in B_r, the solution trajectory $u(t)$ has to leave the ball B_r in finite time.

Ad (2). Let $\Phi(t) := V(u(t))$. Due to assumption $\Phi' \geq \lambda\Phi$, and therefore $\Phi(t) \geq \alpha e^{\lambda t}$. Hence we obtaind $|u(t)| > r$ for large times t similar to the proof of part (1). Since x_k converges to zero, there exist solutions u for arbitrary small initial values x_k, such that $|u(t)| > r$ in finite time t. □

Corollary 5.12. *In particular, the zero–solution is unstable, if*

$$V(x) > 0 \quad \textit{for } x \neq 0 \qquad \textit{and} \qquad V' > 0 \quad \textit{for } x \neq 0\,.$$

This short introduction the Lyapunov–functions is just the starting point for a deeper look into stability theory.

Chapter 6

Partial Differential Equations

Let us recall Definition 1.2: Given an (open) interval $I \subset \mathbb{R}$ and a function $f : I \times \mathbb{R} \times \dots \times \mathbb{R} \to \mathbb{R}$, we call

$$f\left(t,\, u(t),\, \frac{du}{dt}, \dots, \frac{d^k u}{dt^k}\right) = 0 \tag{$*$}$$

a k–th order ordinary differential equation (ODE). If $u \in C^k(I;\mathbb{R})$ and ($*$) holds for all $t \in I$, we call u a *classical solution* of ($*$).

6.1 Definition of PDEs

For the forthcoming, we introduce the following notation.

Definition 6.1 (Multi–index)**.** Let $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$ be a *multi–index* of length n. We define $|\alpha| := \alpha_1 + \dots + \alpha_n$ and

$$D^\alpha u := \frac{\partial^{|\alpha|} u}{\partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n}} = \partial_{x_1}^{\alpha_1} \dots \partial_{x_n}^{\alpha_n} u$$

for a function $u \in \mathcal{C}^{|\alpha|}(\mathbb{R}^n)$.

If the unknown function u is multivariate, i.e. $u : \Omega \subset \mathbb{R}^n \to \mathbb{R}$ depends on more than one variable, we can generalize the notion of an (ordinary) differential equation to

Definition 6.2 (Partial Differential Equation, PDE)**.** Given an open domain $\Omega \subset \mathbb{R}^n$ and a function $F : \Omega \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^{n^2} \times \dots \times \mathbb{R}^{n^k} \to \mathbb{R}$, we call

$$F\left(y,\, u(y),\, Du(y),\, D^2u(y), \dots, D^k u(y)\right) = 0 \tag{$\dagger$}$$

a k–th order PDE, where $u : \Omega \to \mathbb{R}$ is the unknown and $D^k u = \{D^\alpha u : \ |\alpha| = k\}$ is the vector, matrix, . . . of the k–th partial derivatives of u. If the function $u \in C^k(\Omega; \mathbb{R})$, i.e. k–times continuously diff'able, satisfies (†) for all $y \in \Omega$ identically, we call u a *classical solution* to (†).

Example 6.1 (Linear Advection Equation)**.** Let $a \in \mathbb{R}$ and $u : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ satisfy the *linear advection* or *transport equation*

$$u_t + au_x = 0 \ .$$

Here, $u_t = \partial_t u = \dfrac{\partial u}{\partial t}$ and $u_x = \partial_x u = \dfrac{\partial u}{\partial x}$ denote the partial derivative of u with respect to time t and space x. The linear advection equation is a first order PDE. It is easy to check, that for a given initial profile

$$u(t = 0, x) = u_0(x)$$

the solution of the linear advection equation is given by

$$u(t, x) = u_0(x - at) \ .$$

The constant a is also called the *advection velocity.*

Definition 6.3 (Scalar conservation law)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be continuously differentiable. We call the equation

$$u_t + f(u)_x = 0$$

a *scalar conservation law* with *flux* f. Its solution is a function $u : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}$ also called *density.* To solve the scalar conservation law, we need to prescribe an initial density

$$u(0, x) = u_0(x) \ .$$

For a scalar conservation law $u_t + f(u)_x = 0$ with density u and flux f, we call

$$M_{[a,b]}(t) := \int_a^b u(t, x) \, \mathrm{d}x$$

the *mass* inside the interval or domain $[a, b]$. The mass inside an interval changes in time according to the equation

$$\begin{aligned} \frac{d}{dt} M_{[a,b]} &= \frac{d}{dt} \int_a^b u(t, x) \, \mathrm{d}x = \int_a^b \frac{\partial u}{\partial t} \, \mathrm{d}x = -\int_a^b \frac{\partial f(u)}{\partial_x} \\ &= f(u(t, a)) - f(u(t, b)) \ . \end{aligned}$$

Here $f(u(t, a))$ denotes the *influx* into the interval $[a, b]$ at $x = a$ and $-f(u(t, b))$ denotes the *outflux* out of $[a, b]$ through the boundary at $x = b$.

Lemma 6.1. *Let $u \in C(\mathbb{R}_+ \times \mathbb{R})$ be a solution of the scalar conservation law $u_t + f(u)_x = 0$ and assume, that* $\operatorname{supp} u(t,\cdot) = \overline{\{(t,x) : u(t,x) \neq 0\}}$ *is compact for all $t > 0$. Then the* mass of u

$$M(t) := \int_{\mathbb{R}} u(t,x)\,\mathrm{d}x$$

is constant in t, i.e. $M(t) = M(0)$.

Proof.

$$\frac{d}{dt}M(t) = \int_{\mathbb{R}} \frac{\partial u}{\partial t}\,\mathrm{d}x = -\int_{\mathbb{R}} \frac{\partial f(u)}{\partial x}\,\mathrm{d}x = \lim_{R\to\infty} f(u(t,-R)) - f(u(t,R)) = 0\ .$$

□

Example 6.2 (Traffic on a highway)**.** Let u be the density of vehicles on a highway. Then the flux f is given by $f = u \cdot v$, where v is the velocity of the vehicles.

If all vehicles travel, independent of the density, with the same velocity a, then $f = au$ and we recover the linear advection equation

$$\partial_t u + a\partial_x u = 0\ .$$

More realistic is the following model: At low density vehicles travel faster than at high density, i.e. $v = v_{\max} \cdot \dfrac{u_{\max} - u}{u_{\max}}$. Then the flux is given by

$$f = f(u) = \beta u \cdot (u_{\max} - u)$$

and we end up with a *nonlinear, first order PDE*

$$\partial_t u + \beta \cdot \partial_x \left(u \left(u_{\max} - u\right)\right) = 0\ .$$

If we scale the car density u with the maximal density $u_{\max}$, i.e. $u = u_{\max}\tilde{u}$, the space coordinate $x = L\tilde{x}$ and time $t = T\tilde{t}$, we obtain

$$\partial_{\tilde{t}}\tilde{u} + \frac{v_{\max}T}{L}\partial_{\tilde{x}}\left(\tilde{u}(1-\tilde{u})\right) = 0\ .$$

Choosing the time scale $T = L/v_{\max}$ leads to the dimensionless equation (after dropping the tilde)

$$\partial_t u + \partial_x \left[u(1-u)\right] = 0\ .$$

Definition 6.4 (Quasi–linear PDE)**.** A first order PDE $F(y, u, \mathbf{grad}\, u) = 0$ is called *quasi–linear*, if F is linear in $\mathbf{grad}\, u$, but not necessarily in y or u, i.e.

$$F(y, u, \mathbf{grad}\, u) = \langle a(y, u)\, ,\, \mathbf{grad}\, u\rangle - b(y, u)$$

and the corresponding PDE reads as

$$\langle a(y, u)\, ,\, \mathbf{grad}\, u\rangle - b(y, u) = 0\ .$$

Example 6.3. The advection equation $u_t + a u_x = 0$ is quasi–linear. We introduce $y = (t, x)$ and $a(y, u) = (1, a)$, $b = 0$. Then

$$\langle a(y, u)\, ,\, \mathbf{grad}\, u\rangle - b = 1 \cdot u_t + a \cdot u_x = 0\ .$$

The traffic flow equation $\partial_t u + \partial_x u(1 - u) = 0$ is quasi–linear. We introduce $y = (t, x)$ and $a(y, u) = (1, 1 - 2u)$, $b = 0$. Then

$$\langle a(y, u)\, ,\, \mathbf{grad}\, u\rangle - b = 1 \cdot u_t + (1 - 2u) \cdot u_x = 0\ .$$

The PDEs presented so far are of first order. In applications also second order equations play an important role. Here are three classical examples.

Definition 6.5 (Elliptic PDE; Laplace/Poisson equation)**.** Let $\Omega \subset \mathbb{R}^n$ be a domain (open and connected set) and let $f : \Omega \to \mathbb{R}$ be continuous. We call the PDE

$$-\,\Delta\, u = -\left(\partial^2_{x_1} + \cdots + \partial^2_{x_n}\right) u = f$$

Poisson equation/problem. The homogeneous equation

$$-\,\Delta\, u = 0$$

is called *Laplace* equation/problem. To solve this problem, we need to prescribe some *boundary values* for the unknown u at the boundary $\Gamma = \partial\Omega$ of Ω. The following three types of boundary conditions are frequently used

Dirichlet or boundary conditions of the first kind: We prescribe the values of u, i.e.

$$u(y) = g(y) \quad \text{for } y \in \Gamma\ .$$

Neumann or boundary conditions of the second kind: We prescribe the *normal derivative* of u, i.e.

$$\partial_n u(y) = g(y) \quad \text{for } y \in \Gamma\ .$$

The normal derivative $\partial_n u = n \cdot \nabla u = \langle n\, ,\, \nabla u\rangle$ can be interpreted as the *flux* of the quantity u through the boundary.

Robin , mixed or boundary conditions of the third kind: A linear combination of Dirichlet and Neumann, i.e.

$$\beta u + \partial_n u = g \ .$$

Example 6.4 (Poisson equation on the unit circle)**.** Let $\Omega = B_1(0) \subset \mathbb{R}^2$ denote the unit circle and consider the Poisson problem

$$\Delta\, u = 1 \quad \text{in } \Omega$$

supplemented by homogeneous Dirichlet boundary conditions

$$u = 1 \quad \text{on } \partial\Omega.$$

Then, the solution is given by

$$u(x, y) = \frac{1}{4}(r^2 + 3) = \frac{x^2 + y^2 + 3}{4} \, .$$

Definition 6.6 (Parabolic PDE; Diffusion equation)**.** Let $\Omega \subset \mathbb{R}^n$ be a domain with boundary $\Gamma = \partial\Omega$. Let $T > 0$. Let $\alpha > 0$ and $f : [0, T) \times \Omega \to \mathbb{R}$ be continuous. We call the PDE

$$u_t - \alpha\, \Delta\, u = f$$

diffusion (heat) equation. Its solution $u(t, x)$ depends on the *time variable* $t \in [0, T] \subset \mathbb{R}_+$ and the spatial variables $x \in \Omega \subset \mathbb{R}^n$. Note, that the Laplacian is taken with respect to the spatial variables x. To solve the diffusion equation, we have to prescribe both

Initial Condition $u(0, x) = u_0(x)$ for $x \in \Omega$.

Boundary Condition for $t \in [0, T)$ and $x \in \Gamma$. Here we may use Dirichlet, Neumann or Robin type conditions depending on the problem at hand.

The heat equation describes the evolution of a temperature profile $u(t, x)$ over time $t \in [0, T)$. At the starting time $t = 0$, we have given an initial temperature distribution $u_0(x)$ and at the boundary of the domain Ω we have either a fixed temperature $u = g$ (Dirichlet condition), a fixed heat flux $-\alpha\partial_n u = g$ (Neumann condition) or a heat flux, that is proportional to the temperature, e.g. Newton's cooling law $-\alpha\partial_n u = k(u - u_\infty)$. The constant $\alpha > 0$ denotes the *heat conductivity* of the material and $k > 0$ is the heat transfer coefficient through the boundary.

Example 6.5 (Heat equation)**.** Let $\Omega = [0, 2\pi]$ and consider the heat equation

$$u_t - u_{xx} = 0 \quad \text{in } \mathbb{R}_+ \times \Omega$$

with the initial condition

$$u(0, x) = \sin x$$

and the Dirichlet boundary conditions

$$u(t, 0) = 0 = u(t, 2\pi) \ .$$

Then, the solution is given by

$$u(t, x) = e^{-t} \sin x \ .$$

Definition 6.7 (Hyperbolic PDE; Wave Equation)**.** Let $T > 0$ and $f : [0, T) \times \mathbb{R}^n \to \mathbb{R}$ be continuous. We call the PDE

$$u_{tt} - c^2 \, \Delta \, u = f$$

wave equation. To solve the wave equation, we have to prescribe **initial conditions** $u(0, x) = u_0(x)$ and $u_t(0, x) = v_0(x)$ for $x \in \mathbb{R}^n$. If we consider the PDE not on the full space $x \in \mathbb{R}^n$ but just on a domain $\Omega \subset \mathbb{R}^n$, we have to specific additional **boundary conditions** on parts of $\partial\Omega$. For details we refer to the literature.

For the above examples, the solution of the PDEs can be found by educated guesses or try–and–error. For first order PDEs there exists a systematic approach discussed next.

6.2 Method of Characteristics

The scalar conservation law

$$u_t + f(u)_x = 0$$

or its variant called continuity equation with a *source term b*

$$u_t + f(u)_x = b(t, x, u)$$

both supplemented by an initial condition $u(0, x) = u_0(x)$ can be written in quasi–linear from as

$$u_t + a(t, x, u) \cdot u_x = b(t, x, u), \quad u(0, x) = u_0(x) \ .$$

where $a(t, x, u) = f'(u)$. The solution $u = u(t, x)$ can be viewed as a surface over the two–dimensional tx–plane.

Definition 6.8. The *initial value problem (IVP)* for a quasi–linear PDE is given by:
Find a (classical) solution of

$$\partial_t u + a(t,x,u) \cdot u_x = b(t,x,u)$$

where the solution $u(t,x)$ satisfies

$$u(0,x) = u_0(x) \ .$$

To solve a quasi–linear first order PDE

$$\langle A(t,x,u)\, , \, \mathbf{grad}\, u\rangle = \begin{pmatrix} 1 \\ a(t,x,u) \end{pmatrix} \cdot \begin{pmatrix} u_t \\ u_x \end{pmatrix} = b(t,x,u)$$

we use the following idea: The term $\langle A\, , \, \mathbf{grad}\, u\rangle$ denotes the directional derivative of u in direction of the vetor $A = (1,\, a(t,x,u))$. We introduce the parameter s along direction A, and seek the solution $u = u(y)$ along some curves $y = y(s) = (t(s), x(s)) \in \mathbb{R}^2$, the so called *characteristic ground curves.* These characteristic ground curves are determined by the ODE–system

$$\begin{aligned} \frac{d}{ds} t &= 1, & t(0) &= 0 \ , \\ \frac{d}{ds} x &= a(t(s), x(s), u(s)), & x(0) &= x_0 \in \mathbb{R} \ . \end{aligned}$$

Here, the parameter $s \in \mathbb{R}_+$ runs along each of the characteristic ground curves and x_0 denotes the “starting point” of the characteristic. From the first ODE we immediately get $t(s) = s$. The solution $u = u(s, x_0)$ along such a characteristic ground curve is now given by

$$\begin{aligned} \frac{d}{ds} u(s,x_0) = \frac{d}{ds} u(t(s), x(s,x_0)) &= \frac{dt}{ds} \cdot \frac{\partial}{\partial t} u + \frac{dx}{ds} \cdot \frac{\partial}{\partial x} u \\ &= u_t + a(t,x,u) \cdot u_x = b(t,x,u) \end{aligned}$$

i.e. we get a third IVP

$$\frac{d}{ds} u = b(t(s), x(s), u(s)), \quad u(0) = u_0(x_0) \ .$$

Solving the characteristic system consisting of the three IVPs (the first one is trivial in our case)

$$\begin{aligned} \frac{d}{ds} t &= 1, & t(0) &= 0 \ , \\ \frac{d}{ds} x &= a(t(s), x(s), u(s)), & x(0) &= x_0 \in \mathbb{R} \\ \frac{d}{ds} u &= b(t(s), x(s), u(s)), & u(0) &= u_0(x_0) \end{aligned}$$

we obtain the solution $u = u_(s, x_0)$ depending on the two auxiliary parameters (s, x_0) along each of the characteristic curves $(t, x) = (s, x(s, x_0))$ starting at point $(0, x_0)$. If we can invert the implicit equation for the curves, we obtain u as a function of the spatial coordinates (t, x) instead of (s, x_0).

Example 6.6 (Linear Advection)**.** Let's consider the linear advection equation $u_t + cu_x = 0$ with initial condition $u(0, x) = u_0(x)$, where $c \in \mathbb{R}$ is some constant. The characteristic system is given by

$$\begin{aligned} \frac{dt}{ds} &= 1, \quad t(0) = 0 \ , \\ \frac{dx}{ds} &= c, \quad x(0) = x_0 \ , \\ \frac{du}{ds} &= 0, \quad u(0) = u_0(x_0) \ . \end{aligned}$$

Its solution is given by $t = s$, $x(s, x_0) = x_0 + cs = x_0 + ct$ and $u(s) = u_0(x_0)$. Eliminating the curve parameter s and writing x as a function of t, we obtain

$$x = ct + x_0 \quad \text{or} \quad x_0 = x - ct \ .$$

Hence the characteristic ground curves are straight lines $x = x(t)$ and the solution $u(t, x) = u_0(x_0) = u_0(x - ct)$ is constant along those lines.

Example 6.7 (Nonlinear traffic equation)**.** We consider the nonlinear traffic flow model $u_t + (u(1 - x))_x = 0$. Written in quasi–linear form, this equation reads as $u_t + (1 - 2u) \cdot u_x = 0$. Together with the initial condition $u(0, x) = u_0(x)$ the method of characteristics yields the system

$$\begin{aligned} \frac{dt}{ds} &= 1, & t(0) &= 0, & t &= s \ , \\ \frac{du}{ds} &= 0, & u(0) &= u_0 = u_0(x_0), & u(t) &= u_0(x_0) \ , \\ \frac{dx}{ds} &= 1 - 2u, & x(0) &= x_0, & x(t) &= (1 - 2u_0)t + x_0 \ . \end{aligned}$$

We have $x = (1 - 2u_0(x_0)) \cdot t + x_0$; a *nonlinear* equation for x_0 which needs to be solved.

Assume $u_0(x) = \alpha x$. Then $x = (1 - 2\alpha x_0)t + x_0$ and therefore $x_0 = \frac{x-t}{1-2\alpha t}$ for $t \neq 1/(2\alpha)$.
For $t < 1/(2\alpha)$ we obtain

$$u(t, x) = u_0(x_0) = \alpha \frac{x - t}{1 - 2\alpha t} \ .$$

But what happens for $t \geq \frac{1}{2\alpha}$? Does the solution cease to exist?

Let us consider the following nonlinear scalar conservation law with flux $f(u) = u^2/2$

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \quad u(0,x) = u_0(x)\ .$$

This equations is called *Burger's equation.* Using the method of characteristics we obtain $x' = u_0(x_0)$ with initial condition $x(0) = x_0$. The solution reads as $x = u_0(x_0)t + x_0$ which has to be solved for x_0.

Example 6.8 (Shock)**.** Considering the initial condition

$$u_0(x) = \begin{cases} 1: & x < 0 \\ 1 - x: & 0 \leq x \leq 1 \\ 0: & x > 1 \end{cases}$$

we obtain the characteristic ground curves

$$x(t) = \begin{cases} t + x_0: & x_0 < 0 \\ (1 - x_0)t + x_0: & 0 \leq x_0 \leq 1 \\ x_0: & x_0 > 1 \end{cases}$$

If $t < 1$, the characteristic equations are solvable and we obtain the classical solution

$$u(t,x) = \begin{cases} 1: & x < t \\ \frac{1-x}{1-t}: & 0 \leq t \leq x \leq 1 \\ 0: & x > 1 \end{cases}$$

But for $t \geq 1$, the characteristic equations are not solvable anymore; the characteristic ground curves intersect!

What happens? The solution steepens up and finally shows a *jump discontinuity.* This is called a *shock wave.* Does there exist any concept of a solution for $t > 1$?

Example 6.9 (Rarefaction wave)**.** Now, let's consider Burger's equation with a different, increasing initial condition. If u_0 is smooth, the characteristics cover the whole (t,x)–plane; the characteristic system is solvable for all $t \geq 0$ and we obtain a classical solution for all times t. But if the initial solution has a jump, e.g.

$$u_0(x) = \begin{cases} 0: & x < 0 \\ 1: & x \geq 0 \end{cases}$$

we obtain the following ground curves

$$x(t) = \begin{cases} x_0: & x_0 < 0 \\ x_0 + t: & x_0 \geq 0 \end{cases}$$

and the solution reads as

$$u(t,x) = \begin{cases} 0: & x < 0 \\ 1: & x \geq t \end{cases}$$

But what happens in the domain $0 < x < t$, where we have no characteristic ground curves at all? This phenomenon is called a *rarefaction wave.*

But this is just the starting point for the course on *Numerical Methods for Partial Differential Equations*.

Appendix A

Preliminaries

A.1 Eigenvalues and Eigenvectors

In this section we recall some results from Linear Algebra without proving them.

Definition A.1 (Eigenvalue and Eigenvector)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ or resp. $\mathbb{C}^{n\times n}$. We call $\mathbf{0} \neq \boldsymbol{v} \in \mathbb{C}^n$ an *eigenvector* to the *eigenvalue* $\lambda \in \mathbb{C}$, if

$$\mathbf{A}\boldsymbol{v} = \lambda \boldsymbol{v} \,.$$

The set $Eig_\lambda(\mathbf{A}) := \{\boldsymbol{v} : \mathbf{A}\boldsymbol{v} = \lambda\boldsymbol{v}\}$ of all eigenvectors to the eigenvalue λ is called the *eigenspace* to the eigenvalue λ. The eigenspace $Eig_\lambda(\mathbf{A})$ is a vector space. Its dimension $\dim Eig_\lambda(\mathbf{A})$ is called *geometric multiplicity* of the eigenvalue λ.

The set $\sigma(\mathbf{A}) = \{\lambda \in \mathbb{C} : \ \lambda \text{ is an eigenvalue of } \mathbf{A}\}$ consisting of all eigenvalues of $\mathbf{A}$ is called the *spectrum* of $\mathbf{A}$. The largest eigenvalue (by absolute value) defines the *spectral radius* $\rho(\mathbf{A}) := \max\{|\lambda| : \ \lambda \in \sigma(\mathbf{A})\}$.

Theorem A.1. *Let λ be an eigenvalue of $\mathbf{A}$ to the eigenvector $\boldsymbol{v}$. Then*

(1) The eigenvector is not *the zero vector.*

(2) The matrix $\mathbf{A} - \lambda\mathbf{E}$ is singular

$$\det(\mathbf{A} - \lambda\mathbf{E}) = 0 \,,$$

here $\mathbf{E}$ denotes the $n \times n$–identity matrix

(3) The eigenvector $\boldsymbol{v}$ is a non–trivial *solution of the homogeneous linear system*

$$(\mathbf{A} - \lambda\mathbf{E})\boldsymbol{v} = \mathbf{0} \,.$$

(4) The matrix $\mathbf{A} - \lambda\mathbf{E}$ *does* not *have* full *rank.*

(5) The eigenvalue λ *is a root of the* characteristic polynomial

$$\chi_{\mathbf{A}}(x) = \det(\mathbf{A} - x\mathbf{E}) \ .$$

The characteristic polynomial $\chi_{\mathbf{A}}$ of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ or resp. $\mathbb{C}^{n\times n}$ is a polynomial of degree n. Hence there exist n *complex* eigenvalues λ_k, $k = 1 \dots n$ of an $n \times n$–matrix. If the eigenvalue λ_k is a root of $\chi_{\mathbf{A}}$ of multiplicity ν_k, i.e.

$$\chi_{\mathbf{A}}(x) = \prod_{k=1}^{n} (\lambda_k - x)^{\nu_k}$$

then ν_k is called the *algebraic multiplicity* of the eigenvalue λ_k. It holds that $\sum \nu_k = n$.

Let $\lambda \neq \mu$ be two different eigenvalues of $\mathbf{A}$ to eigenvectors $\boldsymbol{v}$ and $\boldsymbol{w}$. Then $\boldsymbol{v}$ and $\boldsymbol{w}$ are linear independent.

Definition A.2 (Diagonalization)**.** A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is called *diagonalizable*, if there exist n linear independent eigenvectors $\boldsymbol{v_1}, \dots, \boldsymbol{v_n} \in \mathbb{C}^n$.

Theorem A.2. *Let* $\mathbf{A}$ *be diagonalizable. Then there exists a diagonal matrix* $\mathbf{D} \in \mathbb{C}^{n\times n}$ *and a regular matrix* $\mathbf{Q} \in \mathbb{C}^{n\times n}$*, such that*

$$\mathbf{A} = \mathbf{Q} \cdot \mathbf{D} \cdot \mathbf{Q}^{-1} \ .$$

Proof. Since $\mathbf{A}$ is diagonalizable, there exist n linear independent eigenvectors $\boldsymbol{v_1}, \dots, \boldsymbol{v_n}$ to the eigenvalues $\lambda_1, \dots, \lambda_n$. Note that $\lambda_j = \lambda_k$ is allowed. We write the eigenvectors as the columns of the matrix $\mathbf{Q}$, i.e.

$$\mathbf{Q} = (\boldsymbol{v_1} \mid \cdots \mid \boldsymbol{v_n}) \in \mathbb{C}^{n\times n} \ .$$

Due to the linear independence of the columns, the matrix $\mathbf{Q}$ is regular, i.e. invertible. The diagonal matrix $\mathbf{D}$ is defined by the eigenvalues

$$\mathbf{D} = \mathbf{diag}\left(\lambda_1, \dots, \lambda_n\right) \ .$$

Then the eigenvalue equations $\mathbf{A}\boldsymbol{v_k} = \lambda_k \boldsymbol{v_k}$ hold for each component, or in matrix form

$$\mathbf{AQ} = \mathbf{QD} \ .$$

Since $\mathbf{Q}$ is invertible we obtain $\mathbf{A} = \mathbf{QDQ}^{-1}$. □

Remark A.1. Not every matrix is diagonalizable. E.g. the matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

has a double eigenvalue $\lambda = 0$ but only (up to linear multiples) one eigenvector $\boldsymbol{v} = \begin{pmatrix} 1 & 0 \end{pmatrix}^T$. Hence the matrix $\mathbf{A}$ is *not diagonalizable.* However, one can show that the so–called JORDAN normalform exists.

Theorem A.3 (JORDAN Normalform)**.** *For every matrix* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *there exist* $\mathbf{Q} \in \mathbb{C}^{n\times n}$*,* $\det \mathbf{Q} \neq 0$*, such that*

$$\mathbf{Q}^{-1}\mathbf{A}\mathbf{Q} = \mathbf{diag}\left(\mathbf{J}_1, \ldots, \mathbf{J}_n\right) \quad \textit{where} \quad \mathbf{J}_k = \begin{pmatrix} \lambda_k & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & \lambda_k \end{pmatrix}.$$

The matrix $\mathbf{J}_k$ *is called* JORDAN*–block to the eigenvalue* λ_k*.*

Theorem A.4 (Further Results)**.** *It holds that*

(1) If $0 \in \sigma(\mathbf{A})$*, then* $\mathbf{A}$ *is* singular*, i.e. not invertible.*

(2) The determinant equals to the product of the eigenvalues

$$\det(\mathbf{A}) = \prod_{\lambda_k \in \sigma(\mathbf{A})} \lambda_k^{\nu_k}$$

(3) The trace equals to the sum of the eigenvalues

$$\operatorname{tr}(\mathbf{A}) = \sum_{\lambda_k \in \sigma(\mathbf{A})} \nu_k \cdot \lambda_k$$

(4) Let $\mathbf{A} = \mathbf{A}^T \in \mathbb{R}^{n\times n}$*. Then* $\sigma(\mathbf{A}) \subset \mathbb{R}$*.*
All eigenvalues of a real symmetric matrix are real.

(5) Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ *or* $\mathbb{C}^{n\times n}$ *be a lower or upper* triangular matrix*, i.e.* $a_{ij} = 0$ *for all* $j < i$ *or* $j > i$*. Then the eigenvalues are the diagonal entries, i.e.* $\sigma(\mathbf{A}) = \{a_{ii} : i = 1 \ldots, n\}$*.*

Definition A.3 (Definiteness)**.** Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric, i.e. $\mathbf{A}^T = \mathbf{A}$. We call $\mathbf{A}$ *positive* or *negative definite*, if $\lambda \geq 0$ or $\lambda \leq 0$ holds for all eigenvalues λ of $\mathbf{A}$.
We call the matrix *strictly positive* or *strictly negative definite*, if $\lambda > 0$ or $\lambda < 0$ holds for all eigenvalues λ.
The matrix is called *indefinite*, if there exist positive and negative eigenvalues.

A.2 Taylor's Theorem

In this section we recall some well–known results from Calculus without proving them.

Definition A.4 (Landau–Symbols)**.** Let f, $g : D \to \mathbb{R}$ be two functions. We call g an *asymptotic upper bound* of f for $x \to x_0$, if there exist $K, \epsilon > 0$ such that $\left|\frac{f(x)}{g(x)}\right| < K$ for all x with $|x - x_0| < \epsilon$.
Analogously, we call g an asymptotic upper bound of f for $x \to \infty$, if there exist $K, R > 0$ such that $\left|\frac{f(x)}{g(x)}\right| < K$ for all x with $x > R$.
If g is an asymptotic upper bound of f, we write $f = \mathcal{O}(g)$ (for $x \to x_0$ resp. $x \to \infty$) and call *"f to be of order big–Oh of g"*.

A function f is called *asymptotically negligible* in relation to g for $x \to x_0$ or $x \to \infty$, if $\lim_{x\to x_0} \frac{f(x)}{g(x)} = 0$ or resp. $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 0$. In this case we write $f = o(g)$ (for $x \to x_0$ or $x \to \infty$) and call "f to be of order small–Oh of g".

Remark A.2. Writing $f = \mathcal{O}(g)$ is just an abbrevition. If $f_1 = \mathcal{O}(g)$ and $f_2 = \mathcal{O}(g)$ holds, this does i.g. *not* imply $f_1 = f_2$. However, if $f = o(g)$ then also $f = \mathcal{O}(g)$.
Mostly, we use statements of the form $f = o(x^k)$ or $\mathcal{O}(x^k)$ for $x \to 0$ or $x \to \infty$, e.g. $\sin x = \mathcal{O}(x)$ for $x \to 0$ and $\ln(x) = o(1/x)$ for $x \to 0$.

Theorem A.5 (Mean Value Theorem (MWT))**.** *Let $f : [a, b] \to \mathbb{R}$ be continuous and differentiable on $]a, b[$. Then there exists $x_0 \in]a, b[$ such that*

$$f'(x_0) = \frac{f(b) - f(a)}{b - a} \ .$$

In a nutshell: At x_0 the tangent is parallel to the secant.

Definition A.5 (Convex, Concave)**.** Let $f : I \to \mathbb{R}$ be twice diff'able. We call the function

convex , if $f''(x) > 0$ for all $x \in I$,

concave , if $f''(x) < 0$ for all $x \in I$.

Theorem A.6 (Taylor)**.** *Let $f : D_f \to \mathbb{R}$ be sufficiently often diff'able in $x_0 \in D_f$ and let $n \in \mathbb{N}$. Then there exists a* unique *polynomial T_n of degree n satisfying*

$$T_n^{(k)}(x_0) = f^{(k)}(x_0) \quad \textit{for } k = 0, 1, 2, \ldots, n \ .$$

This polynomial is called the n–th order Taylor–polynomial *to f at/around x_0. It holds that*

$$T_n(x) = f(x_0) + f'(x_0)\,(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

hence

$$T_n(x) = \sum_{k=0}^{n} \alpha_k (x - x_0)^k \quad \textit{where} \quad \alpha_k = \frac{f^{(k)}(x_0)}{k!} \ .$$

The approximation quality of the TAYLOR–polynomial, i.e. the difference between the function f and its TAYLOR–polynomial can be estimated using the following

Theorem A.7 (LAGRANGE–Remainder)**.** *Let $f : D_f \to \mathbb{R}$ be $(n+1)$–times continuously differentiable and let T_n be the n–th order* TAYLOR*–polynomial at $x_0 \in D_f$. Then*

$$R_n(x) := f(x) - T_n(x)$$

is called remainder*. For all $x \in D_f$ the remainder can be expressed in the* LAGRANGE *form*

$$R_n(x) = \frac{f^{(n+1)}(\tilde{x})}{(n+1)!}(x - x_0)^{n+1}$$

for suitable $\tilde{x} \in [x, x_0]$ or resp. $\tilde{x} \in [x_0, x]$.

Remark A.3. Some remarks to TAYLOR's Theorem.

(1) In the special case of $n = 0$ we obtain the mean value theorem from the Lagrange remainder. The result on the remainder itself can be derived from ROLLE's theorem similarly to the mean value theorem .

(2) Using the LANDAU–symbole, we may write

$$f(x) = T_n(x) + \mathcal{O}\left((x - x_0)^{n+1}\right) \ .$$

(3) TAYLOR–expansion is mostly used for $n = 1$ or $n = 2$.
The case $n = 1$ is also called the *linearization* of f at/around x_0. It holds, that

$$f(x) \approx T_1(x) = f(x_0) + f'(x_0) \cdot (x - x_0) \ .$$

The linear TAYLOR–polynomial (degree $n = 1$) equals to the tangent of f at x_0.
The case $n = 2$ corresponds to the *quadratic approximation* of f at x_0

$$f(x) \approx T_2(x) = f(x_0) + f'(x_0) \cdot (x - x_0) + \frac{f''(x_0)}{2} \cdot (x - x_0) \ .$$

Definition A.6 (Gradient and HESSIAN–matrix)**.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously diff'able.

(1) The vector field

$$\mathbf{grad}\, f(x) := (\partial_{x_1} f(x), \ldots, \partial_{x_n} f(x))$$

is called the *gradient* of f. Usually, the gradient is written as a *row vector.*

(2) The Hessian–matrix $\mathbf{H}_f(x_0) \in \mathbb{R}^{n\times n}$ of f at x_0 is defined as

$$(\mathbf{H}_f(x_0))_{ij} = \partial_{x_i x_j} f(x_0) \ , \quad \text{i.e.} \quad \mathbf{H}_f = \begin{pmatrix} \partial^2_{x_1} f & \partial_{x_1 x_2} f & \ldots & \partial_{x_1 x_n} f \\ \partial_{x_2 x_1} f & \partial^2_{x_2} f & \ldots & \partial_{x_1 x_n} f \\ \vdots & & \ddots & \vdots \\ \partial_{x_n x_1} f & \partial_{x_n x_2} f & \ldots & \partial^2_{x_n} f \end{pmatrix} .$$

Due to Schwarz–theorem, the Hessian is a symmetric matrix.

Theorem A.8 (Taylor–polynomials in $\mathbb{R}^n$)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be three–times cont. diff'able. The linear* Taylor*–polynomial of f at $x^{(0)}$ can be written as*

$$T_1(x) = f(x^{(0)}) + \sum_{i=1}^{n} \partial_{x_i} f(x^{(0)}) \cdot (x_i - x_i^{(0)}) \ .$$

For the second order Taylor*–polynomial it holds that*

$$T_2(x) = T_1(x) + \frac{1}{2} \sum_{i,j=1}^{n} \partial_{x_i x_j} f(x^{(0)}) \cdot (x_i - x_i^{(0)}) \cdot (x_j - x_j^{(0)}) \ .$$

These Taylor*–polynomials yield approximations to $f(x)$ in the neighborhood of x_0*

$$f(x) = T_1(x) + \mathcal{O}(\|x - x_0\|^2) \ ,$$
$$f(x) = T_2(x) + \mathcal{O}(\|x - x_0\|^3) \ .$$

The first–order Taylor*–polynomial T_1 is called* linearization *of f at $x^{(0)}$ and the second–order polynomial T_2 is called* quadratic approximation *of f at $x^{(0)}$.*

Remark A.4. If the function f is sufficiently often differentiable, we can also construct in $\mathbb{R}^n$ Taylor–polynomials of higher orders

$$T_k(x) = T_{k-1}(x) + \frac{1}{k!} \sum_{i_1,\ldots,i_k}^{n} \partial_{x_{i_1} \cdots x_{i_k}} f(x^{(0)}) \cdot (x_{i_1} - x_{i_1}^{(0)}) \cdots (x_{i_k} - x_{i_k}^{(0)}) \ ,$$

here the sum runs over *all* k–th partial derivatives of f.

Theorem A.9 (Remainder in $\mathbb{R}^n$)**.** *Let $f : \mathbb{R}^n \supset \Omega \to \mathbb{R}$ be $(k+1)$–times continuously diff'able and let T_k be the k–th order* Taylor*–polynomial at $x_0 \in \Omega$. Then we call*

$$R_k(x) := f(x) - T_k(x)$$

the remainder*. For all $x \in \Omega$ the remainder can be written in* Lagrange *form*

$$R_k(x) = \sum_{|\alpha|=k+1} \frac{\partial_\alpha f(\tilde{x})}{\alpha!}(x - x_0)^\alpha = \mathcal{O}\left(\|x - x_0\|^{k+1}\right)$$

for some $\tilde{x} \in [x, x_0]$, i.e. $\tilde{x} = \tau x_0 + (1-\tau)(x - x_0)$ where $0 \leq \tau \leq 1$.
The sum runs over all partial derivatives $\partial_\alpha = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_n}^{\alpha_n}$ of order $|\alpha| = \alpha_1 + \cdots + \alpha_n = k+1$. Furthermore, we introduce the notion $\alpha! = \alpha_1 \cdots \alpha_n$ and $x^\alpha = (x_1^{\alpha_1}, \ldots, x_n^{\alpha_n})$.

Remark A.5. In the special case of $f : \mathbb{R}^2 \to \mathbb{R}$ the linear and quadratic Taylor–polynomials around (x_0, y_0) are given by

$$\begin{aligned} T_1(x,y) &= f(x_0,y_0) + \partial_x f(x_0,y_0) \cdot (x - x_0) + \partial_y f(x_0,y_0) \cdot (y - y_0) \\ &= f(x_0) + \langle \mathbf{grad}\, f(x_0)\, ,\, x - x_0 \rangle \end{aligned}$$

and

$$\begin{aligned} T_2(x,y) &= T_1(x,y) + \frac{1}{2}\left[\partial_x^2 f(x_0,y_0) \cdot (x - x_0)^2 \right. \\ &\qquad \left. + 2\,\partial_{xy} f(x_0,y_0) \cdot (x - x_0)(y - y_0) + \partial_y^2 f(x_0,y_0) \cdot (y - y_0)^2\right] \\ &= f(x_0) + \langle \mathbf{grad}\, f(x_0)\, ,\, x - x_0 \rangle + \frac{1}{2}(x - x_0)^T\, \mathbf{H}_f(x_0) \cdot (x - x_0)\ . \end{aligned}$$

The linear Taylor–polynomial T_1 equals to the *tangential plane* to the graph of f at (x_0, y_0).