Numerical Analysis of Ordinary Differential Equations

Guido Kanschat & Robert Scheichl

July 20, 2021

Preface

These notes are a short presentation of the material presented in my lecture. They follow the notes "**Numerik 1:** Numerik gewöhnlicher Differentialgleichungen" by Rannacher (in German) [Ran17b], as well as the books by Hairer, Nørsett, and Wanner [HNW09] and Hairer and Wanner [HW10]. Furthermore, the book by Deuflhard and Bornemann [DB08] was used. Historical remarks are in part taken from the article by Butcher [But96].

We are always thankful for hints and errata.

Thanks go to Dörte Jando, Markus Schubert, Lukas Schubotz, and David Stronczek for their help with writing and editing these notes.

Index for shortcuts

IVP	Initial value problem, s. definition 1.2.7 on page 7
BDF	Backward differencing formula
ODE	Ordinary differential equation
DIRK	Diagonal implicit Runge-Kutta method
ERK	Explicit Runge-Kutta method
IRK	Implicit Runge-Kutta method
LMM	Linear multistep method
VIE	Volterra integral equation, s. Remark 1.2.12 on page 8

Index for symbols

$\mathbb{C}$	The set of complex numbers
e_i	The unit vector of $\mathbb{C}^d$ in direction d
Re	Real part of a complex number
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^d$	The d-dimensional vectorspace of the real d-tuple
u	The exact solution of an ODE or IVP
u_k	The exact solution at time step t_k
y_k	The discrete solution at time step t_k
$\langle x, y \rangle$	The Euclidean scalar product in the space $\mathbb{R}^d$ or $\mathbb{C}^d$
$\lvert x \rvert$	The absolute value of a real number, the modulus of a complex number, or the Euclidean norm in $\mathbb{R}^d$ or $\mathbb{C}^d$, depending on its argument
$\lVert u \rVert$	A norm in a vector space (with exception of the special cases covered by $\lvert x \rvert$)

Contents

Chapter 1

Initial Value Problems and their Properties

1.1 Modeling with ordinary differential equations

Example 1.1.1 (Exponential growth)**.** Bacteria are living on a substrate with ample nutrients. Each bacteria splits into two after a certain time Δt. The time span for splitting is fixed and independent of the individuum. Then, given the amount u_0 of bacteria at time t_0, the amount at $t_1 = t_0 + \Delta t$ is $u_1 = 2u_0$. Generalizing, we obtain

$$u_n = u(t_n) = 2^n u_0, \qquad t_n = t_0 + n\Delta t.$$

After a short time, the number of bacteria will be huge, such that counting is not a good idea anymore. Also, the cell division does not run on a very sharp clock, such that after some time, divisions will not only take place at the discrete times $t_0 + n\Delta t$, but at any time between these as well. Therefore, we apply the continuum hypothesis, that is, u is not a discrete quantity anymore, but a continuous one that can take any real value. In order to accommodate for the continuum in time, we make a change of variables:

$$u(t) = 2^{\frac{t-t_0}{\Delta t}} u_0.$$

Here, we have already written down the solution of the problem, which is hard to generalize. The original description of the problem involved the change of u from one point in time to the next. In the continuum description, this becomes the derivative, which we can now compute from our last formula:

$$\tfrac{d}{dt} u(t) = \frac{\ln 2}{\Delta t} 2^{\frac{t-t_0}{\Delta t}} u_0 = \frac{\ln 2}{\Delta t} u(t).$$

We see that the derivative of u at a certain time depends on u itself at the same time and a constant factor, which we call the growth rate α. Thus, we have arrived at our first differential equation

$$u'(t) = \alpha u(t). \tag{1.1}$$

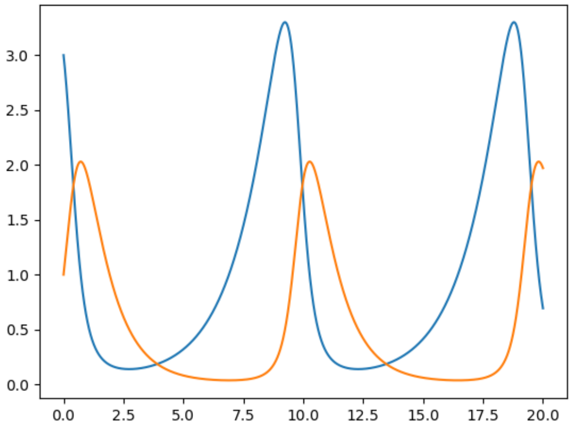


Figure 1.1: Plot of a solution to the predator-prey system with parameters $\alpha = \frac{2}{3}$, $\beta = \frac{4}{3}$, $\delta = \gamma = 1$ and initial values $u(0) = 3$, $v(0) = 1$. Solved with a Runge-Kutta method of order five and step size $h = 10^{-5}$.

What we have seen as well is, that we had to start with some bacteria to get the process going. Indeed, any function of the form

$$u(t) = ce^{\alpha t}$$

is a solution to equation (1.1). It is the initial value u_0, which anchors the solution and makes it unique.

Example 1.1.2 (Predator-prey systems)**.** We add a second species to our bacteria example. Let us replace the bacteria by sardines living in a nutrient rich sea and add sardine-eating tuna. The amount of sardines eaten depends on the likelihood that a sardine and a tuna are in the same place, and on the hunting efficiency β of the tuna. Thus, equation (1.1) is augmented by a negative change in population depending on the product of sardines u and tuna v:

$$u' = \alpha u - \beta uv.$$

In addition, we need an equation for the amount of tuna. In this simple model, we will make two assumptions: first, tuna die of natural causes at a death rate of γ. Second, tuna procreate if there is enough food (sardines), and the procreation rate is proportional to the amount of food. Thus, we obtain

$$v' = \delta uv - \gamma v.$$

Again, we will need initial populations at some point in time to evolve them to later times from that point.

Remark 1.1.3. The predator-prey system (a.k.a. Lotka-Volterra-equations) have periodic solutions. Even though none of these exist in closed form, solutions can be computed numerically (simulated): Volterra became interested in this system when trying to understand why the amount of predatory fish caught in the adriatic sea had increased during World War I, even though during the war years there was a strong decrease of fishing effort. His conclusion, backed by his mathematical model, was that the reason was the abundance of prey, caused by the reduced fishing effort.

A (far too rarely) applied consequence is that in order to diminish the amount of, e.g., foxes one should hunt rabbits, since foxes feed on rabbits.

Example 1.1.4 (Graviational two-body systems)**.** According to Newton's law of universal gravitation, two bodies of masses m_1 and m_2 attract each other with a force

$$\mathbf{F}_1 = G\frac{m_1 m_2}{r^3}\mathbf{r}_1,$$

where $\mathbf{F}_1$ is the force vector acting on m_1 and $\mathbf{r}_1$ is the vector pointing from m_1 to m_2 and $r = |\mathbf{r}_1| = |\mathbf{r}_2|$.

Newton's second law of motion, on the other hand, relates forces and acceleration:

$$\mathbf{F}_i = m_i \mathbf{x}_i'', \quad i = 1, 2,$$

where $\mathbf{x}_i$ is the position of the ith body in space.

Combining these, we obtain equations for the positions of the two bodies:

$$\mathbf{x}_i'' = G\frac{m_{3-i}}{r^3}(\mathbf{x}_i - \mathbf{x}_{3-i}), \qquad i = 1, 2.$$

This is a system of 6 independent variables. However, it can be reduced to three, noting that the distance vector $\mathbf{r}$ is the only variable to be computed for:

$$\mathbf{r}'' = -G\frac{m_1 + m_2}{r^3}\mathbf{r}.$$

Intuitively, it is clear that we need an initial position and an initial velocity for the two bodies. Later on, we will see that this can actually be justified mathematically.

Example 1.1.5 (Celestial mechanics)**.** Now we extend the two-body system to a many-body system. Again, we subtract the center of mass, such that we obtain n sets of 3 equations for an $n+1$-body system. Since forces simply add up, this system becomes

$$\mathbf{x}_i'' = -G\sum_{j\neq i}\frac{m_j}{r_{ij}^3}\mathbf{r}_{ij}. \tag{1.2}$$

Here, $\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i$ and $r_{ij} = |\mathbf{r}_{ij}|$.

Initial data for the solar system can be obtained from

`https://ssd.jpl.nasa.gov/?horizons`

1.2 Introduction to initial value problems

Definition 1.2.1 (Ordinary differential equations): Let $u(t)$ be a function defined on an interval $I \subset \mathbb{R}$ with values in the real or complex numbers or in the space $\mathbb{R}^d$ ($\mathbb{C}^d$). An **ordinary differential equation** (ODE) is an equation for $u(t)$ of the form

$$F\big(t, u(t), u'(t), u''(t), \dots, u^{(n)}(t)\big) = 0. \tag{1.3}$$

Here $F(\dots)$ denotes an arbitrary function of its arguments.

The **order** n of a differential equation is the highest derivative which occurs. If $d > 1$, we talk about **systems of differential equations**.

Remark 1.2.2. A differential equation (DE), which is not ordinary, is called **partial**. These are equations or systems of equations, which involve partial **derivatives with respect to several independent variables**. While the functions in an ordinary differential equation may be dependent on additional parameters, derivatives are only taken with respect to one variable. Often, but not exclusively, this variable is time. This manuscript only deals with ordinary differential equations, and so the adjective will be omitted in the following.

Definition 1.2.3: An **explicit differential equation** of first order is a equation of the form

$$\begin{aligned} u'(t) &= f(t, u(t)) \qquad (1.4)\\ \text{or shorter:} \qquad u' &= f(t, u). \end{aligned}$$

A differential equation of order n is called explicit, if it is of the form

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

Lemma 1.2.4: Every differential equation (of arbitrary order) can be written as a system of first-order differential equations. If the equation is explicit, then the system is explicit.

Proof. We introduce the additional variables $u_0(t) = u(t)$, $u_1(t) = u'(t)$ to $u_{n-1}(t) = u^{(n-1)}(t)$. Then, the differential equation in (1.3) can be reformulated as the system

$$\begin{pmatrix} u_0'(t) - u_1(t) \\ u_1'(t) - u_2(t) \\ \vdots \\ u_{n-2}'(t) - u_{n-1}(t) \\ F\big(t, u_0(t), u_1(t), \dots, u_{n-1}(t), u_{n-1}'(t)\big) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{1.5}$$

In the case of an explicit equation, the system has the form

$$\begin{pmatrix} u_0'(t) \\ u_1'(t) \\ \vdots \\ u_{n-2}'(t) \\ u_{n-1}'(t) \end{pmatrix} = \begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_{n-1}(t) \\ f\big(t, u_0(t), u_1(t), \dots, u_{n-1}(t)\big) \end{pmatrix}. \tag{1.6}$$

□

Example 1.2.5. The differential equation

$$u'' + \omega^2 u = f(t) \tag{1.7}$$

can be transformed into the system

$$\begin{aligned} u_0' - u_1 &= 0, \\ u_1' + \omega^2 u_0 &= f(t). \end{aligned} \tag{1.8}$$

The transformation is not uniquely determined. In this example, a more symmetric system can be obtained:

$$\begin{aligned} u_0' - \omega u_1 &= 0, \\ u_1' + \omega u_0 &= f(t). \end{aligned} \tag{1.9}$$

From a numerical perspective, system (1.9) should be chosen over (1.8) to avoid loss of significance or overflow, i.e. if $|\omega| \ll 1$ or $|\omega| \gg 1$.

Definition 1.2.6: A differential equation of the form (1.4) is called **autonomous**, if the right hand side f is not explicitly dependent on t, i.e.

$$u' = f(u). \tag{1.10}$$

Each differential equation can be transformed into an autonomous differential equation. This is called **autonomization**.

$$U = \begin{pmatrix} u \\ t \end{pmatrix}, \qquad F(U) = \begin{pmatrix} f(t,u) \\ 1 \end{pmatrix}, \qquad U' = F(U)$$

A method that provides the same solution for the autonomous DE as for the original IVP, is called **invariant under autonomization**.

Differential equations usually provide sets of solutions from which we have to choose a solution. An important selection criterion is setting an initial value which leads to a well-posed problem (see below).

Definition 1.2.7: Given a point $(t_0, u_0) \in \mathbb{R} \times \mathbb{R}^d$ and a function $f(t,u)$ with values in $\mathbb{R}^d$, defined in a neighborhood $I \times U \subset \mathbb{R} \times \mathbb{R}^d$ of (t_0, u_0). Then, an **initial value problem** (IVP) is defined as follows: find a function $u(t)$, such that

$$u'(t) = f\big(t, u(t)\big) \tag{1.11a}$$

$$u(t_0) = u_0 \tag{1.11b}$$

Definition 1.2.8: A continuously differentiable function $u(t)$ with $u(t_0) = u_0$ is called a **local solution** of the IVP (1.11), if there exists a neighborhood $J \subset \mathbb{R}$ of t_0, such that $u(t)$ and $f(t, u(t))$ are defined and equation (1.11a) holds for all $t \in J$.

Remark 1.2.9. We introduced the IVP deliberately in a "local" form because the local solution term is the most useful one for our purposes. Due to the fact that the neighborhood J in the definition above can be arbitrarily small, we will have to deal with the extension to larger intervals below.

Remark 1.2.10. Through the substitution of $t \mapsto \tau$ with $\tau = t - t_0$ it is possible to transform every IVP at the point t_0 to an IVP in 0. We will make use of this fact and soon always assume $t_0 = 0$.

Lemma 1.2.11: Let f be continuous in both arguments. Then, the function $u(t)$ is a solution of the initial value problem (1.11) if and only if it is a solution of the **Volterra integral equation** (VIE)

$$u(t) = u_0 + \int_{t_0}^{t} f\big(s, u(s)\big)\,\mathrm{d}s. \tag{1.12}$$

Remark 1.2.12. The formulation as an integral equation allows a more general notion of solution, because the problem is well-posed for functions $f(t, u)$ that are merely integrable with respect to t. (In that case, the solution u would be just absolutely continuous and not continuously differentiable.) Both the theoretical analysis of the IVP and the numerical methods in this lecture notes (with the exception of the BDF methods) are in fact considering the associated integral equation (1.12) and not the IVP (1.11).

Theorem 1.2.13 (Peano's local existence theorem): Let $\alpha, \beta > 0$ and let the function $f(t, x)$ be continuous on the closed set

$$\overline{D} = \big\{(t, x) \in \mathbb{R} \times \mathbb{R}^d \;\big|\; |t - t_0| \le \alpha,\ |x - u_0| \le \beta\big\}. \tag{1.13}$$

There exists a solution $u(t) \in C^1(I)$ of (1.11) on the interval $I = [t_0 - T, t_0 + T]$ with

$$T = \min\left(\alpha, \frac{\beta}{M}\right),\ M = \max_{(t,x)\in\overline{D}} |f(t, x)|.$$

The proof of this theorem is of little consequence for the remainder of these notes. For its verification, we refer to textbooks on the theory of ordinary differential equations or to [Ran17b, Satz 1.1].

Remark 1.2.14. The Peano existence theorem does not make any statements about the uniqueness of a solution and guarantees only local existence. The second limitation is addressed by the following theorem. Uniqueness will be discussed in Section 1.4.

Theorem 1.2.15 (Peano's continuation theorem): Let $f(t,x)$ be continuous on the open set $D \subset \mathbb{R} \times \mathbb{R}^d$. For every $(t_0, u_0) \in D$, there exists an interval $I = (t_-, t_+)$ such that the local solution u from Theorem 1.2.13 can be extended to the left and to the right of t_0 to a continuously differentiable solution $u \in C^1(I)$ of (1.11) as long as the graph of u does not touch the boundary of D. The interval I may be unbounded and the graph $\{(t, u(t)) : t \in I\}$ may also be unbounded for $t \to t_+$ or $t \to t_-$.

Proof. W.l.o.g. we only focus on the extension of the local solution from Theorem 1.2.13 to the right, i.e. to $[t_0, t_+)$. Since D is open, there exist $\alpha_0, \beta_0 > 0$ such that

$$\overline{D}_0 := \left\{(t,x) \in \mathbb{R} \times \mathbb{R}^d : |t - t_0| \leq \alpha_0,\ |x - u_0| \leq \beta_0\right\} \subset D.$$

It follows from Theorem 1.2.13 that there exists a local solution u^0 of (1.11) on $[t_0, t_0 + T_0]$ with

$$T_0 = \min(\alpha_0, \beta_0/M_0) > 0 \quad \text{and} \quad M_0 = \max_{(t,x)\in\overline{D}_0} |f(t,x)|.$$

Now, let $(t_1, u_1) := (t_0 + T_0, u^0(t_0 + T_0)) \in \overline{D}_0 \subset D$. We can apply Theorem 1.2.13 again to show the existence of a local solution u^1 of (1.11) with initial condition $u^1(t_1) = u_1$ on $[t_1, t_2]$, where $t_2 = t_1 + T_1$,

$$T_1 = \min(\alpha_1, \beta_1/M_1) > 0 \quad \text{and} \quad M_1 = \max_{(t,x)\in\overline{D}_1} |f(t,x)|,$$

for $\alpha_1, \beta_1 > 0$ and the cylinder

$$\overline{D}_1 := \left\{(t,x) \in \mathbb{R} \times \mathbb{R}^d : |t - t_1| \leq \alpha_1,\ |x - u_1| \leq \beta_1\right\} \subset D.$$

Due to the continuity of $f(t,x)$ on D, the local solution pieces u^0 on $[t_0, t_1]$ and u^1 on $[t_1, t_2]$ can be combined to a continuously differentiable solution $u \in C^1[t_0.t_2]$ of (1.11).

Obviously, this process can be continued until the graph of the solution approaches the boundary of D. In fact, if the sequence $(t_k, u_k) \in D$ converges to an interior point $(t_*, u(t_*)) \in D$, it is possible to apply Theorem 1.2.13 again to extend the interval beyond t_* until the boundary of D is reached. □

Example 1.2.16. The IVP

$$u' = 2\sqrt{|u|}, \qquad u(0) = 0,$$

has solutions $u(t) = t^2$ and $u(t) = 0$ that both exist for all $t \in \mathbb{R}$ (global existence, but no uniqueness).

Example 1.2.17. The IVP

$$u' = -u^2, \qquad u(0) = 1.$$

has the unique solution $1/(1+t)$. This solution has a singularity for $t \to -1$ (no global existence, but uniqueness). However, it exists for all $t > -1$ and thus in particular for all $t > 0 = t_0$, which is all that matters for an IVP.

1.3 Linear ODEs and Grönwall's inequality

1.3.1. The study of linear differential equations turns out to be particularly simple and results obtained here will provide us with important statements for general non-linear IVPs. Therefore, we pay particular attention to the linear case.

Definition 1.3.2: An IVP according to Definition 1.2.7 is called **linear** if the right hand side f is an affine function of u and the IVP can be written in the form

$$u'(t) = A(t)u(t) + c(t) \qquad \forall t \in \mathbb{R} \tag{1.14a}$$
$$u(t_0) = u_0 \tag{1.14b}$$

with a continuous matrix function $A : \mathbb{R} \to \mathbb{C}^{d\times d}$.
If in addition $c(t) \equiv 0$, we call the IVP **homogeneous**.

Definition 1.3.3: Let the matrix function $A : I \to \mathbb{C}^{d\times d}$ be continuous. Then the function defined by

$$M(t) = \exp\left(-\int_{t_0}^{t} A(s)\,\mathrm{d}s\right) \tag{1.15}$$

is called **integrating factor** of the equation (1.14a). (See Appendix A.2.1 for a definition and properties of the **matrix exponential** in (1.15).)

Corollary 1.3.4. *The integrating factor $M(t)$ has the properties*

$$M(t_0) = \mathbb{I} \tag{1.16}$$
$$M'(t) = -M(t)A(t). \tag{1.17}$$

Lemma 1.3.5: Let $M(t)$ be the integrating factor of the equation (1.14a) defined in (1.15). Then, the function

$$u(t) = M(t)^{-1}\left(u_0 + \int_{t_0}^{t} M(s)c(s)\,\mathrm{d}s\right) \tag{1.18}$$

is a solution of the IVP (1.14) that exists for all $t \in \mathbb{R}$.

Proof. We consider the auxiliary function $w(t) = M(t)u(t)$ with the integrating factor $M(t)$ defined as in eqn. (1.15). It follows by using the product rule that

$$w'(t) = M(t)u'(t) + M'(t)u(t) = M(t)(u'(t) - A(t)u(t)). \tag{1.19}$$

Using the differential equation (1.14a), we obtain

$$w'(t) = M(t)c(t).$$

This can be integrated directly to obtain

$$w(t) = u_0 + \int_{t_0}^{t} M(s)c(s)\,\mathrm{d}s,$$

where we have used (1.16) such that $w(t_0) = M(t_0)u(t_0) = u_0$.

According to Lemma A.2.3 about the matrix exponential, $M(t)$ is invertible for all t. Thus we can apply $M(t)^{-1}$ to $w(t)$ to obtain the solution $u(t)$ of (1.14) as given in equation (1.18).

The global solvability follows since the solution is defined for arbitrary $t \in \mathbb{R}$. □

Example 1.3.6. The equation in Example 1.2.5 is linear and can be written in the form of (1.14) with

$$A(t) = A = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \quad \text{and} \quad c(t) = \begin{pmatrix} 0 \\ f(t) \end{pmatrix}.$$

Let now $f(t) \equiv 0$, $t_0 = 0$ and $u(0) = u_0$. It is easy to see A has eigenvalues $i\omega$ and $-i\omega$, so that we can write

$$A = C^{-1} \begin{pmatrix} \omega i & 0 \\ 0 & -\omega i \end{pmatrix} C$$

with a suitable transformation matrix C $\boxed{\text{DIY}}$. Using the properties of the matrix exponential, the integrating factor is

$$M(t) = e^{-At} = C^{-1} \begin{pmatrix} e^{-i\omega t} & 0 \\ 0 & e^{i\omega t} \end{pmatrix} C = \begin{pmatrix} \cos\omega t & \sin\omega t \\ -\sin\omega t & \cos\omega t \end{pmatrix}.$$

Thus, the solution is

$$u(t) = \begin{pmatrix} \cos\omega t & -\sin\omega t \\ \sin\omega t & \cos\omega t \end{pmatrix} u_0\,.$$

The missing details in this argument and the case for an inhomogeneity $f(t) = \cos\alpha t$ are left as an exercise $\boxed{\text{DIY}}$.

Remark 1.3.7. If the function $c(t)$ in (1.14a) is only integrable, the function $u(t)$ defined in (1.18) is absolutely continuous and thus differentiable almost everywhere. The chain rule (1.19) is applicable in all points of differentiability and $w(t)$ solves the Volterra integral equation corresponding to (1.14). Thus, the representation formula (1.18) holds generally for solutions of linear Volterra integral equations.

Lemma 1.3.8 (Grönwall): Let $w(t)$, $a(t)$ and $b(t)$ be scalar-valued, nonnegative, integrable functions, such that $a(t)w(t)$ is integrable. Furthermore, let $b(t)$ be monotonically non-decreasing and let $w(t)$ satisfy the integral inequality

$$w(t) \le b(t) + \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s, \qquad t \ge t_0. \tag{1.20}$$

Then, for almost all $t \ge t_0$ there holds:

$$w(t) \le b(t) \exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right). \tag{1.21}$$

Proof. Using the integrating factor

$$m(t) = \exp\left(-\int_{t_0}^{t} a(s)\,\mathrm{d}s\right), \qquad \frac{1}{m(t)} = \exp\left(\int_{t_0}^{t} a(s)\,\mathrm{d}s\right),$$

we introduce the auxiliary function

$$v(t) = m(t)\int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s,$$

This function is absolutely continuous, and since $m'(t) = -a(t)m(t)$, we have almost everywhere

$$v'(t) = m(t)a(t)\left[w(t) - \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s\right].$$

Using assumption (1.20), the bracket on the right can be bounded by $b(t)$. Thus,

$$v'(t) \le m(t)a(t)b(t)$$

and since by definition $v(t_0) = 0$, it follows that

$$v(t) \le \int_{t_0}^{t} m(s)a(s)b(s)\,\mathrm{d}s,$$

which implies, using the definition of $v(t)$, that

$$\int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s = \frac{1}{m(t)}v(t) \le \frac{1}{m(t)}\int_{t_0}^{t} m(s)a(s)b(s)\,\mathrm{d}s.$$

Finally, since $b(t)$ is nondecreasing we obtain almost everywhere

$$\begin{aligned}\int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s &\le \frac{b(t)}{m(t)}\int_{t_0}^{t} a(s)\exp\left(-\int_{t_0}^{s} a(r)\,\mathrm{d}r\right)\,\mathrm{d}s \\ &= \frac{b(t)}{m(t)}\Bigg[-\underbrace{\exp\left(-\int_{t_0}^{s} a(r)\,\mathrm{d}r\right)}_{m(s)}\Bigg]_{t_0}^{t} \\ &= \frac{b(t)}{m(t)}\big(m(t_0) - m(t)\big) = \frac{b(t)}{m(t)} - b(t).\end{aligned}$$

Combining this bound with the integral inequality (1.20), we obtain

$$w(t) \le b(t) + \int_{t_0}^{t} a(s)w(s)\,\mathrm{d}s = \frac{b(t)}{m(t)},$$

which proves the lemma. □

Remark 1.3.9. As we can see from the form of assumption (1.20) and estimate (1.21), the purpose of Grönwall's inequality is to construct a majorant for $w(t)$ that satisfies a linear IVP. The bound is particularly simple when $a, b \ge 0$ are constant.

Corollary 1.3.10: If two solutions $u(t)$ and $v(t)$ of the linear differential equation (1.14a) coincide in a point t_0, then they are identical.

Proof. The difference $w(t) = v(t) - u(t)$ solves the integral equation

$$w(t) = \int_{t_0}^{t} A(s)w(s)\,\mathrm{d}s.$$

Hence, for an arbitrary vector norm $\|\cdot\|$ (and induced matrix norm also denoted by $\|\cdot\|$), we can obtain the following integral inequality

$$\|w(t)\| \le \int_{t_0}^{t} \|A(s)w(s)\|\,\mathrm{d}s \le \int_{t_0}^{t} \|A(s)\|\|w(s)\|\,\mathrm{d}s$$

Now, applying Grönwall's inequality (1.21) with $a(t) = \|A(t)\|$ and $b(t) = 0$, we can conclude that $\|w(t)\| = 0$ and therefore $u(t) = v(t)$, for all t. □

Corollary 1.3.11. *The representation formula* (1.18) *in Lemma 1.3.5 defines the unique solution to the IVP* (1.14)*. In particular, solutions of linear IVPs are always defined for all* $t \in \mathbb{R}$.

Example 1.3.12. Let $A \in \mathbb{C}^{d\times d}$ be diagonalizable with (possibly repeated) eigenvalues $\lambda_1, \dots, \lambda_d$ and corresponding eigenvectors $\psi^{(1)}, \dots, \psi^{(d)}$. The linear IVP

$$\begin{aligned} u' &= Au, \\ u(0) &= u_0. \end{aligned}$$

has unique solution $u(t) = e^{At}u_0$. Using the properties of the matrix exponential (see Appendix A.2.1), with $\Psi \in \mathbb{C}^{d\times d}$ denoting the matrix with ith column $\psi^{(i)}$, we get

$$u(t) = e^{\Psi^{-1}\operatorname{diag}(\lambda_1,\dots,\lambda_d)\Psi t}u_0 = \Psi^{-1}\operatorname{diag}\left(e^{\lambda_1 t}, \dots, e^{\lambda_d t}\right)\Psi u_0.$$

Lemma 1.3.13: The solutions of the homogeneous, linear differential equation

$$u'(t) = A(t)u(t) \tag{1.22}$$

with $A : \mathbb{R} \to \mathbb{R}^{d\times d}$ and $u : \mathbb{R} \to \mathbb{R}^d$, define a vector space H of dimension d.

Let $\{\psi^{(i)}\}_{i=1,\dots,d}$ be a basis of $\mathbb{R}^d$. Then the solutions $\varphi^{(i)}(t)$ of the equation (1.22) with initial values $\varphi^{(i)}(0) = \psi^{(i)}$ form a basis of the solution space H. The vectors $\{\varphi^{(i)}(t)\}_{i=1,\dots,d}$ are linear independent for all $t \in \mathbb{R}$.

Proof. At first we observe that, due to linearity of the derivative and the right hand side, for two solutions $u(t)$ and $v(t)$ of the equation (1.22) and for $\alpha \in \mathbb{R}$, $\alpha u(t) + v(t)$ is also a solution of (1.22) with initial condition $\alpha u(0) + v(0) \in \mathbb{R}^d$. Therefore, the vector space structure follows from the vector space structure of $\mathbb{R}^d$.

Let now $\{\varphi^{(i)}(t)\}_{i=1,\dots,d}$ be solutions of the IVP with linearly independent initial values $\{\psi^{(i)}\}_{i=1,\dots,d}$. As a consequence the functions are linearly independent as well.

Assume that $w(t)$ is a solution of equation (1.22) that cannot be written as a linear combination of the functions $\{\varphi^{(i)}(t)\}_{i=1,\dots,d}$. Then, $w(0)$ is not a linear combination of the vectors $\{\psi^{(i)}\}_{i=1,\dots,d}$. Because otherwise, if there exists $\{\alpha_i\}_{i=1,\dots,d}$ with $w(0) = \sum \alpha_i \psi^{(i)}$, then $w(t) = \sum \alpha_i \varphi^{(i)}(t)$ due to the uniqueness of any solution of equation (1.22) proven in Corollary 1.3.10, which would lead to a contradiction. However, since $\{\psi^{(i)}\}_{i=1,\dots,d}$ was assumed to form a basis of $\mathbb{R}^d$, that implies $w(0) = 0$ and thus $w \equiv 0$. It follows that the solution space has dimension d and that $\{\varphi^{(i)}(t)\}_{i=1,\dots,d}$ forms a basis.

Since $t \in \mathbb{R}$ was arbitrary above, the $\varphi^{(i)}(t)$ are linearly independent for all $t \in \mathbb{R}$. □

Corollary 1.3.14: Let $A : \mathbb{R} \to \mathbb{R}^{d\times d}$ and $c : \mathbb{R} \to \mathbb{R}^d$ be continuous and let $u_p(t)$ be the particular solution of the inhomogeneous, linear ODE

$$u'(t) = A(t)u(t) + c(t) \tag{1.23}$$

given by (1.18) with initial condition $u_p(0) = 0$ (i.e., with $t_0 = u_0 = 0$). Then every other solution of (1.23) is of the form $u(t) = u_p(t) + v(t)$ with $v \in H$.

Proof. Let u be another solution of (1.23). Then $v := u - u_p$ satisfies

$$v' = u' - u_p' = Au + c - Au_p - c = Av$$

and thus $v \in H$ (cf. Lemma 1.3.13). □

1.4 Well-posedness of the IVP

Definition 1.4.1: A mathematical problem is called **well-posed** if the following **Hadamard conditions** are satisfied:

1. A solution exists.
2. The solution is unique.
3. The solution depends continuously on the data.

The third condition in this form is purely qualitative. Typically, in order to characterize problems with good stability properties, we will require Lipschitz continuity, which has a more quantitative character.

Example 1.4.2. The IVP

$$u' = \sqrt[3]{u}, \qquad u(0) = 0,$$

has infinitely many solutions of the form

$$u(t) = \begin{cases} 0 & \text{for } t \in [0, c], \\ \left(\frac{2}{3}(t-c)\right)^{3/2} & \text{for } t > c, \end{cases}$$

with $c \in \mathbb{R}$. Thus, the solution is not unique and therefore, the IVP is not well-posed.

Let now the initial value be nonzero, but slightly positive. Then, the solution $u(t) \approx \left(\frac{2}{3}t\right)^{3/2}$ is unique. In contrast, when the initial value is slightly negative, no real-valued solution exists. Hence, a small perturbation of the initial condition has a dramatic effect on the solution; this is what Condition 3 for a well-posed problem in Definition 1.4.1 excludes.

Definition 1.4.3: The function $f(t,y)$ satisfies a uniform **Lipschitz condition** on the domain $D = I \times \Omega \subset \mathbb{R} \times \mathbb{R}^d$, if it is Lipschitz continuous with respect to y, i.e., there exists a constant $L > 0$, such that

$$\forall t \in I;\, x, y \in \Omega \ : \ |f(t,x) - f(t,y)| \le L|x-y| \tag{1.24}$$

It satisfies a local Lipschitz condition if (1.24) holds for all compact subsets of D.

Example 1.4.4. Let $f(t,y) \in C^1(\mathbb{R} \times \mathbb{R}^d)$ and let all partial derivatives with respect to the components of u be bounded such that

$$\max_{\substack{t\in\mathbb{R}\\ y\in\mathbb{R}^d\\ 1\le i,j\le d}} \left|\frac{\partial}{\partial y_i} f_j(t,y)\right| \le K.$$

Then, f satisfies the Lipschitz condition (1.24) with $L = Kd$. Indeed, by using the Fundamental Theorem of Calculus, we see that

$$\begin{aligned} f_j(t,y) - f_j(t,x) &= \int_0^1 \frac{d}{ds} f_j\big(t, x + s(y-x)\big)\,\mathrm{d}s \\ &= \int_0^1 \sum_{i=1}^d \frac{\partial}{\partial y_i} f_j\big(t, x + s(y-x)\big)(y_i - x_i)\,\mathrm{d}s. \end{aligned}$$

Now, exploiting the fact that $|Ax| \le \|A\|_F |x|$, where $\|A\|_F := \sqrt{\sum_{i,j=1}^d a_{ij}^2}$ is the Frobenius norm of the matrix A, we get

$$\begin{aligned} |f(t,y) - f(t,x)| &\le \int_0^1 \left|\sum_{i=1}^d \frac{\partial f}{\partial y_i}\big(t, x + s(y-x)\big)(y_i - x_i)\right| \mathrm{d}s \\ &\le \int_0^1 \left[\sum_{i,j=1}^d \left|\frac{\partial f}{\partial y_i}\big(t, x + s(y-x)\big)\right|^2\right]^{1/2} |y-x|\,\mathrm{d}s \le Kd|y-x|. \end{aligned}$$

Theorem 1.4.5 (Stability): Let $f(t,y)$ and $g(t,y)$ be two continuous functions on a cylinder $D = I \times \Omega$ where the interval I contains t_0 and Ω is a convex set in $\mathbb{R}^d$. Furthermore, let f admit a Lipschitz condition with constant L on D. Let u and v be solutions to the IVPs

$$u' = f(t,u) \quad \forall t \in I, \qquad u(t_0) = u_0, \tag{1.25}$$

$$v' = g(t,v) \quad \forall t \in I, \qquad v(t_0) = v_0. \tag{1.26}$$

Then

$$|u(t) - v(t)| \le e^{L|t-t_0|}\left[|u_0 - v_0| + \int_{t_0}^t \max_{x\in\Omega} |f(s,x) - g(s,x)|\,\mathrm{d}s\right]. \tag{1.27}$$

Proof. Both $u(t)$ and $v(t)$ solve their respective Volterra integral equations. Taking the difference, we obtain

$$\begin{aligned} u(t) - v(t) &= u_0 - v_0 + \int_{t_0}^{t} \big[f(s, u(s)) - g(s, v(s))\big] \,\mathrm{d}s \\ &= u_0 - v_0 + \int_{t_0}^{t} \big[f(s, u(s)) - f(s, v(s))\big] \,\mathrm{d}s + \int_{t_0}^{t} \big[f(s, v(s)) - g(s, v(s))\big] \,\mathrm{d}s. \end{aligned}$$

Thus, its norm admits the integral inequality

$$\begin{aligned} |u(t) - v(t)| &\le |u_0 - v_0| + \int_{t_0}^{t} |f(s, v(s)) - g(s, v(s))| \,\mathrm{d}s + \int_{t_0}^{t} |f(s, u(s)) - f(s, v(s))| \,\mathrm{d}s \\ &\le \underbrace{|u_0 - v_0| + \int_{t_0}^{t} \max_{x\in\Omega} |f(s, x) - g(s, x)| \,\mathrm{d}s}_{=:\, b(t)} + \int_{t_0}^{t} L|u(s) - v(s)| \,\mathrm{d}s. \end{aligned}$$

This inequality is in the form of the assumption in Grönwall's lemma with $a \equiv L$ and $w(t) := |u(t) - v(t)|$, and its application yields the stability result (1.27). □

Theorem 1.4.6 (Picard-Lindelöf): Let $f(t, y)$ be continuous and satisfy the Lipschitz condition (1.24) with constant L on a domain D that contains the closed set

$$\overline{D}_0 = \{(t, y) \in \mathbb{R} \times \mathbb{R}^d : |t - t_0| \le \alpha, |y - u_0| \le \beta\}.$$

Then f is also bounded on $\overline{D}_0$ with $M := \max_{(t,y)\in\overline{D}_0} |f(t, y)| < \infty$ and the IVP

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

is uniquely solvable on the interval $I = [t_0 - T, t_0 + T]$ where $T = \min\{\alpha, \frac{\beta}{M}\}$.

Proof. W.l.o.g., we assume $t_0 = 0$ and let

$$I := [-T, T] \text{ and } \Omega = \big\{y \in \mathbb{R}^d : |y - u_0| \le \beta\big\}.$$

The Volterra integral equation (1.12) allows us to define the operator

$$F(u)(t) := u_0 + \int\limits_{0}^{t} f(s, u(s)) \,\mathrm{d}s. \tag{1.28}$$

Obviously, u is a solution of the Volterra integral equation (1.12) if and only if u is a fixed point of F, i.e., $u = F(u)$. We can obtain such a fixed-point by the iteration $u^{(k+1)} = F(u^{(k)})$ with some initial guess $u^{(0)} : I \to \Omega$.

From the boundedness of f, we obtain for all $t \le T$ that

$$\Big|u^{(k+1)}(t) - u_0\Big| = \left|\int_{0}^{t} f(s, u^{(k)}(s)) \,\mathrm{d}s\right| \le \int_{0}^{t} |f(s, u^{(k)}(s))| \,\mathrm{d}s \le TM \le \beta.$$

Thus, it follows by an inductive argument that $u^{(k)} : I \to \Omega$ is well-defined for all $k \in \mathbb{N}$.

We now show that under the given assumptions, F is a contraction and then apply the Banach Fixed-Point Theorem. We follow the technique in [Heu86, §117] and choose on the space $\mathcal{C}(I)$ of continuous functions defined on I, the weighted maximum-norm

$$\|u\|_e := \max_{t\in I} e^{-2Lt}|u(t)|.$$

Then, for all $u,\ v \in \mathcal{C}(I)$,

$$\begin{aligned}
|F(u)(t) - F(v)(t)| &= \left| u_0 - u_0 + \int_0^t (f(s,u(s)) - f(s,v(s)))\,\mathrm{d}s \right| \\
&\le \int_0^t \big|f\big(s,u(s)\big) - f\big((s,v(s)\big)\big|\,\mathrm{d}s \;\le\; \int_0^t L|u(s)-v(s)|\underbrace{e^{-2Ls}e^{2Ls}}_{=1}\,\mathrm{d}s \\
&\le\; L\|u-v\|_e \int_0^t e^{2Ls}\,\mathrm{d}s \;=\; L\|u-v\|_e \frac{e^{2Lt}-1}{2L} \;\le\; \frac{1}{2}e^{2Lt}\|u-v\|_e
\end{aligned}$$

and by multiplying both sides with e^{-2Lt} it follows that

$$\|F(u) - F(v)\|_e \le \frac{1}{2}\|u-v\|_e.$$

Thus, we have shown that F is a contraction on $\mathcal{C}(I)$ with respect to the norm $\|\cdot\|_e$. Therefore, we can apply Theorem A.3.1, the Banach Fixed-Point Theorem, and conclude that F has exactly one fixed-point, which completes the proof. □

Remark 1.4.7. The norm $\|\cdot\|_e$ has been chosen with regard to Grönwall's inequality, which was not used explicitly in the proof. It is equivalent to the norm $\|\cdot\|_\infty$ because e^{-2Lt} is strictly positiv and bounded. One could have performed the proof also with respect to the ordinary maximum norm $\|\cdot\|_\infty$ (with some extra calculations).

Remark 1.4.8. As in Theorem 1.2.13, the Theorem of Picard-Lindelöf again only gives local existence and uniqueness of the solution on $I_0 = [t_0 - T, t_0 + T]$. However, as in Theorem 1.2.15 this local solution can be extended (uniquely) beyond $t_0 - T$ and $t_0 + T$ if f is bounded and Lipschitz on a larger time interval. Since T is chosen in such a way in Theorem 1.4.6 that the graph of u does not leave the domain, the local solution always ends at an interior point $(t_1, u_1) \in D$ with $t_1 := t_0 + T$. Thus, one can apply Theorem 1.4.6 again to obtain a unique solution of the IVP $u' = f(t,u)$ with initial condition $u(t_1) = u_1$ on an interval $I_1 = [t_1, t_2]$ with $t_2 > t_1$. Continuing in this way, one obtains a solution on $I_0 \cup I_1 \cup I_2 \cup \ldots$. Similarly, the solution can also be extended to the left of $t_0 - T$.

If f satisfies a Lipschitz condition everywhere then this leads to the following corollary.

Corollary 1.4.9. *Let the function $f(t,y)$ admit the Lipschitz condition* (1.24) *on all of $\mathbb{R} \times \mathbb{C}^d$. Then, the IVP has a unique solution for all $t \in \mathbb{R}$.*

Proof. In the proof of Theorem 1.4.6, the boundedness of f was used in order to guarantee that $u(t) \in \Omega$ for any t. This is not necessary anymore, since $\Omega = \mathbb{C}^d$. The fixed point argument does not rely on boundedness of the set. (A more detailed proof will be part of one of the *Exercise Sheets*.) □

Chapter 2

Explicit One-Step Methods and Convergence

2.1 Introduction

Example 2.1.1 (Euler's method)**.** We begin this section with the method that serves as the prototype for a whole class of schemes for solving IVPs numerically.

(As always for problems with infinite dimensional solution spaces, numerical solution refers to finding an approximation by means of a discretization method and the study of the associated discretisation error.)

Consider the following problem:

Given an IVP of the form (1.11) with $t_0 = 0$, calculate the value $u(T)$ at time $T > 0$.

Note first that at the initial point 0, not only the value $u(0) = u_0$ of u, but also the derivative $u'(0) = f(0, u_0)$ are known. Thus, near 0 we are able to approximate the solution $u(t)$ (in blue in Figure 2.1) by a straight line $y(t)$ (in red in Figure 2.1, left) using a first-order Taylor series expansion, i.e.

$$u(t) \approx y(t) = u(0) + tu'(0) = u_0 + tf(0, u_0) \ .$$

The figure suggests that in general the accuracy of this method may not be very good for t far from 0. The first improvement is that we do not draw the line through the whole interval from 0 to T. Instead, we insert intermediate points and apply the method to each subinterval, using the result of the previous interval as the initial point for the next. As a result we obtain a continuous chain of straight lines (in red in Figure 2.1, right) and the so-called **Euler method** (details below).

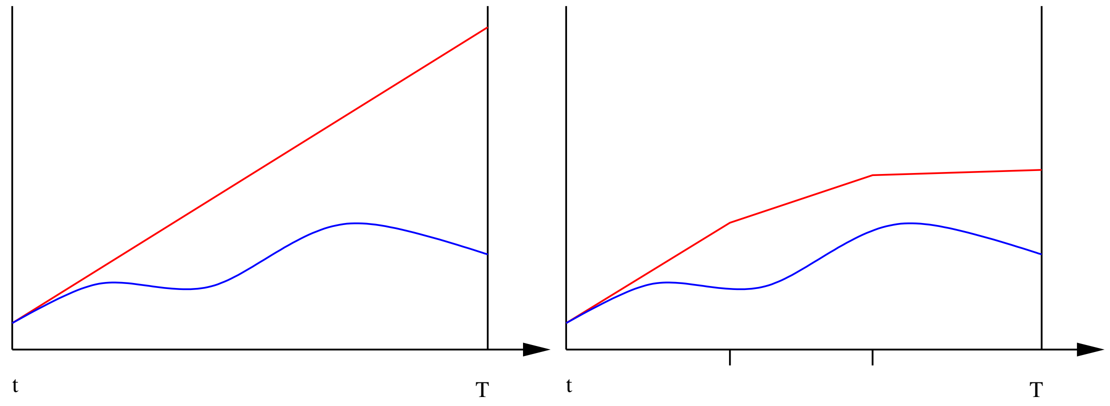
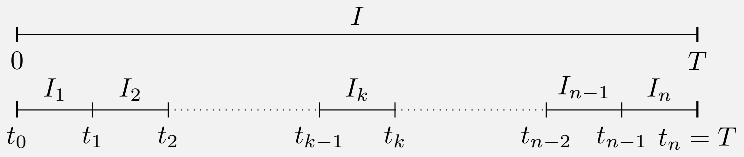


Figure 2.1: Derivation of the Euler method. Left: approximation of the solution of the IVP by a line that agrees in slope and value with the solution at $t = 0$. Right: Euler method with three subintervals.

Definition 2.1.2: On a time interval $I = [0, T]$, we define a partitioning in n subintervals, also known as **time steps**. Here we choose the following notation:



The time steps $I_k = [t_{k-1}, t_k]$ have **step size** $h_k = t_k - t_{k-1}$. A partitioning in n time steps implies $t_n = T$. The term "k-th time step" is used both for the interval I_k and for the point in time t_k (which one is meant will be clear from the context). Very often, we will consider **uniform time steps** and in that case the step size is denoted by h and $h_k = h$, for all k.

Definition 2.1.3 (Notation)**.** In the following chapters we will regularly compare the solution of an IVP with the results of discretization methods. Therefore, we introduce the following convention for notation and symbols.

The solution of the IVP is called the **exact** or **continuous solution**. to emphasize that it is the solution of the non-discretized problem. We denote it in general by u and in addition we use the abbreviation

$$u_k = u(t_k).$$

If u is vector-valued we also use the alternative superscript $u^{(k)}$ and write $u_i^{(k)}$ for the ith component of the vector $u(t_k)$.

The **discrete solution** is in general denoted by y and we write y_k or $y^{(k)}$ for the value of the discrete solution at time t_k. In contrast to the continuous solution, y is only defined at discrete time steps (except for special methods discussed later).

Definition 2.1.4 (Explicit one-step methods): An **explicit one-step method** is a method which, given u_0 at $t_0 = 0$ computes a sequence of approximations $y_1 \dots, y_n$ to the solution of an IVP at the time steps $t_1, \dots, t_n$ using an update formula of the form

$$y_k = y_{k-1} + h_k F_{h_k}(t_{k-1}, y_{k-1}). \tag{2.1}$$

The function $F_{h_k}()$ is called **increment function.**[a] We will often omit the index h_k on $F_{h_k}()$ because it is clear that the method is always applied to time intervals. The method is called **one-step method** because the value y_k explicitly depends only on the values y_{k-1} and $f(t_{k-1}, y_{k-1})$, not on previous values.

[a] The adjective 'explicit' is here in contrast to 'implicit' one-step methods, where the increment function depends also on y_k and equation (2.1) typically leads to a nonlinear equation for y_k.

Remark 2.1.5. For one-step methods every step is per definition identical. Therefore, it is sufficient to define and analyze methods by stating the dependence of y_1 on y_0, which then can be transferred to the general step from y_{n-1} to y_n. The general one-step method above then reduces to

$$y_1 = y_0 + h_1 F_{h_1}(t_0, y_0).$$

This implies that the values y_k with $k \geq 2$ are computed through formula (2.1) with the respective h_k and the same increment function (but evaluated at t_{k-1}, y_{k-1}).

Example 2.1.6: The simplest choice for the increment function is $F_{h_k}(t, y) := f(t, y)$, leading to the **Euler method**

$$y_1 = y_0 + hf(t_0, y_0). \tag{2.2}$$

Consider, for example, the (scalar, homogeneous, linear) IVP

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

which has exact solution $u(t) = e^t$. In that case, the Euler method (with uniform time steps) reads

$$y_1 = (1 + h)y_0.$$

The results for $h = 1$ and $h = 1/2$ are:

exact		$h = 1$			$h = 1/2$		
t	$u(t)$	k	y_k	$\lvert u_k - y_k\rvert$	k	y_k	$\lvert u_k - y_k\rvert$
0	1	0	1		0	1	
1	2.71828	1	2	0.718	2	2.25	0.468
2	7.38906	2	4	3.389	4	5.0625	2.236
k	2.71828^k	k	2^k		$2k$	2.25^k	

The error is growing in time. The approximation of the solution is improved by reducing h from 1 to $1/2$. The goal of the following error analysis will be to establish those dependencies.

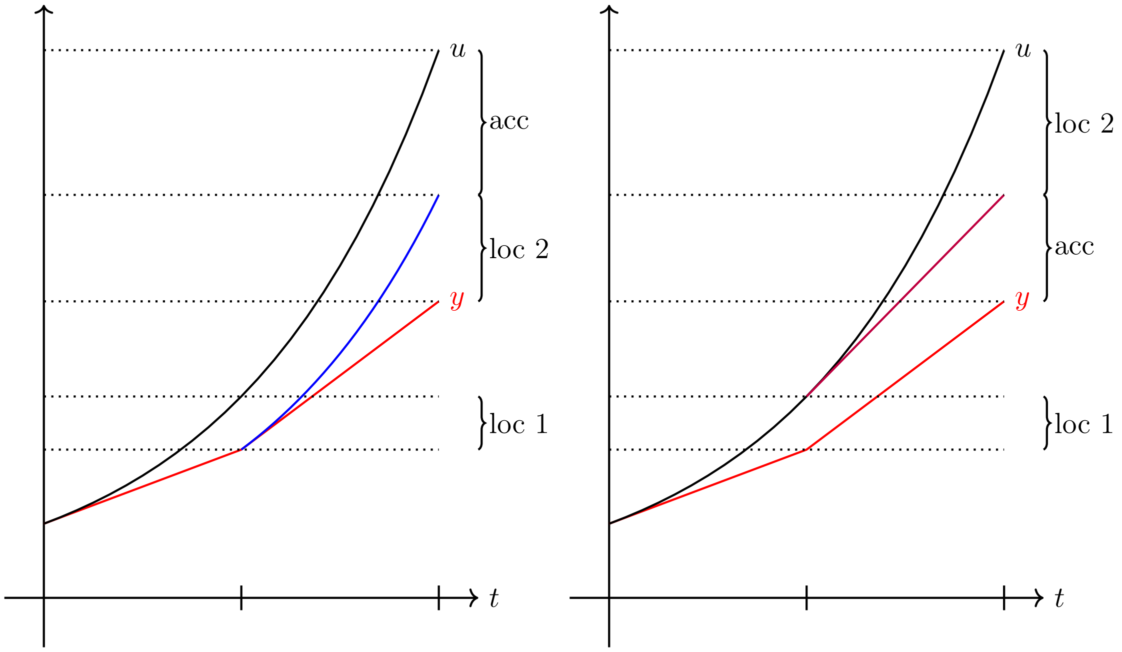


Figure 2.2: Local and accumulated errors. Exact solution in black, the Euler method in red. On the left, in blue the exact solution of an IVP on the second interval with initial value y_1. On the right, in purple the second step of the Euler method, but with exact initial value u_1.

2.2 Error analysis

Remark 2.2.1. In Figure 2.1, we observe that, at a given time t_{k+1}, the error consists of two parts: (i) due to replacing the differential equation by the discrete method on the interval I_k and (ii) due the initial value y_k already being inexact. This is illustrated more clearly in Figure 2.2. Therefore, in our analysis we split the error into the local error and an accumulated error. The local error compares continuous and discrete solutions on a single interval with the same initial value. In the analysis, we will have the options of using the exact (right figure) or the approximated initial value (left figure).

Definition 2.2.2: Let u be a solution of the differential equation $u' = f(t,u)$ on the interval $I = [t_0, t_n] = [0, T]$. Then, the **global error** of a discrete method F_{h_n} is

$$|u(t_n) - y(t_n)|, \tag{2.3}$$

i.e., the difference between the solution u_n of the differential equation at t_n and the result of the one-step method at t_n.

Definition 2.2.3: Let u be a solution of the differential equation $u' = f(t,u)$ on the interval $I_k = [t_{k-1}, t_k]$. Then, the **local error** of a discrete method F_{h_k} is

$$\eta_k = \eta_k(u) = u_k - \big[u_{k-1} + h_k F_{h_k}(t_{k-1}, u_{k-1})\big], \tag{2.4}$$

i.e., the difference between $u_k = u(t_k)$ and the result of one time step (2.1) with this method with exact initial value $u_{k-1} = u(t_{k-1})$.

The **truncation error** is the quotient of the local error and h_k:

$$\tau_k = \tau_k(u) = \frac{\eta_k}{h_k} = \frac{u_k - u_{k-1}}{h_k} - F_{h_k}(t_{k-1}, u_{k-1}). \tag{2.5}$$

The one-step method $F_{h_k}(t,y)$ is said to have **consistency of order** p, if for all sufficiently regular functions f there exists a constant c independent of $h := \max_{k=1}^n h_k$ such that for $h \to 0$:

$$\max_{k=1}^{n} |\tau_k| \;\le\; ch^p \tag{2.6}$$

Example 2.2.4 (Euler method)**.** To find the order of consistency of the Euler method, where $F_{h_k}(t,y) = f(t,y)$, consider Taylor expansion of u at t_{k-1}:

$$u(t_k) = u(t_{k-1}) + h_k u'(t_{k-1}) + \frac{1}{2} h_k^2 u''(\zeta), \quad \text{for some} \;\; \zeta \in I_k.$$

As a result the truncation error reduces to:

$$\begin{aligned}
\tau_k &= \frac{u_k - u_{k-1}}{h_k} - F_{h_k}(t_{k-1}, u(t_{k-1})) \\
&= \frac{h_k f(t_{k-1}, u_{k-1}) + \frac{1}{2} h_k^2 u''(\zeta)}{h_k} - f(t_{k-1}, u_{k-1}) \;\;=\;\; \frac{1}{2} u''(\zeta)\, h_k\,.
\end{aligned}$$

If $f \in C^1(D)$ on a compact set D around the graph of u, we can bound the right hand side:

$$\begin{aligned}
|\tau_k| \le \frac{1}{2} \max_{t \in I_k} |u''(t)|\, h_k &= \frac{1}{2} \max_{t \in I_k} \left| \frac{\partial f}{\partial t}(t, u(t)) + \nabla_y f(t, u(t)) u'(t) \right| h_k \\
&\le \underbrace{\frac{1}{2} \max_{(t,y) \in D} \left| \frac{\partial f}{\partial t}(t, y) + \nabla_y f(t,y) f(t,y) \right|}_{=:\, c} h_k\,.
\end{aligned}$$

Here, we use the assumption that f is sufficiently smooth to conclude that the Euler method is consistent of order 1 (slightly more than Lipschitz continuous).

Next we consider stability of explicit one-step methods. To prove this, we first need a discrete version of Grönwall's inequality.

Lemma 2.2.5 (Discrete Grönwall inequality): Let (w_k), (a_k), (b_k) be non-negative sequences of real numbers, such that (b_k) is monotonically nondecreasing. Then, it follows from

$$w_0 \le b_0 \quad \text{and} \quad w_n \le \sum_{k=0}^{n-1} a_k w_k + b_n, \quad \text{for all} \;\; n \ge 1, \tag{2.7}$$

that

$$w_n \le \exp\left(\sum_{k=0}^{n-1} a_k\right) b_n. \tag{2.8}$$

Proof. Let $k \in \mathbb{N}$ and define the functions $w(t)$, $a(t)$, and $b(t)$ such that for all $t \in [k-1,k)$

$$w(t) = w_{k-1}, \quad a(t) = a_{k-1}, \quad b(t) = b_{k-1}.$$

These functions are bounded and piecewise continuous on any finite interval. Thus, they are integrable on $[0,n]$. Therefore, the continuous Grönwall inequality of Lemma 1.3.8 applies and proves the result. □

Theorem 2.2.6 (Discrete stability): If $F_{h_k}(t,y)$ is Lipschitz continuous in y for any $t = t_k$, $k < n$, with constant L, then the corresponding one-step method is **discretely stable** in the following sense: Suppose (x_k) and (z_k) are two arbitrary sequences, e.g. from two numerical approximations of the IVP with F_{h_k} with different initial conditions u_0 and v_0, or the approximate and the exact solution at $t = t_k$. Then

$$|x_n - z_n| \le e^{LT}\left(|x_0 - z_0| + \sum_{k=1}^{n} |\eta_k(x) - \eta_k(z)|\right)$$

Proof. Subtracting the equations

$$\begin{aligned} \eta_k(x) &= x_k - x_{k-1} - h_k F_{h_k}(t_{k-1}, x_{k-1}), \\ \eta_k(z) &= z_k - z_{k-1} - h_k F_{h_k}(t_{k-1}, z_{k-1}), \end{aligned}$$

and exploiting the Lipschitz continuity of F_{h_k}, we obtain

$$\begin{aligned} |x_k - z_k| &= \bigl|x_{k-1} - z_{k-1} + \eta_k(x) - \eta_k(z) + h_k\bigl(F_{h_k}(t_{k-1}, x_{k-1}) - F_{h_k}(t_{k-1}, z_{k-1})\bigr)\bigr| \\ &\le |x_{k-1} - z_{k-1}| + |\eta_k(x) - \eta_k(z)| + Lh_k|x_{k-1} - z_{k-1}|. \end{aligned}$$

Recursive application yields

$$|x_n - z_n| \le |x_0 - z_0| + \sum_{k=1}^{n} |\eta_k(x) - \eta_k(z)| + \sum_{k=1}^{n} Lh_k |x_{k-1} - z_{k-1}|.$$

The estimate now follows from the discrete Grönwall inequality in Lemma 2.2.5, choosing $w_n = |x_n - z_n|$, $a_{n-1} = Lh_n$ and $b_n = |x_0 - z_0| + \sum_{j=k}^{n} |\eta_k(x) - \eta_k(z)|$. □

Corollary 2.2.7 (One-step methods with finite precision): Let the one-step method F_{h_k} be run in finite precision floating-point arithmetic on a computer, yielding a (perturbed) sequence (z_k). Let (y_k) be the mathematically correct solution of the one-step method. Then, the difference equation (2.1) is fulfilled only up to machine accuracy eps, i.e., there exists a $c > 0$:

$$\begin{aligned} |y_0 - z_0| &\leq c|z_0| \,\mathrm{eps}\,, \\ |\eta_k(y) - \eta_k(z)| = |\eta_k(z)| &\leq c|z_k| \,\mathrm{eps}\,. \end{aligned}$$

Then, the error between the true solution of the one-step method y_n and the computed solution z_n is bounded by

$$|y_n - z_n| \leq c\, e^{LT} n \max_{k=0}^{n} |z_k| \,\mathrm{eps}.$$

Theorem 2.2.8 (Convergence of one-step methods): Let the one-step method F_{h_k} be consistent of order p and Lipschitz continuous in its second argument, for all $t = t_k$, $k < n$. Furthermore, let $y_0 = u_0$ and let $h = \max_{k=1}^{n} h_k$. Then, the global error of the one-step method converges with order p as $h \to 0$ and

$$|u_n - y_n| \leq cTe^{LT} h^p, \tag{2.9}$$

where the constant c is independent of h.

Proof. Since F_{h_k} is consistent of order p, we have, for all $k = 1, \dots, n$,

$$|\eta_k(u) - \underbrace{\eta_k(y)}_{=0}| = h_k |\tau_k(u)| \leq ch^p\, h_k, \tag{2.10}$$

where c is the constant in (2.6), which is independent of h.

Now, since F_{h_k} is Lipschitz continuous in its second argument, we can apply the discrete stability result in Theorem 2.2.6 with $x_k = y_k$ and $z_k = u_k$ and use the bound in (2.10) to obtain

$$|u_n - y_n| \leq e^{LT} \sum_{k=1}^{n} |\eta_k(u) - \eta_k(y)| \leq e^{LT}\, ch^p \sum_{k=1}^{n} h_k = cT\, e^{LT} h^p.$$

□

Corollary 2.2.9. *If f is in C^1 in a compact set D around the graph of u over $[0, T]$, then the convergence order of the global error in the Euler method is one.*

Important! General approach:

$$\boxed{\text{CONSISTENCY} \;+\; \text{STABILITY} \;=\; \text{CONVERGENCE}} \tag{2.11}$$

2.3 Runge-Kutta methods

2.3.1. Since the IVP is equivalent to the Volterra integral equation (1.12), we can consider its numerical solution as a quadrature problem. However, the difficulty is that the integrand is not known. It will need to be approximated on each interval from values at earlier times, leading to a class of methods for IVP, called Runge-Kutta methods.

We present the formula again only for the calculation of y_1 from y_0 on the interval from t_0 to $t_1 = t_0 + h$. The formula for a later time step k is obtained by replacing y_0, t_0 and h by y_{k-1}, t_{k-1} and h_k, respectively to obtain y_k. (The coefficients a_{ij}, b_j, c_j remain fixed.)

Definition 2.3.2: An **explicit Runge-Kutta method (ERK)** is a one-step method that uses s evaluations of f with the representation

$$g_i = y_0 + h\sum_{j=1}^{i-1} a_{ij}k_j\,, \qquad i = 1,\dots,s, \tag{2.12a}$$

$$k_i = f\left(t_0 + c_ih, g_i\right)\,, \qquad i = 1,\dots,s, \tag{2.12b}$$

$$y_1 = y_0 + h\sum_{i=1}^{s} b_ik_i\,, \tag{2.12c}$$

i.e., with increment function

$$F_h(t, y_0) := \sum_{i=1}^{s} b_if\left(t_0 + c_ih, g_i\right).$$

The values $t_0 + c_ih$ are the quadrature points on the interval $[t_0, t_1]$. The values g_i are approximations to the solution $u(t_0 + c_ih)$ at the quadrature points, obtained via recursive extrapolation using the evaluations of f at previous quadrature points. Since the method uses s intermediate approximations of u on $[t_0, t_1]$, it is called an s-stage method.

Remark 2.3.3. In typical implementations, the intermediate values g_i are not stored separately. However, they are useful for highlighting the structure of the method.

Definition 2.3.4 (Butcher tableau): It is customary to write Runge-Kutta methods in the form of a **Butcher tableau**, containing only the coefficients of equation (2.12) in the following matrix form:

$$\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array} \tag{2.13}$$

Remark 2.3.5. The first row of the tableau should be read such that $c_1 = 0$, $g_1 = y_0$ and that k_1 is computed directly by $f(t_0, y_0)$. The method is *explicit* since the computation of

k_i only involves coefficients with index less than i. The row below the line is the short form of formula (2.12c) and lists the quadrature weights in the increment function $F_h(t, y_0)$.

Considering the coefficients a_{ij} as the entries of a square $s \times s$ matrix A, we see that A is strictly lower triangular. This is the defining feature of an explicit RK method. We will later see RK methods that also have entries on the diagonal or even in the upper part. Those methods will be called *"implicit"*, because the computation of a stage value also involves the values at the current or future stages. We will also write $b = (b_1, \dots, b_s)^T$ and $c = (0, c_2, \dots, c_s)^T$, such that the Butcher tableau in (2.13) simplifies to

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

Example 2.3.6. The Euler method has the Butcher tableau:

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

That leads to the already known formula:

$$y_1 = y_0 + hf(t_0, y_0)$$

The values $b_1 = 1$ and $c_1 = 0$ indicate that this is a quadrature rule with a single point at the left end of the interval. As a quadrature rule, such a rule is exact for constant polynomials and thus of order 1.

Example 2.3.7 (Two-stage methods): The **modified Euler method** is a variant of Euler's method of the following form:

$$\begin{aligned} k_1 &= f(t_0, y_0) \\ k_2 &= f\left(t_0 + \frac{1}{2}h_1, y_0 + h_1\frac{1}{2}k_1\right) \\ y_1 &= y_0 + h_1 k_2 \end{aligned} \qquad \begin{array}{c|cc} 0 & & \\ \frac{1}{2} & \frac{1}{2} & \\ \hline & 0 & 1 \end{array}$$

The so-called **Heun method of order 2** is characterized through the equation

$$\begin{aligned} k_1 &= f(t_0, y_0) \\ k_2 &= f(t_0 + h_1, y_0 + h_1 k_1) \\ y_1 &= y_0 + h\left(\frac{1}{2}k_1 + \frac{1}{2}k_2\right) \end{aligned} \qquad \begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

Remark 2.3.8. The modified Euler method uses one quadrature node at $t_0 + \frac{h}{2} = \frac{t_0+t_1}{2}$ and an approximation to $f(t_0 + \frac{h}{2}, u(t_0 + \frac{h}{2}))$ in F_h, corresponding to the midpoint quadrature rule. The Heun method of order 2 is constructed based on the trapezoidal rule. Both quadrature rules are of second order, and so are these one-step methods. Both methods were discussed by Runge in his article of 1895 [Run95].

Lemma 2.3.9: For f sufficiently smooth, the Heun method of order 2 and the modified Euler method have consistency order two.[a]

[a] Here and in the following proofs of consistency order, we will always assume that all necessary derivatives of f exist and are bounded and simply write "f is sufficiently smooth".

Proof. The proof uses Taylor expansion of the continuous solution u_1 and the discrete solution y_1 around t_0 with $y_0 = u_0$. W.l.o.g. we choose $t_0 = 0$. Considering first only the case $d = 1$ and the abbreviations

$$f_t = \partial_t f(t_0, u_0) \quad \text{and} \quad f_y = \partial_y f(t_0, u_0)$$

and replacing $u'(t_0) = f(t_0, u_0) = f$, we obtain

$$\begin{aligned} u_1 = u(h) = u_0 + hf(t_0, u_0) + \frac{h^2}{2}\big(f_t + f_y f\big) \\ + \frac{h^3}{6}\big(f_{tt} + 2f_{ty}f + f_{yy}f^2 + f_y f_t + f_y^2 f\big) + \dots \end{aligned} \tag{2.14}$$

For the discrete solution of the modified Euler step on the other hand, Taylor expanding f around (t_0, u_0) leads to

$$\begin{aligned} y_1 &= u_0 + hf\left(t_0 + \frac{h}{2}, u_0 + \frac{h}{2}f(t_0, u_0)\right) \\ &= u_0 + hf + \frac{h^2}{2}\big(f_t + f_y f\big) + \frac{h^3}{8}\big(f_{tt} + 2f_{ty}f + f_{yy}f^2\big) + \dots . \end{aligned}$$

Thus, $|\tau_1| = h^{-1}|u_1 - y_1| = \mathcal{O}(h^2)$. Since the truncation error at the kth step can be estimated identically, the method has consistency order two. The proof for the Heun method is left as an exercise.

For $d > 1$, the derivatives with respect to y are no longer scalars, but tensors of increasing rank, or equivalently multilinear operators. Thus, to be precise in d dimensions, $\partial_y f(t_0, u_0)$ is a $d \times d$ matrix that is applied to the vector $f(t_0, u_0)$ and we should write more carefully

$$f_y(f) = \partial_y f(t_0, u_0) f(t_0, u_0).$$

Similarly, $\partial_{yy} f(t_0, u_0)$ is a $d \times d \times d$ rank-3 tensor, or more simply a bilinear operator and to stress this we write $f_{yy}(f, f)$ instead of $f_{yy}f^2$. (However, we will not dwell on this issue in this course.) □

Example 2.3.10: A suitable three stage Runge-Kutta method is

$$\begin{aligned} k_1 &= f(t_0, y_0) \\ k_2 &= f\left(t_0 + \frac{1}{2}h, y_0 + \frac{1}{2}hk_1\right) \\ k_3 &= f(t_0 + h, y_0 - hk_1 + 2hk_2) \\ y_1 &= y_0 + h\left(\frac{1}{6}k_1 + \frac{4}{6}k_2 + \frac{1}{6}k_3\right) \end{aligned}
\qquad
\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2} & & \\ 1 & -1 & 2 & \\ \hline & \frac{1}{6} & \frac{4}{6} & \frac{1}{6} \end{array}$$

This method is obviously based on the Simpson rule.

Remark 2.3.11. These Taylor series expansions become tedious very fast. For autonomous ODEs $u' = f(u)$ the analysis simplifies significantly. The Runge-Kutta method (2.12) reduces to

$$
\begin{aligned}
g_i &= y_0 + h\sum_{j=1}^{i-1} a_{ij} f(g_j), \quad i = 1,\dots,s \\
y_1 &= y_0 + h\sum_{j=1}^{s} b_j f(g_j).
\end{aligned}
\tag{2.15}
$$

Each (non-autonomous) ODE can be autonomized (see Def. 1.2.6) using the transformation

$$
U' := \begin{pmatrix} u' \\ t' \end{pmatrix} = \begin{pmatrix} f(t,u) \\ 1 \end{pmatrix} =: F(U). \tag{2.16}
$$

Lemma 2.3.12: An ERK is invariant under autonomization, i.e. its coefficients remain unchanged, if and only if

$$
\sum_{j=1}^{i-1} a_{ij} = c_i, \quad i = 1,\dots,s, \quad \text{and} \quad \sum_{j=1}^{s} b_j = 1\,. \tag{2.17}
$$

Proof. Considering the last components of the vector g_i in (2.15) when applied to the autonomized ODE (2.16) with right hand side $F(\cdot)$, we obtain

$$
t_0 + h\sum_{j=1}^{i-1} a_{ij}\,.
$$

For the ERK to be invariant under autonomization, we require that f is evaluated at t_0+hc_i in the ith stage leading to the first condition in (2.17). Similarly, the second condition in (2.17) follows from the last component of y_1, when applying (2.15) to (2.16). □

Lemma 2.3.13: An ERK that is invariant under autonomization with s stages is consistent of *first order,* if and only if

$$
b_1 + \dots + b_s = 1; \tag{2.18a}
$$

it is consistent of *second order*, if and only if in addition we have

$$
b_1c_1 + \dots + b_sc_s = 1/2 \tag{2.18b}
$$

it is consistent of *third order*, if and only if in addition we have

$$
b_1c_1^2 + \dots + b_sc_s^2 = 1/3, \tag{2.18c}
$$

$$
\sum\nolimits_{i,j=1}^{s} b_i a_{ij} c_j = 1/6; \tag{2.18d}
$$

it is consistent of *fourth order*, if and only if in addition we have

$$
b_1c_1^3 + \dots + b_sc_s^3 = 1/4, \tag{2.18e}
$$

$$
\sum\nolimits_{i,j=1}^{s} b_i a_{ij} c_j^2 = 1/12, \tag{2.18f}
$$

$$
\sum\nolimits_{i,j,k=1}^{s} b_i a_{ij} a_{jk} c_k = 1/24, \tag{2.18g}
$$

$$
\sum\nolimits_{i,j=1}^{s} b_i c_i a_{ij} c_j = 1/8. \tag{2.18h}
$$

Proof. We consider the autonomous ODE $u' = f(u)$ and $u(t_0) = u_0$, where we assume w.l.o.g. again $t_0 = 0$. As in the proof of lemma 2.3.9, we first expand $u_1 = u(t_1)$ around t_0, using $u'(t_0) = f(u_0) = f$. Also, since f now only depends on one argument, we abbreviate

$$\frac{\mathrm{d}}{\mathrm{d}t} f(u(t))\Big|_{t=t_0} = \nabla f(u_0) f(u_0) =: f'(f), \quad \frac{\mathrm{d}^2}{\mathrm{d}t^2} f(u(t))\Big|_{t=t_0} =: f''(f,f) + f'(f'(f)), \quad \dots$$

Thus, we obtain

$$\begin{aligned} u_1 = u_0 + hf + \frac{h^2}{2} f'(f) + \frac{h^3}{6}\Big(f''(f,f) + f'(f'(f))\Big) \\ + \frac{h^4}{24}\Big(f'''(f,f,f) + 3f''(f'(f),f) + f'(f''(f,f)) + f'(f'(f'(f)))\Big) + \mathcal{O}(h^5). \end{aligned} \tag{2.19}$$

To expand y_1 around $t_0 = 0$ we consider it as a function $y_1(h)$ of the stepsize h. The stage values g_i are also considered as functions $g_i(h)$ of h. (Careful! The g_i, and thus also y_1, depend implicitly on h and are <u>not</u> affine!)

First note that

$$y_1(0) = u_0 \quad \text{and} \quad g_i(0) = u_0, \quad \text{for all} \;\; i = 1, \dots, s. \tag{2.20}$$

To compute the derivatives of y_1 and g_i at $h = 0$, let $q \geq 1$ and note that for an arbitrary function $\varphi = \varphi(h)$, applying Leibniz's rule (the product rule for higher derivatives), gives

$$\begin{aligned} \frac{d^q}{dh^q}\Big(h\varphi(h)\Big)\Big|_{h=0} &= \left[h\varphi^{(q)}(h) + \binom{q}{1}\underbrace{h'}_{=1}\varphi^{(q-1)}(h) + \binom{q}{2}\underbrace{h''}_{=0}\varphi^{(q-2)}(h) + 0\right]_{h=0} \\ &= q\varphi^{(q-1)}(0). \end{aligned} \tag{2.21}$$

Using (2.21) and the definition of an ERK for an autonomous ODE in (2.15) we get

$$y_1^{(q)}(0) = 0 + \sum_{i=1}^{s} b_i \frac{d^q}{dh^q}\Big(hf(g_i(h))\Big)\Big|_{h=0} = q\sum_{i=1}^{s} b_i \frac{d^{q-1}}{dh^{q-1}} f(g_i(h))\Big|_{h=0}, \tag{2.22}$$

$$g_i^{(q)}(0) = 0 + \sum_{j=1}^{s} a_{ij} \frac{d^q}{dh^q}\Big(hf(g_j(h))\Big)\Big|_{h=0} = q\sum_{j=1}^{s} a_{ij} \frac{d^{q-1}}{dh^{q-1}} f(g_j(h))\Big|_{h=0}. \tag{2.23}$$

(where for simplicity we have set $a_{ij} = 0$, for $j \geq i$).

Finally, we need to apply the chain rule to compute the derivatives of $f(g_i(h))$ needed in (2.22) and (2.23). First for $q = 1$, it follows from (2.17), (2.20) and (2.23) – using again the shorthand notation for the higher derivatives of f as above – that

$$\frac{d}{dh} f(g_i(h))\Big|_{h=0} = \nabla f(g_i(0)) g_i'(0) = f'\left(\sum_{j=1}^{s} a_{ij} f(g_j(0))\right) = \sum_{j=1}^{s} a_{ij} f'(f) = c_i f'(f)$$

Similarly, for $q = 2$, we get

$$\begin{aligned} \frac{d^2}{dh^2} f(g_i(h))\Big|_{h=0} &= \left[f''(g_i'(h), g_i'(h)) + f'(g_i''(h))\right]\Big|_{h=0} \\ &= f''\left(\sum_{j=1}^{s} a_{ij} f(g_j(0)), \sum_{j=1}^{s} a_{ij} f(g_j(0))\right) + f'\left(2\sum_{j=1}^{s} a_{ij} \frac{d}{dh} f(g_j(h))\Big|_{h=0}\right) \\ &= c_i^2 f''(f,f) + 2\sum_{j=1}^{s} a_{ij} c_j f'(f'(f)). \end{aligned}$$

Substituting these formulae into (2.22), we finally obtain

$$\begin{aligned}
y_1'(0) &= \left(\sum_{i=1}^{s} b_i\right) f,\\
y_1''(0) &= 2\left(\sum_{i=1}^{s} b_i c_i\right) f'(f),\\
y_1'''(0) &= 3\left(\sum_{i=1}^{s} b_i c_i^2\right) f''(f,f) + 6\left(\sum_{i,j=1}^{s} b_i a_{ij} c_j\right) f'(f'(f)).
\end{aligned}$$

Considering now the Taylor series expansion of $y_1(h)$ around $h = 0$, i.e.,

$$y_1(h) = y_1(0) + hy_1'(0) + \frac{h^2}{2}y_1''(0) + \frac{h^3}{6}y_1'''(0) + \frac{h^4}{24}y_1^{(4)}(0) + \mathcal{O}(h^5)$$

and comparing coefficients with the coefficients in the expansion of u_1 in (2.19), we obtain the order conditions in (2.18a-d).

The case $q = 3$ and thus the 4th derivative $y_1^{(4)}(0)$ can be derived in a similar way. Comparing the coefficients in front of the 4th-order term we then also obtain the order conditions in (2.18e-h). This is left as an exercise $\boxed{\text{DIY}}$. □

Remark 2.3.14. Butcher introduced a graph theoretical method for order conditions based on trees. While this simplifies the process of deriving these conditions for higher order methods considerably, it is beyond the scope of this course.

Example 2.3.15 (The classical Runge-Kutta method of 4th order):

$$\begin{aligned}
k_1 &= f(t_0, y_0)\\
k_2 &= f\left(t_0 + \frac{1}{2}h, y_0 + \frac{1}{2}hk_1\right)\\
k_3 &= f\left(t_0 + \frac{1}{2}h, y_0 + \frac{1}{2}hk_2\right)\\
k_4 &= f(t_0 + h, y_0 + hk_3)\\
y_1 &= y_0 + h(\frac{1}{6}k_1 + \frac{2}{6}k_2 + \frac{2}{6}k_3 + \frac{1}{6}k_4)
\end{aligned}$$

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

Like the 3-stage method in Example 2.3.10, this formula is based on Simpson's quadrature rule, but it uses two approximations for the value at the midpoint of the interval and is of order 4.

Remark 2.3.16 (Order conditions and quadrature)**.** The order conditions derived by recursive Taylor expansion have a very natural interpretation via the analysis of quadrature formulae for the Volterra integral equation, where $c_i h$, $i = 1, \ldots, s$, are the quadrature points on $[0, h]$ and the other coefficients are quadrature weights.

First, we observe that

$$h\sum_{i=1}^{s} b_i f(c_i h, g_i) \quad \text{approximates} \quad \int_0^h f(s, u(s))\,\mathrm{d}s.$$

In this view, conditions (2.18a)–(2.18c) and (2.18e) state that the quadrature formula $\sum_i b_i f(c_i h)$ is exact for polynomials f of degree up to 3. This implies (see Numerik 0) that the convergence of the quadrature rule is of 4th order.

Equally, we deduce from formula (2.12a) for g_i that

$$h \sum_{j=1}^{i-1} a_{ij} f(c_j h, g_j) \quad \text{approximates} \quad \int_0^{c_i h} f(s, u(s))\,\mathrm{d}s.$$

The condition (2.17), which guarantees that the method is autonomizable, simply translates to the quadrature rule being exact for constant functions.

For higher order, the accuracy of the value of g_i only implicitly enters the accuracy of the Runge-Kutta method as an approximation of the integrand in another quadrature rule. Thus, we actually look at approximations of nested integrals of the form

$$\int_0^h \varphi(s) \int_0^s \psi(r)\,\mathrm{d}r\,\mathrm{d}s.$$

Condition (2.18d) for 3rd order states, that this condition must be true for linear polynomials $\psi(r)$ and constant $\varphi(s)$; thus, after the inner integration again any polynomial of second order in the outer rule. Equally, conditions (2.18h) and (2.18f) for fourth order state this for linear polynomials $\psi(r)$ with linear $\varphi(s)$ and for quadratic polynomials $\psi(r)$ with constant $\varphi(s)$, respectively. Finally, condition (2.18g) states that the quadrature has to be exact for any linear polynomial $\varphi(\tau)$ in

$$\int_0^h \int_0^s \int_0^r \varphi(\tau)\,\mathrm{d}\tau\,\mathrm{d}r\,\mathrm{d}s.$$

Remark 2.3.17 (Butcher barriers)**.** The maximal order of an explicit Runge-Kutta method is limited through the number of stages, or vice versa, a minimum number of stages is required for a certain order. The **Butcher barriers** state that in order to achieve consistency of order p one requires s stages, where p and s are related as follows:

p	1	2	3	4	5	6	7	8
s	p	p	p	p	p+1	p+1	p+2	p+3
# cond.	1	2	4	8	17	37	85	200

The Butcher barriers for $p \geq 9$ are not known yet.

Lemma 2.3.18: Let $f(t,y)$ admit a uniform Lipschitz condition on $[0,T] \times \Omega$ with $\{u(t) : t \in [0,T]\} \subset \Omega$. Then every ERK that is invariant under autonomization admits a uniform Lipschitz condition on $[0,T] \times \Omega$.

Proof. The increment function of an ERK is

$$F_h(t,y) = \sum_{i=1}^{s} b_i f\Big(t + c_i h, g_i(y;h)\Big), \tag{2.24}$$

with g_i defined recursively by

$$g_i(y;h) = y + h \sum_{j=1}^{i-1} a_{ij} f\Big(t + c_j h, g_j(y;h)\Big).$$

Let L_0 be the Lipschitz constant of f and let $q := hL_0$ and assume for the moment that $a_{ij} \geq 0$, $1 \leq i,j \leq s$. Then, for any $x, y \in \Omega$, using (2.17), we get

$$\begin{aligned}
|g_1(y;h) - g_1(x;h)| &= 1\ |y - x| =: L_1|y - x| \\
|g_2(y;h) - g_2(x;h)| &= \Big|y - x + ha_{21}\Big(f\big(t + c_2h, g_1(y;h)\big) - f\big(t + c_2h, g_1(x;h)\big)\Big)\Big| \\
&\leq (1 + ha_{21}L_0)|y - x| = (1 + qc_2)|y - x| =: L_2|y - x| \\
|g_3(y;h) - g_3(x;h)| &\leq \Big(1 + hL_0\big(a_{31} + a_{32}(1 + ha_{21}L_0)\big)\Big)|y - x| \\
&\leq \big(1 + qc_3(1 + qc_2)\big)|y - x| =: L_3|y - x| \\
&\ \vdots \\
|g_s(y;h) - g_s(x;h)| &\leq \Big(1 + qc_s\big(1 + \ldots(1 + qc_2)\big)\ldots\Big)|y - x| =: L_s|y - x|.
\end{aligned}$$

Since $c_i \leq 1$, for all $i = 2, \ldots, s$, we can bound

$$L_i \leq L_s \leq \Big(1 + q\big(1 + \ldots(1 + q)\big)\ldots\Big) = 1 + q + q^2 + \ldots + q^{s-1} = \frac{1 - q^s}{1 - q}$$

Moreover, if $q = hL_0 \leq 1$ we have $L_s \leq s$ and if $q \leq 1/2$ we have $L_s \leq 2$.

Using the Lipschitz conditions for the g_i together with (2.24) and (2.17) we finally get

$$|F_h(t,y) - F_h(t,x)| \leq \sum_{j=1}^{s} b_j L_0 L_j |x - y| \leq L_0 L_s |x - y|.$$

Thus, the increment function F_h admits a Lipschitz condition with constant

$$L := L_0 \frac{1 - (hL_0)^s}{1 - hL_0}$$

for general step size h and with constant $L = 2L_0$ for $h \leq (2L_0)^{-1}$.

In general, i.e. when there exist i, j such that $a_{ij} < 1$, the Lipcschitz constant for g_i has to be changed to $L_i' = \Big(1 + qc_i'\big(1 + \ldots(1 + qc_2')\big)\ldots\Big)$ with $c_i' = \sum_{j=1}^{i-1} |a_{ij}|$ and the Lipschitz constant for F_{h_k} to $L_s' L_0$. □

Corollary 2.3.19. *Let $f(t,y)$ admit a uniform Lipschitz condition on $[0,T] \times \Omega$ with $\{u(t) : t \in [0,T]\} \subset \Omega$ and let $F_{h_k}(.,.)$ be an ERK that is invariant under autonomization. Then consistency of order p implies convergence with order p and*

$$|u_n - y_n| \leq ce^{LT}h^p, \tag{2.25}$$

where L is the Lipschitz constant of F_h from Lemma 2.3.18 and the constant c is independent of h.

Proof. Follows directly from Lemma 2.3.18 and Theorem 2.2.8. □

f LIPSCHITZ ⇒ F_h LIPSCHITZ ⇒ F_h LOCALLY STABLE

CONSISTENCY + STABILITY = CONVERGENCE

2.4 Estimates of the local error and time step control

2.4.1. The analysis in the last section used a crude a priori bound of the local error based on high-order derivatives of the right hand side $f(t,u)$. In the case of a complex nonlinear system, such an estimate is bound to be inefficient, since it involves global bounds on the derivatives. Obviously, the local error cannot be computed exactly either, because that would require or imply the knowledge of the exact solution.

In this section, we discuss two methods that allow an estimate of the truncation error from computed solutions. These estimates are local in nature and therefore usually much sharper. Thus, they can be used to control the step size, which in turn gives good control over the balance of accuracy and effort.

The idea is to use two numerical estimates of u_k that converge with different order, so as to estimate the leading order term of the local error of the lower-order method. Given this estimate for the local error, we can then devise an algorithm for step-size control that guarantees that the local error of a one-step method remains below a threshold ε in every time step.

Nevertheless, in these estimates there is the implicit assumption that the true solution u is sufficiently regular and the step size is sufficiently small, such that the local error already 'follows' the theoretically predicted order.

Algorithm 2.4.2 (Adaptive step size control)**.** Let y_k and $\widetilde{y}_k$ be two approximations of u_k of consistency order p and $\widetilde{p} \ge p+1$, respectively, and let $\varepsilon > 0$ be given.

1. Given y_{k-1}, compute y_k and $\widetilde{y}_k$ with time step size h_k
   (both starting from the value y_{k-1} at t_{k-1}).
2. Compute
$$h_{\text{opt}} = h_k \left(\frac{\varepsilon}{|y_k - \widetilde{y}_k|} \right)^{\frac{1}{p+1}}. \tag{2.26}$$
3. If $|y_k - \widetilde{y}_k| > \varepsilon$, and thus $h_{\text{opt}} < h_k$, the time step is rejected: Choose $h_k = h_{\text{opt}}$ and recompute y_k and $\widetilde{y}_k$.
4. If the time step was accepted, let $h_{k+1} = \min(h_{\text{opt}}, T - t_k)$.
5. Increase k by one and return to Step 1.

Remark 2.4.3. When t_k is close to T, then the choice $h_{k+1} = h_{\text{opt}}$ in Step 4, may lead to $h_{k+2} = T - t_{k+1} \approx \text{eps}$ (machine epsilon) in the following step. To avoid round-off errors, it is advisable to choose $h_{k+1} = T - t_k$ already for $T - t_k \le c h_{\text{opt}}$, where c is a moderate constant of size around 1.1. This way we avoid that $h_{k+2} \approx \text{eps}$.

Remark 2.4.4. This algorithm controls and equilibrates the local error. It does not control the accumulated global error. The global error estimate still retains the exponential term. Global error estimation techniques involve considerably more effort and are beyond the scope of this course.

The algorithm does not provide an estimate for the leading-order term in the local error of $\widetilde{y}_k$. However, since it is a higher order approximation than y_k, we should use $\widetilde{y}_k$ as the approximation of u at t_k and as the initial value for the next time step.

Let us now discuss two techniques to compute higher-order approximations $\widetilde{y}_k$ of u_k.

2.4.1 Extrapolation methods

2.4.5. Here, we estimate the local error by a method called Richardson extrapolation (cf. Numerik 0). It is based on computing two approximations with the same method, but different step size. In particular, we will use an approximation y_2 with two steps of size h and an approximation Y_1 with one step of size $2h$, both starting with the same initial value at t_0.

Theorem 2.4.6: Let y_2 be the approximation of $u_2 = u(t_0 + 2h)$ obtained after two steps of an ERK with step size h and let Y_1 be the approximation of u_2 after one step of the same method with step size $2h$, both starting from u_0 at t_0. If f is sufficiently smooth and the ERK is consistent of order p, then we can define

$$\widetilde{y}_2 = \frac{2^p y_2 - Y_1}{2^p - 1}, \tag{2.27}$$

and have

$$|u_2 - \widetilde{y}_2| = O(h^{p+2}). \tag{2.28}$$

Proof. The exact form of the leading-order term in the local error of an ERK of order p can be obtained by explicitly calculating the leading order term in the Taylor expansion in Lemma 2.3.13, i.e. there exist constant vectors $\zeta_k = \zeta_k\left(f_{k-1}, f'_{k-1}, \dots, f^{(p)}_{k-1}\right) \in \mathbb{R}^d$, for $k = 1, 2$, independent of h such that

$$\eta_k(u) = \zeta_k h^{p+1} + \mathcal{O}(h^{p+2}),$$

where $f^{(j)}_{k-1}$ denotes the jth derivative of f evaluated at (t_{k-1}, y_{k-1}), for $k = 1, 2$.

Moreover, since $t_1 = t_0 + h$ and $y_1 = y_0 + \mathcal{O}(h)$, we can also deduce via Taylor expansion that $f^{(j)}_1 = f^{(j)}_0 + \mathcal{O}(h)$ so that $\zeta_2 = \zeta_1 + \mathcal{O}(h)$ and thus

$$\eta_k(u) = \zeta_1 h^{p+1} + \mathcal{O}(h^{p+2}), \quad k = 1, 2. \tag{2.29}$$

Furthermore, we can also use Taylor series expansion of $F_h(t_1, y_1)$ around (t_1, u_1) to obtain

$$\begin{aligned} F_h(t_1, y_1) &= F_h(t_1, u_1) + \nabla_y F_h(t_1, u_1)\eta_1(u) + \mathcal{O}\left(|\eta_1(u)|^2\right) \\ &= F_h(t_1, u_1) + h^{p+1}\nabla_y F_h(t_1, u_1)\zeta_1 + \mathcal{O}(h^{p+2}). \end{aligned} \tag{2.30}$$

Thus, following the same proof technique as in Theorem 2.2.6, we obtain for the error after two steps of size h,

$$\begin{aligned} u_2 - y_2 &= \underbrace{u_0 - y_0}_{=0} + \sum_{k=1}^{2}\Big[\eta_k(u) - \underbrace{\eta_k(y)}_{=0} + hF_h(t_{k-1}, u_{k-1}) - hF_h(t_{k-1}, y_{k-1})\Big] \qquad (2.31) \\ &= \sum_{k=1}^{2}\eta_k(u) + h\left[F_h(t_1, u_1) - F_h(t_1, y_1)\right] \\ &= 2\zeta_1 h^{p+1} + \mathcal{O}(h^{p+2}) + h\left[h^{p+1}\nabla_y F_h(t_1, u_1)\zeta_1 + \mathcal{O}(h^{p+2})\right] = 2\zeta_1 h^{p+1} + \mathcal{O}(h^{p+2}) \end{aligned}$$

On the other hand,

$$u_2 - Y_1 = \zeta_1(2h)^{p+1} + \mathcal{O}(h^{p+2}) = 2^{p+1}\zeta_1 h^{p+1} + \mathcal{O}(h^{p+2}). \tag{2.32}$$

Taking 2^p times equation (2.31) and subtracting equation (2.32), we can eliminate the leading order term and obtain

$$\mathcal{O}(h^{p+2}) = 2^p(u_2 - y_2) - (u_2 - Y_1) = (2^p - 1)u_2 - (2^P y_2 - Y_1) = (2^p - 1)(u_2 - \widetilde{y}_2)$$

which completes the proof. □

Remark 2.4.7. Thus, $\widetilde{y}_2$ provides an approximation of u_2 with consistency order $p+1 > p$ and can be used in Algorithm 2.4.2 above to control the step size in each step. In particular, $\widetilde{y}_{k+1}$ can be computed cheaply from y_{k+1} and Y_k via formula (2.27) (with index $k+1$ instead of 2). As mentioned in Remark 2.4.4, in practice we expect a better global accuracy, if we use $\widetilde{y}_{k-1}$ instead of y_{k-1} as the initial value at t_{k-1} for computing y_{k+1} and Y_k.

However, in general the computation of Y_1 requires $s-1$ additional evaluations of f, since the stage values will differ from those of y_1 and y_2, leading to a total of $3s-1$ function evaluations for two time steps for this $p+1$-order method. An alternative, that uses the optimal number of stage values s for a $p+1$-order method and reuses all stage values of the lower-order method will be discussed in the next section.

2.4.2 Embedded Runge-Kutta methods

Instead of estimating the local error by doubling the step size, embedded Runge-Kutta methods use two methods of different order to achieve the same effect. The key to efficiency is here, that the computed stages g_i are the same for both methods, and only the quadrature weights b_i differ.

Definition 2.4.8 (Embedded Runge-Kutta methods)**.** An embedded s-stage Runge-Kutta method with orders of consistence p and $\widetilde{p}$ computes two approximations y_k and $\widetilde{y}_k$ of u_k with the same function evaluations. For this purpose, the stage values g_i and k_i at $t_0 + c_i h_k$ are identical for all $i = 1, \dots, s$, i.e. both methods have the same coefficients a_{ij} and c_i. To compute the final approximations at time step t_k, we use two different quadrature rules, i.e.

$$\begin{aligned} y_k &= y_{k-1} + h_k \sum b_i k_i \,, \\ \widetilde{y}_k &= y_{k-1} + h_k \sum \widetilde{b}_i k_i \,, \end{aligned} \tag{2.33}$$

such that y_k and $\widetilde{y}_k$ are consistent of order p and $\widetilde{p} > p$, respectively. Typically, $\widetilde{p} = p + 1$.

Definition 2.4.9: The Butcher tableau for the embedded method has the form:

$$\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & \vdots & \ddots & & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\
\hline
 & \widetilde{b}_1 & \widetilde{b}_2 & \cdots & \widetilde{b}_{s-1} & \widetilde{b}_s \\
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array}$$

Remark 2.4.10. For higher order methods or functions $f(t,u)$ with complicated evaluation, most of the work lies in the computation of the stages. Thus, the additional quadrature for the computation of y_k is almost for free. Nevertheless, due to the different orders of approximation, $\widetilde{y}_k$ is much more accurate and we obtain

$$u_k - y_k = \widetilde{y}_k - y_k + \mathcal{O}(h^p). \tag{2.34}$$

Thus, $\widetilde{y}_k - y_k$ is a good estimate for the local error in y_k. This is the error which is used in step size control below. However, as in the Richardson extrapolation above, we use the more accurate value $\widetilde{y}_k$ as the final approximation at t_k and as the initial value for the next time step, even if we do not have a computable estimate for its local error.

Definition 2.4.11 (Dormand-Prince 45): The embedded Runge-Kutta method of orders 5 for $\widetilde{y}_k$ and 4 for y_k due to Dormand and Prince has the Butcher tableau

$$\begin{array}{c|ccccccc}
0 & & & & & & & \\
1/5 & 1/5 & & & & & & \\
3/10 & 3/40 & 9/40 & & & & & \\
4/5 & 44/45 & -56/15 & 32/9 & & & & \\
8/9 & \frac{19372}{6561} & \frac{-25360}{2187} & \frac{64448}{6561} & \frac{-212}{729} & & & \\
1 & \frac{9017}{3168} & \frac{-355}{33} & \frac{46732}{5247} & \frac{49}{176} & \frac{-5103}{18656} & & \\
1 & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & \frac{-2187}{6784} & \frac{11}{84} & \\
\hline
\widetilde{y}_k & \frac{35}{384} & 0 & \frac{500}{1113} & \frac{125}{192} & \frac{-2187}{6784} & \frac{11}{84} & 0 \\
y_k & \frac{5179}{57600} & 0 & \frac{7571}{16695} & \frac{393}{640} & \frac{-92097}{339200} & \frac{187}{2100} & \frac{1}{40}
\end{array}$$

It has become a standard tool for the integration of IVP and it is the backbone of `ode45` in Matlab.

Chapter 3

Implicit One-Step Methods and Long-Term Stability

In the first chapter, we studied methods for the solution of IVP and the analysis of their convergence with shrinking step size h. We could gain a priori error estimates from consistency and stability for sufficiently small h.

All of those error estimates are based on Grönwall's inequality and contain a term of the form e^{LT}. This increases fast with increasing length of the time interval $[0, T]$ and thus, the analysis is unsuitable for the study of long-term integration. The exponential term will eventually outweigh any term of the form h^p.

On the other hand, for instance our solar system has been moving on stable orbits for several billion years and we do not observe an exponential increase of velocities. Thus, there are in fact applications for which the simulation of long time periods is worthwhile and where exponential growth of the discrete solution would be extremely disturbing.

This chapter first studies conditions on differential equations with bounded long term solutions, and then discusses numerical methods mimicking this behavior.

3.1 Monotonic initial value problems

Example 3.1.1. We consider for $\lambda \in \mathbb{C}$ the (scalar) linear initial value problem

$$\begin{aligned} u' &= \lambda u \\ u(0) &= 1. \end{aligned} \tag{3.1}$$

Splitting $\lambda = \operatorname{Re}(\lambda) + i \operatorname{Im}(\lambda)$ into its real and imaginary part, the (complex valued) solution to this problem is

$$u(t) = e^{\lambda t} = e^{\operatorname{Re}(\lambda)t}\big(\cos(\operatorname{Im}(\lambda)t) + i\,\sin(\operatorname{Im}(\lambda)t)\big).$$

The behavior of $u(t)$ for $t \to \infty$ is determined by the real part of λ:

$$\begin{aligned} \operatorname{Re}(\lambda) < 0: &\qquad u(t) \to 0 \\ \operatorname{Re}(\lambda) = 0: &\qquad |u(t)| = 1 \\ \operatorname{Re}(\lambda) > 0: &\qquad u(t) \to \infty \end{aligned} \tag{3.2}$$

Moreover, the solution is bounded for λ with non-positive real part for all points in time t.

Remark 3.1.2. Since we deal in the following again and again with eigenvalues of real-valued matrices and these eigenvalues can be complex, we will always consider complex valued IVP hereafter.

Remark 3.1.3. Due to Grönwall's inequality and the stability Theorem 1.4.5, the solution to the IVP above admits the estimate $|u(t)| \le e^{|\lambda| t}|u(0)|$. This is seen easily by applying the comparison function $v(t) \equiv 0$. As soon as $\lambda \neq 0$ has a non-positive real part, this estimate is still correct but very pessimistic and therefore useless for large t. Since problems with bounded long-term behavior are quite important in applications, we will have to introduce an improved notation of stability.

Definition 3.1.4: The function $f(t,y)$ satisfies on its domain $D \subset \mathbb{R} \times \mathbb{C}^d$ a **one-sided Lipschitz condition** if the inequality

$$\mathrm{Re}\langle f(t,y) - f(t,x), y - x\rangle \le \nu |y-x|^2 \tag{3.3}$$

holds with a constant $\nu \in \mathbb{R}$ for all $(t,x), (t,y) \in D$. Moreover, such a function is called **monotonic** if $\nu = 0$, thus

$$\mathrm{Re}\langle f(t,y) - f(t,x), y - x\rangle \le 0. \tag{3.4}$$

An ODE $u' = f(u)$ is called monotonic if its right hand side f is monotonic.

Remark 3.1.5. The term monotonic from the previous definition is consistent with the term *monotonically decreasing*, which we know from scalar, real-valued functions. We can see this by observing that, for $f : \mathbb{R} \to \mathbb{R}$ and $y > x$:

$$\big(f(t,y) - f(t,x)\big)(y-x) \le 0 \quad \Leftrightarrow \quad f(t,y) - f(t,x) \le 0.$$

Theorem 3.1.6: Let $u(t)$ and $v(t)$ be the solutions of the equations

$$u' = f(t,u) \quad \text{and} \quad v' = f(t,v)$$

with initial values $u(t_0) = u_0$ and $v(t_0) = v_0$ in $\mathbb{C}^d$, respectively. Let the function $f : \mathbb{R} \times \mathbb{C}^d \to \mathbb{C}^d$ be continuous and let the one-sided Lipschitz condition (3.3) hold. Then we have for $t > t_0$:

$$|v(t) - u(t)| \le e^{\nu(t-t_0)}|v(t_0) - u(t_0)|. \tag{3.5}$$

Proof. We consider the auxiliary function $m(t) = |v(t) - u(t)|^2$ and its derivative

$$\begin{aligned} m'(t) &= 2\mathrm{Re}\big\langle v'(t) - u'(t), v(t) - u(t)\big\rangle \\ &= 2\mathrm{Re}\big\langle f\big(t, v(t)\big) - f\big(t, u(t)\big), v(t) - u(t)\big\rangle \\ &\le 2\nu |v(t) - u(t)|^2 \\ &= 2\nu m(t). \end{aligned}$$

According to Grönwall's inequality (Lemma 1.3.8 on page 11) we obtain for $t > t_0$:

$$m(t) \le m(t_0) e^{2\nu(t-t_0)}.$$

Taking the square root yields the stability estimate (3.5). □

Remark 3.1.7. As in example 3.1.1, we obtain from the stability estimate, that for the difference of two solutions $u(t)$ and $v(t)$ of the differential equation $u' = f(t,u)$ (with different initial conditions) we obtain in the limit $t \to \infty$:

$$\begin{aligned} \nu < 0: &\qquad |v(t) - u(t)| \to 0 \\ \nu = 0: &\qquad |v(t) - u(t)| \le |v(t_0) - u(t_0)| \end{aligned} \tag{3.6}$$

Lemma 3.1.8: Let $A(t) \in \mathbb{C}^{d\times d}$ be a diagonalizable matrix function with eigenvalues $\lambda_j(t)$, $j = 1, \dots, d$. Then the linear function $f(t,y) := A(t)y$ admits the one-sided Lipschitz condition (3.3) on all of $\mathbb{R} \times \mathbb{C}^d$ with the constant

$$\nu = \max_{\substack{j=1,\dots,d \\ t\in\mathbb{R}}} \operatorname{Re}(\lambda_j(t)).$$

Furthermore, the linear differential equation $u' = Au$ with $u(t) \in \mathbb{C}^d$ is monotonic if and only if

$$\operatorname{Re}(\lambda_j(t)) \le 0, \quad \text{for all } t \in \mathbb{R}. \tag{3.7}$$

(This is the vector-valued form of Example 3.1.1.)

Proof. For the right hand side of the ODE, we have

$$\operatorname{Re}\langle A(t)y - A(t)x, y - x\rangle = \operatorname{Re}\frac{\langle A(t)(y-x), y-x\rangle}{|y-x|^2}|y-x|^2 \le \max_{j=1,\dots,d} \operatorname{Re}(\lambda_j(t))|y-x|^2.$$

This shows that $\nu \le \max_{j=1,\dots,d;\, t\in\mathbb{R}} \operatorname{Re}(\lambda_j(t))$. If we now insert for $y - x$ an eigenvector of eigenvalue $\lambda_j(t)$ for which the maximum is attained, then we obtain the equality and therefore $\nu = \max_{j=1,\dots,d;\, t\in\mathbb{R}} \operatorname{Re}(\lambda_j)$. □

3.1.1 Stiff initial value problems

Example 3.1.9. We consider the IVP

$$u' = Au \quad \text{with} \quad A := \begin{pmatrix} -21 & 19 & -20 \\ 19 & -21 & 20 \\ 40 & -40 & -40 \end{pmatrix} \quad \text{and} \quad u(0) = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}. \tag{3.8}$$

The eigenvalues of A are $\lambda_1 = -2$ and $\lambda_{2,3} = -40 \pm 40i$. The exact solution is

$$u(t) = \begin{pmatrix} \frac{1}{2}e^{-2t} + \frac{1}{2}e^{-40t}[\cos 40t + \sin 40t] \\ \frac{1}{2}e^{-2t} - \frac{1}{2}e^{-40t}[\cos 40t + \sin 40t] \\ -e^{-40t}[\cos 40t - \sin 40t] \end{pmatrix} \to 0 \quad \text{as} \quad t \to \infty.$$

For small time $0 \le t \le 0.2$ all three components are changing rapidly due to the trigonometric terms, since the factor e^{-40t} in front of them is still fairly big. Thus, it is necessary to choose small time step sizes $h \ll 1$.

For $t > 0.2$, we have $u_3 \approx 0$ and $u_1 \approx u_2$, and both those components change fairly slowly, so we could choose a larger time step size $h \geq 0.1$.

However, if we consider the explicit Euler method applied to (3.8) we get

$$y^{(n)} = y^{(n-1)} + hAy^{(n-1)}$$

and thus

$$y^{(n)} = (I + hA)^n \, u_0 \, .$$

Now, if we choose a time step size of $h = 0.01$ the matrix $I + hA$ has eigenvalues $\mu_j = 0.98, 0.6 + 0.4i, 0.6 - 0.4i$, so that

$$|y^{(n)}| = |(I + hA)^n \, u_0| \leq \|I + hA\|^n |u_0| = 0.98^n \sqrt{2} \;\to\; 0 \quad \text{as} \quad t \to \infty,$$

which is, at least qualitatively, the correct behaviour.

For $h = 0.1$, $I + hA$ has eigenvalues $\mu_j = 0.8, -3+4i, -3-4i$. It is easy to see that the first eigenvector is $v_1 = \frac{1}{\sqrt{2}}(1, 1, 0)^T$; the other two eigenvectors are orthogonal to v_1. Thus, if we apply $(I + hA)^n$ to the second or third eigenvector, v_2 or v_3, we get

$$|(I + hA)^n v_j| = |-3 \pm 4i|^n |v_j| = 5^n \;\to\; \infty \quad \text{as} \quad n \to \infty, \quad \text{for} \;\; j = 2, 3.$$

Since u_0 contains components in the direction of v_2 and v_3, this means that $|y^{(n)}| \to \infty$ as $n \to \infty$, very much in contrast to the behaviour of the exact solution $u(t) \to 0$ for $t \to \infty$.

So, even when $u_3 \approx 0$ and $u_2 - u_1 \approx 0$ and the perturbations are very small, the instability of the explicit Euler method with time step size $h = 0.1$ will lead to an exponential increase in these perturbations.

Remark 3.1.10. The important message here is that from a point of view of approximation error (or consistency), it would be possible to increase the time step significantly at later times, but due to stability problems with the explicit Euler method we cannot increase h beyond a certain stability threshold.

This phenomenon only arises for monotonic ODEs, or for ODEs that satisfy a one-sided Lipschitz condition with constant $0 < \nu \ll L$ and that are monotonic for all $t \geq t^*$, for some $t^* \geq t_0$. The consistency error is closely linked to the Lipschitz constant L of f, while the stability is linked to the ratio of L and the constant ν in the one-sided Lipschitz condition. In the following definition, we will only focus on monotonic IVPs.

Definition 3.1.11: Let f be Lipschitz continuous with constant $L > 0$ and one-sided Lipschitz continuous with constant $\nu \in \mathbb{R}$. An initial value problem is called **stiff**, if it has the following characteristic properties:

1. The right hand side of the ODE is monotonic.
2. The time scales on which different solution components are evolving differ a lot, i.e.,
$$L \gg |\nu| \, .$$
3. The time scales which are of interest for the application are much longer than the fastest time scales of the equation, i.e.,
$$e^{\nu T} \gg e^{-LT} \approx 0 \, . \tag{3.9}$$

Remark 3.1.12. Note that for the linear IVP in Lemma 3.1.8 and when f is monotonic, we have

$$L := \max_{\substack{j=1,\dots,d\\ t\in\mathbb{R}}} |\lambda_j(t)| \geq \max_{\substack{j=1,\dots,d\\ t\in\mathbb{R}}} |\mathrm{Re}(\lambda_j(t))| \quad \text{and} \quad |\nu| := \min_{\substack{j=1,\dots,d\\ t\in\mathbb{R}}} |\mathrm{Re}(\lambda_j(t))| .$$

Remark 3.1.13. Even though we used the term definition, the notion of "stiffness of an IVP" has something vague or even inaccurate about it. In fact that is due to the very nature of the problems and cannot be fixed. Instead we are forced to sharpen our understanding by means of a few examples.

Example 3.1.14. First of all we will have a look at equation (3.8) in Example 3.1.9. Studying the eigenvalues of the matrix A, we clearly see that $\nu = -2$ and thus the problem is monotonic. We can also find that the Lipschitz constant is $L = \|A\| \approx 72.5$ so that the second condition holds as well.

According to the discussion of example 3.1.9, the third condition depends on the purpose of the computation. If we want to compute the solution at time $T = 0.01$, we would not denote the problem as stiff. On the other hand, if one is interested on the solution at time $T \geq 1$, on which the terms containing e^{-40T} are already below typical machine accuracy, the problem is stiff indeed. Here, we have seen that Euler's method requires disproportionately small time steps.

Remark 3.1.15. The definition of stiffness and the discussion of the examples reveal that numerical methods are needed, which are not just convergent for time steps $h \to 0$ but also for fixed step size h, even in the presence of time scales clearly below h. In this case, methods still have to produce solutions with correct limit behavior for $t \to \infty$.

Example 3.1.16. The **implicit Euler method** is defined by the one-step formula

$$y_1 = y_0 + hf(t_1, y_1) \quad \Leftrightarrow \quad y_1 - hf(t_1, y_1) = y_0 , \tag{3.10}$$

which in general involves solving a nonlinear system of equations. Applied to our linear example (3.8), we get

$$y^{(n)} = (I - hA)^{-1} y^{(n-1)} \quad \Rightarrow \quad y^{(n)} = (I - hA)^{-n} u_0$$

For all $h > 0$, the real part of the eigenvalues of the matrix $I - hA$ is

$$\mathrm{Re}(\mu_j) = \frac{1}{1+2h}, \frac{1}{1+40h}, \frac{1}{1+40h} ,$$

which are all strictly less than 1, such that we get

$$|y^{(n)}| \to 0 \quad \text{as} \quad n \to \infty,$$

independently of h. Thus, although the implicit Euler method requires in general the solution of a nonlinear system in each step, it allows for much larger time steps than the explicit Euler method, when applied to a stiff problem.

For a visualization see the programming exercise on the last problem sheet and the appendix.

3.2 A-, B- and L-stability

3.2.1. In this section, we will investigate desirable properties of one-step methods for stiff IVP (3.11). We will first study linear problems of the form

$$u' = Au \qquad u(t_0) = u_0. \tag{3.11}$$

and the related notion of A-stability in detail. From the conditions for stiffness we derive the following problem characteristics:

1. All eigenvalues of the matrix A lie in the left half-plane of the complex plane. With (3.2) all solutions are bounded for $t \to \infty$.
2. There are eigenvalues close to zero and eigenvalues with a large negative real part.
3. We are interested in time spans which make it necessary, that the product $h\lambda$ is allowed to be large, for an arbitrary eigenvalue and an arbitrary time step size.

For this case we now want to derive criteria for the boundedness of the discrete solution for $t \to \infty$. The important part is not to derive an estimate holding for $h \to 0$, but one that holds for any value of $h\lambda$ in the left half-plane of the complex numbers.

Definition 3.2.2: Consider the (general) one-step method

$$y_1 = y_0 + hF_h(t_0, y_0, y_1),$$

applied to the scalar, linear test problem $u'(t) = \lambda u(t)$. Then

$$y_1 = R(h\lambda)u_0, \tag{3.12}$$

and

$$y^{(n)} = R(h\lambda)^n u_0, \tag{3.13}$$

for some function $R : \mathbb{C} \to \mathbb{C}$, which is denoted the **stability function** of the one-step method F_h. The **stability region** of the one-step method is the set

$$S = \big\{z \in \mathbb{C} \big| |R(z)| \le 1\big\}. \tag{3.14}$$

Example 3.2.3 (explicit Euler)**.**

$$\begin{aligned} &y_1 = y_0 + h\lambda y_0 = (1 + h\lambda)y_0 \\ \Rightarrow \quad &R(z) = 1 + z \end{aligned} \tag{3.15}$$

The stability region for the explicit Euler is a circle with radius 1 and centre (-1,0) in the complex plane (see Figure 3.1 left).

Example 3.2.4 (Implicit Euler)**.**

$$\begin{aligned} &y_1 = y_0 + h\lambda y_1 \quad \Leftrightarrow (1 - h\lambda)y_1 = y_0 \\ \Rightarrow \quad &R(z) = \frac{1}{1 - z} \end{aligned} \tag{3.16}$$

The stability region for the implicit Euler is the complement of a circle with radius 1 and centre (1,0) in the complex plane (see Figure 3.1 right).

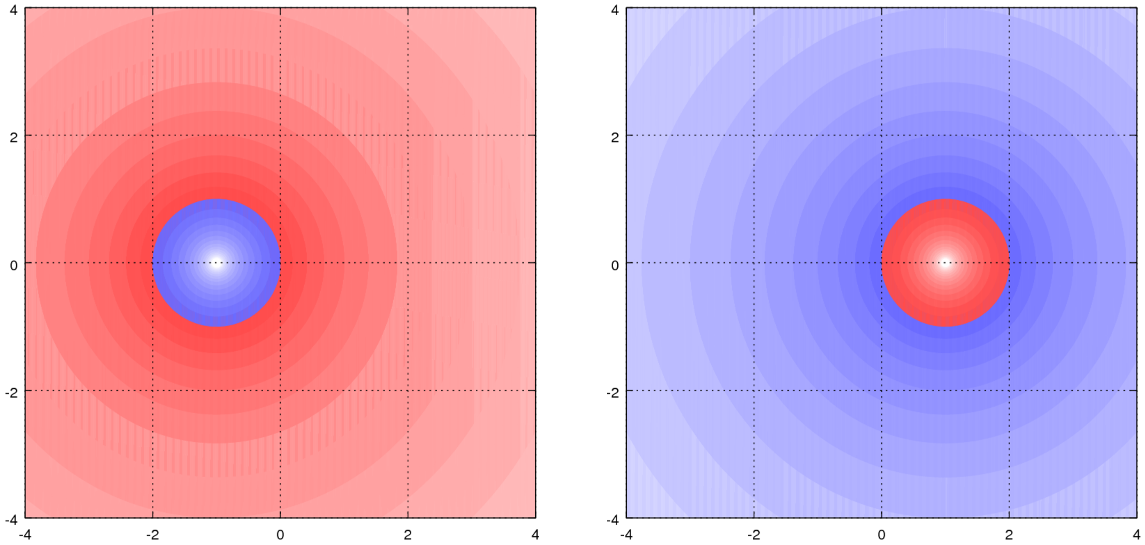


Figure 3.1: Stability regions for explicit and implicit Euler (blue stable, red unstable)

Definition 3.2.5 (A-stability): A method is called **A-stable**, if its stability region contains the left half-plane of $\mathbb{C}$. Hence,

$$\{z \in \mathbb{C} \mid \operatorname{Re}(z) \le 0\} \subset S \tag{3.17}$$

Theorem 3.2.6: Consider the linear, autonomous IVP

$$u' = Au, \qquad u(t_0) = u_0$$

with a diagonalizable matrix A and initial value $y^{(0)} = u_0$. The stability of a one-step method with stability region S applied to this vector-valued problem is inherited from the scalar equation.
In particular, let $\left(y^{(k)}\right)_{k=0}^{\infty}$ be the sequence of approximations, generated by an A-stable one-step method with step size h for this IVP. If all eigenvalues of A have a non-positive real part, then the sequence is uniformly bounded for all h.

Remark 3.2.7. The term “A-stability” was deliberately chosen neutrally by Dahlquist. In particaluar, note that A-stability does **not** stand for asymptotic stability.

Proof. Since A is diagonalizable, there exists an invertible matrix $V \in \mathbb{C}^{d\times d}$ and a diagonal matrix $\Lambda \in \mathbb{C}^{d\times d}$ such that $A = V^{-1}\Lambda V$. Let $w := Vu$. Then

$$w' = (Vu)' = Vu' = VAu = \Lambda V u = \Lambda w \tag{3.18}$$

and $w_0 = w(t_0) = Vu(t_0)$, and so the system of ODEs decouples into d independent ODEs

$$w_\ell' = \lambda_\ell w_\ell, \quad w_\ell(t_0) = (w_0)_\ell\,, \quad \ell = 1, \dots, d.$$

Similarly, the stage values of a Runge-Kutta (RK) method decouple into d independent, decoupled components:

$$
\begin{aligned}
g_i = y_0 + h\sum_{j=1}^{s} a_{ij} V^{-1}\Lambda V g_j \quad &\Rightarrow \\
\gamma_i := V g_i = V y_0 + h\sum_{j=1}^{s} a_{ij}\Lambda V g_j &= w_0 + h\sum_{j=1}^{s} a_{ij}\Lambda\gamma_j
\end{aligned}
$$

$$
\text{or equivalently} \qquad (\gamma_i)_\ell = (w_0)_\ell + h\sum_{j=1}^{s} a_{ij}\lambda_\ell(\gamma_j)_\ell\,, \quad \ell = 1,\dots,d.
$$

Finally, if we denote by $\eta_j := V y_j$ the transformed numerical solution at the jth time step, we get for the next iterate

$$
\eta_1 = V y_1 = V y_0 + h\sum_{i=1}^{s} b_i V g_i = \eta_0 + h\sum_{i=1}^{s} b_i\gamma_i
$$

Thus, the RK applied to a vector valued problem decouples into d decoupled scalar problems solved by the same RK. But for each of the scalar problems, the definition of A-stability implies boundedness of the solution, if $\mathrm{Re}(\lambda_\ell) \le 0$ for all $\ell = 1,\dots,d$, and thus

$$
|y^{(k)}| = |V\eta^{(k)}| \le \|V\||\eta^{(k)}| < \infty.
$$

□

Theorem 3.2.8: No explicit Runge-Kutta method is A-stable.

Proof. We show that for such methods $R(z)$ is a polynomial. It is known for polynomials that the absolute value of their value goes to infinity, if the absolute value of the argument goes to infinity. Thus, there exists $z \in \{z \in \mathbb{C} \mid \mathrm{Re}(z) \le 0\}$ such that $|R(z)| > 1$ and thus $z \notin S$ which implies the result of the theorem.

Consider an arbitrary ERK applied to the scalar problem $u' = \lambda u$, $u(t_0) = u_0$. From equation (2.12b) it follows that $k_i = \lambda g_i$, for all $i = 1,\dots,s$. If we insert that into the equation (2.12a), we obtain

$$
g_i = y_0 + h\lambda\sum_{j=1}^{i-1} a_{ij} g_j.
$$

With $g_1 = y_0$ and $z = h\lambda$ one has

$$
\begin{aligned}
g_2 &= y_0 + a_{21} z y_0 = (1 + a_{21} z) y_0 \\
g_3 &= y_0 + a_{31} z g_1 + a_{32} z g_2 = y_0 + a_{31} z y_0 + a_{32} z (1 + a_{21} z) y_0 \\
&= (1 + a_{31} z + a_{32} z (1 + a_{21} z)) y_0.
\end{aligned}
$$

Therefore, one shows easily per induction that g_j is a polynomial of order $j - 1$ in z. Substituting into formula (2.12c) it follows that $R(z)$ is a polynomial of order $s - 1$. □

Remark 3.2.9. The notion of A-stability is only applicable to linear problems with diagonalizable matrices. Now we are considering its extension to nonlinear problems with monotonic right hand sides.

Definition 3.2.10: A one-step method applied to a monotonic initial value problem $u' = f(t, u)$ with arbitrary initial values y_0 and $\tilde{y}_0$ is called **B-stable** if

$$|y_1 - \tilde{y}_1| \le |y_0 - \tilde{y}_0| \tag{3.19}$$

independent of the time step size h.

Theorem 3.2.11: Consider the IVP

$$u' = f(t, u) \quad \text{with} \quad u(t_0) = u_0$$

with f monotonic and such that $f(t, 0) = 0$, for all $t \in \mathbb{R}$. Let $\left(y^{(k)}\right)_{k=0}^{\infty}$ be the sequence generated by a B-stable one-step method F_h with initial value $y^{(0)} = u_0$ that satisfies $F_h(t, 0) = 0$, for all $t \in \mathbb{R}$. Then the sequence is uniformly bounded for $k \to \infty$ independent of the time step size h.

Proof. The theorem follows immediately by setting $\tilde{y}_0 = 0$ and iterating over the definition of B-stability, since the assumptions of the theorem guarantee that $\tilde{y}_k = 0$, for all k. □

Note that $f(t, 0) = 0$ implies $F_h(t, 0) = 0$ for all Runge-Kutta methods.

Corollary 3.2.12: Any B-stable method is A-stable.

Proof. Apply the method to the scalar, linear problem $u' = \lambda u$, which is monotonic for $\mathrm{Re}(\lambda) \le 0$. Now, the definition of B-stability implies $|R(z)| \le 1$, and thus, the method is A-stable. □

An undesirable feature of complex differentiable functions in the context of stability of Runge-Kutta methods is the fact, that $\lim_{z\to\infty} R(z)$ is well-defined on the Riemann sphere, independent of the path chosen to approach this limit in the complex plane. Thus, for any real number x, we have

$$\lim_{x\to\infty} R(x) = \lim_{x\to\infty} R(ix). \tag{3.20}$$

Thus, a method, which has exactly the left half-plane of $\mathbb{C}$ as its stability domain, seemingly a desirable property, has the undesirable property that components in eigenspaces corresponding to very large negative eigenvalues, and thus decaying very fast in the continuous problem, are decaying very slowly if such a method is applied.

This gave rise to the following notion of L-stability. However, note that L-stable methods are not always better than A-stable ones. Similarly, it is also not always necessary to require A-stability. Judgment must be applied according to the problem being solved.

Definition 3.2.13: An A-stable one-step method is called **L-stable**, if

$$\lim_{\mathrm{Re}(z)\to-\infty} |R(z)| = 0. \tag{3.21}$$

Some authors refer to L-stable methods as **strongly A-stable**.

3.3 General Runge-Kutta methods

3.3.1. According to theorem 3.2.8, an explicit Runge-Kutta method cannot be A- or B-stable. Thus, they are not suitable for long term integration of stiff IVPs. The goal of this chapter is the study of methods not suffering from this limitation. The cure will be implicit methods, where stages may not only depend on known values from the past, but also on the value to be computed and future stage values.

We point out immediately that the main drawback of these methods is the fact that they typically require the solution of nonlinear systems of equations and thus involve much higher computational effort. Therefore, careful judgment should be applied to determine whether a problem is really stiff or whether it is better to use an explicit method.

Definition 3.3.2: A (general) **Runge-Kutta method** is a one-step method of the form

$$g_i = y_0 + h\sum_{j=1}^{s} a_{ij}k_j \qquad i = 1,\dots,s \tag{3.22a}$$

$$k_i = f(t_0 + hc_i, g_i) \qquad i = 1,\dots,s \tag{3.22b}$$

$$y_1 = y_0 + h\sum_{i=1}^{s} b_i k_i \tag{3.22c}$$

where $a_{ij} \neq 0$ for all i, j in general. The method is called

Explicit (ERK) if $a_{ij} = 0$, for all $j \geq i$,

Diagonal Implicit (DIRK) if $a_{ij} = 0$, for all $j > i$,

Singly Diagonal Implicit (SDIRK) if DIRK and $a_{11} = a_{22} = \ldots = a_{ss}$,

Implicit (IRK) in all other cases.

Remark 3.3.3. The corresponding Butcher tableaus are

$$\begin{array}{c|cccc}
0 & a_{11} & a_{12} & \dots & a_{1s} \\
c_2 & a_{21} & a_{22} & \dots & a_{2s} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
c_s & a_{s1} & \cdots & a_{s,s-1} & a_{s,s} \\
\hline
 & b_1 & \cdots & b_{s-1} & b_s
\end{array}
\qquad
\begin{array}{c|cccc}
0 & a_{11} & & & \\
c_2 & a_{21} & a_{22} & & \\
\vdots & \vdots & \ddots & \ddots & \\
c_s & a_{s1} & \cdots & a_{s,s-1} & a_{s,s} \\
\hline
 & b_1 & \cdots & b_{s-1} & b_s
\end{array}
\qquad
\begin{array}{c|cccc}
0 & a_{11} & & & \\
c_2 & a_{21} & a_{11} & & \\
\vdots & \vdots & \ddots & \ddots & \\
c_s & a_{s1} & \cdots & a_{s,s-1} & a_{11} \\
\hline
 & b_1 & \cdots & b_{s-1} & b_s
\end{array}$$

IRK **DIRK** **SDIRK**

Example 3.3.4 (Two-stage SDIRK)**.** The following two SDIRK methodsare of order three:

$$\begin{array}{c|cc}
\frac12 - \frac{\sqrt3}{6} & \frac12 - \frac{\sqrt3}{6} & 0 \\
\frac12 + \frac{\sqrt3}{6} & \frac{\sqrt3}{3} & \frac12 - \frac{\sqrt3}{6} \\
\hline
 & \frac12 & \frac12
\end{array}
\qquad
\begin{array}{c|cc}
\frac12 + \frac{\sqrt3}{6} & \frac12 + \frac{\sqrt3}{6} & 0 \\
\frac12 - \frac{\sqrt3}{6} & -\frac{\sqrt3}{3} & \frac12 + \frac{\sqrt3}{6} \\
\hline
 & \frac12 & \frac12
\end{array} \tag{3.23}$$

Lemma 3.3.5: Let $\mathbb{I}$ be the $s \times s$ identity matrix and let $e := (1, \dots, 1)^T \in \mathbb{R}^s$. The stability function of a (general) s-stage Runge-Kutta method with coefficients

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1s} \\ \vdots & & \vdots \\ a_{s1} & \cdots & a_{ss} \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_s \end{pmatrix}$$

is given by the two expressions

$$R(z) \;=\; 1 + zb^T\bigl(\mathbb{I} - zA\bigr)^{-1}e \;=\; \frac{\det\bigl(\mathbb{I} - zA + zbe^T\bigr)}{\det\bigl(\mathbb{I} - zA\bigr)} \tag{3.24}$$

Proof. Applying the method to the scalar test problem with $f(u) = \lambda u$, the definition of the stages g_i leads to the system of linear equations

$$g_i = y_0 + h\sum_{j=1}^{s} a_{ij}\lambda g_j, \quad i = 1, \dots, s.$$

In matrix notation, with $z = h\lambda$, we obtain $(\mathbb{I} - zA)g = (y_0, \dots, y_0)^T$, where g is the vector $(g_1, \dots, g_s)^T$. Equally, we obtain

$$\begin{aligned} R(z)y_0 = y_1 &= y_0 + h\sum_{i=1}^{s} b_i\lambda g_i \\ &= y_0 + zb^T g \\ &= y_0 + zb^T(\mathbb{I} - zA)^{-1}\begin{pmatrix} y_0 \\ \vdots \\ y_0 \end{pmatrix} = \bigl(1 + zb^T(\mathbb{I} - zA)^{-1}e\bigr)\,y_0. \end{aligned}$$

In order to prove the second representation, we write the whole Runge-Kutta method as a single system of equations of dimension $s + 1$:

$$\begin{pmatrix} \mathbb{I} - zA & 0 \\ -zb^T & 1 \end{pmatrix}\begin{pmatrix} g \\ y_1 \end{pmatrix} = y_0\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Applying Cramer's rule yields the result. $\boxed{\text{DIY}}$ □

Example 3.3.6: Stability functions of the modified Euler method, of the classical Runge-Kutta method of order 4 and of the Dormand-Prince method of order 5 are

$$\begin{aligned} R_2(z) &= 1 + z + \tfrac{z^2}{2} \\ R_4(z) &= 1 + z + \tfrac{z^2}{2} + \tfrac{z^3}{6} + \tfrac{z^4}{24} \\ R_5(z) &= 1 + z + \tfrac{z^2}{2} + \tfrac{z^3}{6} + \tfrac{z^4}{24} + \tfrac{z^5}{120} + \tfrac{z^6}{600} \end{aligned}$$

respectively. $\boxed{\text{DIY}}$ Their stability regions are shown in Figure 3.2.

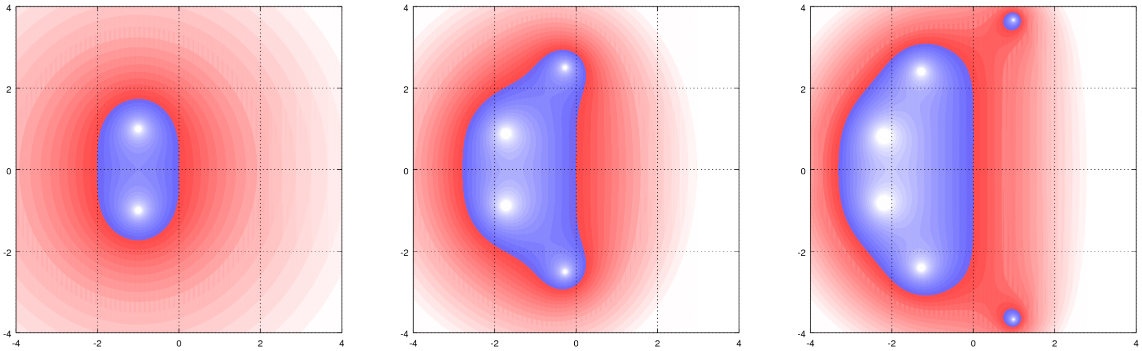


Figure 3.2: Stability regions of the modified Euler method, the classical Runge-Kutta method of order 4 and the Dormand/Prince method of order 5 (blue stable, red unstable)

Definition 3.3.7: The ϑ-scheme is the one-step method, defined for $\vartheta \in [0,1]$ by

$$y_1 = y_0 + h\big((1-\vartheta)f(y_0) + \vartheta f(y_1)\big). \tag{3.25}$$

It is an RKM with the Butcher Tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1-\vartheta & \vartheta \\ \hline & 1-\vartheta & \vartheta \end{array}. \tag{3.26}$$

Three special cases are distinguished:

$$\begin{array}{r|l} \vartheta = 0 & \text{explicit Euler method} \\ \vartheta = 1 & \text{implicit Euler method} \\ \vartheta = 1/2 & \text{Crank-Nicolson method} \end{array}$$

Theorem 3.3.8: The ϑ-scheme is A-stable for $\vartheta \geq 1/2$.

Proof. $\boxed{\text{DIY}}$ (The stability regions for different ϑ are shown in figure 3.3.) □

3.3.1 Existence and uniqueness of discrete solutions

While it was clear that the steps of an explicit Runge-Kutta method can always be executed, implicit methods require the solution of a possibly nonlinear system of equations. The solvability of such a system is not always clear. We will investigate several cases here: First, Lemma 3.3.9 based on a Lipschitz condition on the right hand side. Since this result suffers from a severe step size constraint, we add Lemma 3.3.10 for DIRK methods based on right hand sides with a one-sided Lipschitz condition. Finally, we present Theorem 3.3.11 for general Runge-Kutta methods with one-sided Lipschitz condition.

Recall the definition of the usual maximum row-sum norm of a matrix A:

$$\|A\|_\infty := \max_{i=1,\dots,s} \sum_{j=1}^{s} |a_{ij}|\,.$$

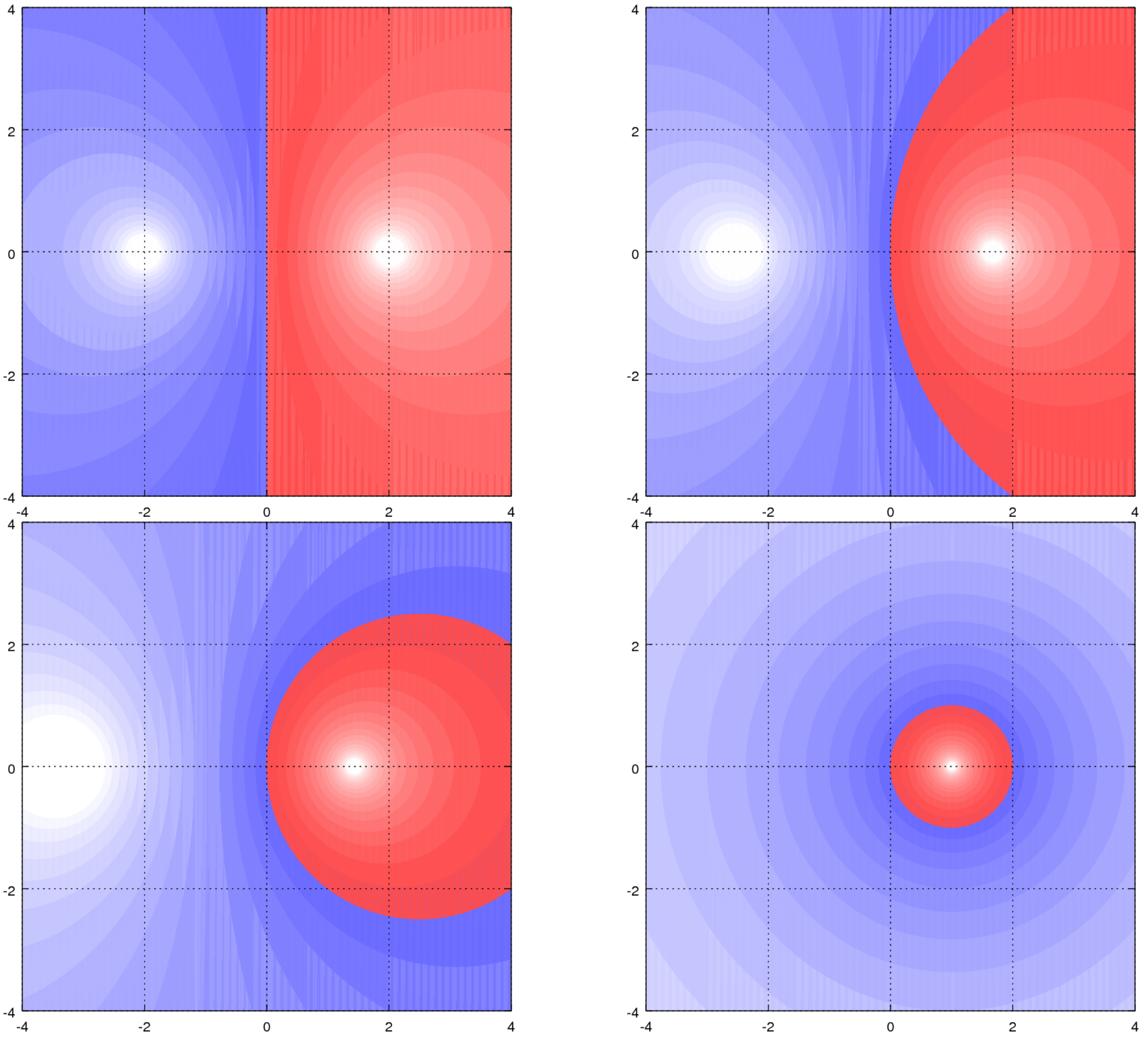


Figure 3.3: Stability regions of the ϑ-scheme with $\vartheta = 0.5$ (Crank-Nicolson), $\vartheta = 0.6$, $\vartheta = 0.7$, and $\vartheta = 1$ (implicit Euler).

Lemma 3.3.9: Let $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be continuous and satisfy the Lipschitz condition with constant L. If

$$hL\|A\|_\infty < 1 \tag{3.27}$$

then, for any $y_0 \in \mathbb{R}^d$, the Runge-Kutta method (3.22) has a unique solution $y_1 \in \mathbb{R}^d$.

Proof. We prove existence and uniqueness by a fixed-point argument. To this end, given $y_0 \in \mathbb{R}^d$, we define the matrix of stage values $K = [k_1, \dots, k_s] \in \mathbb{R}^{d\times s}$ in (3.22).

Given some initial $K^{(0)} \in \mathbb{R}^{d\times s}$, we consider the fixed-point iteration $K^{(m)} = \Psi(K^{(m-1)})$, $m = 1, 2, \dots$, defined columnwise by

$$k_i^{(m)} = \Psi_i(k^{(m-1)}) = f\Bigg(t_0 + c_i h, y_0 + h\sum_{j=1}^{s} a_{ij}k_j^{(m-1)}\Bigg), \quad i = 1, \dots, s,$$

which clearly has the matrix of stage values K as a fixed point. Using on $\mathbb{R}^{d\times s}$ the norm $\|K\| = \max_{i=1,\dots,s}|k_i|$, where $|.|$ is the regular Euclidean norm on $\mathbb{R}^d$, it follows from the Lipschitz continuity of f in its second argument that

$$\|\Psi(K) - \Psi(K')\| \le \Bigg(hL \max_{i=1,\dots,s}\sum_{j=1}^{s}|a_{ij}|\Bigg)\|K - K'\|.$$

Under assumption (3.27), the term in parentheses is strictly less than one and thus, the mapping Ψ is a contraction. Then the Banach fixed-point theorem (cf. theorem A.3.1) yields the unique existence of y_1 . □

Lemma 3.3.10: Let $f : \mathbb{R} \times \mathbb{C}^d \to \mathbb{C}^d$ be continuous, differentiable in its second argument and satisfy the one-sided Lipschitz condition (3.3) with constant ν. Consider an arbitrary DIRK method with $a_{ii} > 0$. If for all $i = 1, \dots, s$

$$h\nu a_{ii} < 1, \tag{3.28}$$

then, for any $y_0 \in \mathbb{C}^d$, each of the (decoupled) nonlinear equations in (3.22a) has a solution $g_i \in \mathbb{C}^d$. In particular, the solution exists unconditionally if the IVP is monotonic, i.e. if $\nu \le 0$.

Proof. The proof simplifies compared to the general case of an IRK, since each stage depends explicitly on the previous stages and implicitly only on itself. Thus, we can write

$$g_i = y_0 + v_i + ha_{ii}f(g_i) \quad \text{with} \quad v_i = h\sum_{j=1}^{i-1} a_{ij}f(g_j). \tag{3.29}$$

For linear IVPs with $f(t, y) := My$ with diagonalizable system matrix M, we have

$$(I - ha_{ii}M)\, g_i = y_0 + v_i\,.$$

Since $\nu = \max_{j=1,\dots,d} \mathrm{Re}(\lambda_j(M))$ (cf. lemma 3.1.8), assumption (3.28) implies that all eigenvalues of $(I - ha_{ii}M)$ have positive real part. Thus, the inverse exists and we obtain a unique solution.

In the nonlinear case, we use a homotopy argument. To this end, we introduce the parameter $\tau \in [0,1]$ and set up the family of equations

$$g(\tau) = y_0 + \tau v_i + ha_{ii} f(g(\tau)) + (\tau - 1) ha_{ii} f(y_0).$$

For $\tau = 0$ this equation has the solution $g(0) = y_0$, and for $\tau = 1$ the solution $g(1) = g_i$. Now, provided g' is bounded on $[0,1]$, we can conclude that a solution exists, since

$$g(1) = g(0) + \int_0^1 g'(s)\,\mathrm{d}s\,. \tag{3.30}$$

To show that g' is bounded, note first that since f was assumed to be differentiable in the second argument

$$\langle f_y(t,y)h + o(|h|), h\rangle = \langle f(t, y+h) - f(x), h\rangle \le \nu |h|^2$$

Dividing by $|h|^2$ and taking the limit as $|h| \to 0$, we obtain with $\widehat{h} = h/|h|$ that

$$\left\langle f_y(t,y)\widehat{h}, \widehat{h}\right\rangle \le \nu \quad \Leftrightarrow \quad \langle f_y(t,y)h, h\rangle \le \nu |h|^2, \quad \text{for all} \quad h \in \mathbb{C}^d.$$

Hence, with

$$g'(\tau) = v_i + ha_{ii} f_y\big(t, g(\tau)\big) g'(\tau) + ha_{ii} f(y_0)$$

we obtain

$$\begin{aligned} |g'(\tau)|^2 &= \big\langle v_i + ha_{ii} f(y_0), g'(\tau)\big\rangle + ha_{ii}\big\langle f_y(t, g(\tau)) g'(\tau), g'(\tau)\big\rangle \\ &\le |v_i + ha_{ii} f(y_0)| |g'(\tau)| + ha_{ii}\nu |g'(\tau)|^2. \end{aligned}$$

Now subtracting the second term on the right hand side and dividing by $1 - ha_{ii}\nu$, which by assumption is positive, it follows that

$$|g'(\tau)|^2 \le \frac{|v_i + ha_{ii} f(y_0)|}{1 - ha_{ii}\nu} |g'(\tau)|,$$

which implies that $g'(\tau)$ is either zero or bounded for all $\tau \in [0,1]$.

Thus, we have proved existence of the stage values g_i. □

If the DIRK method in lemma 3.3.10 is A- or B-stable, then the g_i are unique.

Theorem 3.3.11: Let f be continuously differentiable and let it satisfy the one-sided Lipschitz condition (3.3) with constant ν. If the Runge-Kutta matrix A is invertible and if there exists a diagonal matrix $D = \mathrm{diag}(d_1, \dots, d_s)$ with positive entries, such that

$$h\nu < \frac{\langle x, A^{-1}x\rangle_D}{\langle x, x\rangle_D}, \quad \forall x \in \mathbb{R}^s, \tag{3.31}$$

then the nonlinear system (3.22a) has a solution $(g_1, ..., g_s)$, where $\langle x, y\rangle_D = \langle Dx, y\rangle$.

Proof. We omit the proof here and refer to [HW10, Theorem IV.14.2] □

3.3.2 Considerations on the implementation of Runge-Kutta methods

3.3.12. As we have seen in the proof of lemma 3.3.9, implicit Runge-Kutta methods require the solution of a nonlinear system of size $s \cdot d$, where s is the number of stages and d the dimension of the system of ODEs. DIRK methods are simpler and only require the solution of systems of dimension d. Thus, we should prefer this class of methods, weren't it for the following theorem.

Theorem 3.3.13: A B-stable DIRK method has at most order 4.

Proof. See [HW10, Theorem IV.13.13]. □

Remark 3.3.14. In each step of an IRK, we have to solve a (non-)linear system for the quantities g_i. In order to reduce round-off errors, it is advantageous to solve for $z_i = g_i - y_0$. Especially for small time steps, z_i is expected to be much smaller than g_i. Thus, we have to solve the system

$$z_i = h \sum_{j=1}^{s} a_{ij} f(t_0 + c_j h, y_0 + z_j), \quad i = 1, \dots, s. \tag{3.32}$$

Using the Runge-Kutta matrix A, we rewrite this as

$$\begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix} = A \begin{pmatrix} hf(t_0 + c_1 h, y_0 + z_1) \\ \vdots \\ hf(t_0 + c_s h, y_0 + z_s) \end{pmatrix}. \tag{3.33}$$

We can avoid further function evaluations by then computing

$$y_1 = y_0 + b^T A^{-1} z, \tag{3.34}$$

which again is numerically much more stable than evaluating f (with a possibly large Lipschitz constant).

3.4 Construction of Runge-Kutta methods via quadrature

We finish our discussion of Runge-Kutta methods by describing a systematic way to construct stable, high-order implicit Runge-Kutta methods.

Definition 3.4.1 (Simplifying order conditions):

$$B(p): \quad \sum_{i=1}^{s} b_i c_i^{q-1} = \frac{1}{q} \qquad q = 1, \dots, p \tag{3.35a}$$

$$C(p): \quad \sum_{j=1}^{s} a_{ij} c_j^{q-1} = \frac{c_i^q}{q} \qquad \begin{array}{c} q = 1, \dots, p \\ i = 1, \dots, s \end{array} \tag{3.35b}$$

$$D(p): \quad \sum_{i=1}^{s} b_i a_{ij} c_i^{q-1} = \frac{b_j}{q}(1 - c_j^q) \qquad \begin{array}{c} q = 1, \dots, p \\ j = 1, \dots, s \end{array} \tag{3.35c}$$

Theorem 3.4.2: Consider a (general) Runge-Kutta method that satisfies condition $B(p)$ in (3.35a), condition $C(\xi)$ in (3.35b), and condition $D(\eta)$ in (3.35c) with $\xi \geq p/2 - 1$ and $\eta \geq p - \xi - 1$. Then the method has consistency order p.

Proof. For the proof, we refer to [HNW09, Ch. II, Theorem 7.4]. Here, we only observe, that

$$\int_0^1 \tau^{q-1}\,\mathrm{d}\tau = \frac{1}{q}, \qquad \int_0^{c_i} \tau^{q-1}\,\mathrm{d}\tau = \frac{c_i^q}{q}.$$

Thus, condition $B(p)$ in (3.35a) simply states that the quadrature rule with quadrature points $c_1, \ldots, c_s$ and quadrature weights $b_1, \ldots, b_s$ for the integral $\int_0^1 f(\tau)\,\mathrm{d}\tau$ is exact for all polynomials f up to order $p-1$. Similarly, condition $C(\xi)$ in (3.35b) states that the quadrature rule with quadrature points c_j and quadrature weights a_{ij}, for $j = 1, \ldots, s$, for the integral $\int_0^{c_i} f(\tau)\,\mathrm{d}\tau$ is exact for all polynomials f up to order $\xi - 1$. □

3.4.1 Gauss-, Radau-, and Lobatto-quadrature

3.4.3. In this subsection, we review some of the basic facts of quadrature formulas based on orthogonal polynomials (cf. Numerik 0 for details).

Definition 3.4.4: Let $p_n(x)$ be the (shifted) Legendre polynomial of degree n on $[0, 1]$, up to scaling. These can be compactly defined by

$$p_n(x) = \frac{\mathrm{d}^n}{\mathrm{d}x^n} x^n (x-1)^n.$$

A quadrature formula for $\int_0^1 f(x)\mathrm{d}x$ that uses the n roots of p_n as its quadrature points and the integrals of the Lagrange interpolating polynomials at those points as its weights is called a **Gauss rule**.

Lemma 3.4.5: The Gauss quadrature formula $Q_{n-1}^{[a,b]}(f)$ with n points for approximating the integral $\int_a^b f\,\mathrm{d}x$ is exact for polynomials of degree $2n-1$. If $f \in C^{2n}[a, b]$ and $h := b - a$ then

$$\left| Q_{n-1}^{[a,b]}(f) - \int_a^b f\,\mathrm{d}x \right| = \mathcal{O}(h^{2n+1}).$$

Proof. See Numerik 0. (Please note that in Numerik 0 we numbered the quadrature nodes $x_0, \ldots, x_{n-1}$ and thus n here is $n-1$ in the notes to Numerik 0.) □

Remark 3.4.6. An important alternative set of quadrature formulae are the Radau and Lobatto formulas.

The **Radau quadrature** formulae are similar to the Gauss rules, but they use one end point of the interval $[0, 1]$ and the roots of orthogonal polynomials of degree $n-1$ as their abscissas. We distinguish left and right Radau quadrature formulae, depending on which end is included. **Lobatto quadrature** formulae use both end points and the roots of a polynomial of degree $n-2$. The polynomials are

$$\begin{aligned}
&\text{Radau left} && p_n(x) = \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\bigl(x^n(x-1)^{n-1}\bigr), && (3.36)\\
&\text{Radau right} && p_n(x) = \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}}\bigl(x^{n-1}(x-1)^{n}\bigr), && (3.37)\\
&\text{Lobatto} && p_n(x) = \frac{\mathrm{d}^{n-2}}{\mathrm{d}x^{n-2}}\bigl(x^{n-1}(x-1)^{n-1}\bigr). && (3.38)
\end{aligned}$$

A Radau quadrature formula with n points is exact for polynomials of degree $2n-2$. A Lobatto quadrature formula with n points is exact for polynomials of degree $2n-3$. The quadrature weights of these formulae are positive.

3.4.2 Collocation methods

3.4.7. An alternative to solving IVP in individual points in time, is to develop methods, which first approximate the solution function through a polynomial.

However, as we have seen in Numerik 0, polynomials are not suited for high-order interpolation over large intervals. Therefore, as for composite quadrature rules, we use polynomial interpolation over subintervals in the form of Runge-Kutta methods. The subintervals correspond to the time steps and the quadrature points are the stages.

Definition 3.4.8: The **collocation polynomial** $y(t) \in \mathbb{P}_s$ of an s-stage **collocation method** with pairwise different support points $c_1, \dots, c_s$ is defined uniquely through the $s+1$ conditions:

$$y(t_0) = y_0 \tag{3.39a}$$

$$y'(t_0 + c_i h) = f\bigl(t_0 + c_i h, y(t_0 + c_i h)\bigr) \quad i = 1, \dots, s. \tag{3.39b}$$

The value at the next time step is then defined as

$$y_1 = y(t_0 + h). \tag{3.39c}$$

Lemma 3.4.9: An s-stage collocation method with the points c_1 to c_s defines a Runge-Kutta method, as defined in definition 3.3.2, with the coefficients c_i and

$$a_{ij} = \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau, \qquad b_i = \int_0^1 L_i(\tau)\,\mathrm{d}\tau, \tag{3.40}$$

where

$$L_j(\tau) = L_j^{(s-1)}(\tau) = \prod_{\substack{k=1\\k\neq j}}^{s} \frac{\tau - c_k}{c_j - c_k}, \quad j = 1, \dots, s,$$

are the Lagrange interpolation polynomials associated to the point set $\{c_1, \dots, c_s\}$.

Proof. The polynomial $y'(\tau)$ is of degree $s-1$ and thus uniquely defined by the s interpolation conditions in equation (3.39b). Setting $y'(t_0+c_ih) = f\big(t_0+c_ih, y(t_0+c_ih)\big) = k_i$ we obtain

$$y'(t_0+\tau h) = \sum_{j=1}^{s} k_j \cdot L_j(\tau), \tag{3.41}$$

where $L_1,\dots,L_s$ are the Lagrange interpolation polynomials. By integration we obtain:

$$g_i = y(t_0+c_ih) = y_0 + h\int\limits_0^{c_i} y'(t_0+\tau h)\,\mathrm{d}\tau = y_0 + h\sum_{j=1}^{s} k_j \int_0^{c_i} L_j(\tau)\,\mathrm{d}\tau, \tag{3.42}$$

which, by comparison with (3.22a), defines the coefficients a_{ij}. Integrating (3.41) from 0 to 1 instead, we obtain the coefficients b_j by comparison with (3.22c). □

Lemma 3.4.10: An implicit s-stage Runge-Kutta method, with pairwise different support points c_i, is a collocation method if and only if simplifying conditions $B(s)$ (3.35a) and $C(s)$ in (3.35b) are satisfied. Thus, an s-stage collocation method is of order (at least) s.

Proof. Consider an s-stage RK method. Condition $B(s)$ leads to a system of s conditions for the s coefficients $b_1,\dots,b_s$. The system matrix is the transpose of the Vandermonde matrix V with entries $V_{i,q} := c_i^{q-1}$ which (for pairwise different c_i) is invertible. Therefore these coefficients are defined uniquely. Similarly, for each $i = 1,\dots,s$, condition $C(s)$ leads to a uniquely solvable system of s conditions for the s coefficients $a_{i,j}$, $j = 1,\dots,s$, with the same system matrix. Thus, all the coefficients are defined uniquely.

On the other hand, (3.35b) yields for $q < s$:

$$\sum_{j=1}^{s} a_{ij}c_j^q = \frac{c_i^{q+1}}{q+1} = \int_0^{c_i} t^q\,\mathrm{d}t.$$

As a consequence of linearity we have

$$\sum_{j=1}^{s} a_{ij}p(c_j) = \int_0^{c_i} p(t)\,\mathrm{d}t, \qquad \forall p \in \mathcal{P}_{s-1}.$$

Applying this to the Lagrange interpolation polynomials $L_j(t)$, we obtain the coefficients of equation (3.40), which were in turn computed from the collocation polynomial, proving the equivalence.

It follows from theorem 3.4.2 that a Runge-Kutta method that satisfies $B(s)$ and $C(s)$ has consistency order (at least) s. □

Theorem 3.4.11: Consider a collocation method with s pairwise different support points c_i and define

$$\pi(\tau) = \prod_{i=1}^{s} (\tau - c_i). \tag{3.43}$$

If π is orthogonal on $[0,1]$ to all polynomials of degree $r-1$ for $r \le s$, then the collocation method (3.39) is of consistency order $p = s + r$.

Proof. We have already shown in the proof of Lemma 3.4.10, that for any collocation method with s stages, $B(s)$ and $C(s)$ hold.

The condition on π implies that on the interval $[0,1]$ the quadrature rule is in fact exact for polynomials of degree $s+r-1$ (cf. Numerik 0 for the case $r = s$), so that we have $B(s+r)$. Therefore, to prove conistency order $p = s + r$ it remains to show $D(r)$.

First, we observe that due to $C(s)$ and $B(s+r)$, for all $p \le s$ and $q \le r$, we have

$$\sum_{j=1}^{s} \left(\sum_{i=1}^{s} b_i a_{ij} c_i^{q-1} \right) c_j^{p-1} = \sum_{i=1}^{s} b_i c_i^{q-1} \frac{c_i^p}{p} = \frac{1}{p} \sum_{i=1}^{s} b_i c_i^{p+q-1} = \frac{1}{p(p+q)}.$$

Furthermore, since $B(s+r)$ we have for the same p and q:

$$\sum_{j=1}^{s} b_j \big(1 - c_j^q\big) c_j^{p-1} = \sum_{j=1}^{s} \big(b_j c_j^{p-1} - b_j c_j^{p+q-1} \big) = \frac{1}{p} - \frac{1}{p+q} = \frac{q}{p(p+q)}.$$

Subtracting $\frac{1}{q}$ times the second result from the first we get

$$0 = \frac{1}{p(p+q)} - \frac{1}{p(p+q)} = \sum_{j=1}^{s} c_j^{p-1} \underbrace{\left(\sum_i b_i c_i^{q-1} a_{ij} - \frac{1}{q} b_j \big(1 - c_j^q\big) \right)}_{=:\zeta_j}.$$

This holds for $p = 1, \dots, s$ and thus amounts to a homogeneous, linear system in the variables ζ_j with system matrix V^T. Thus, $\zeta_j = 0$ and the theorem holds. □

Corollary 3.4.12. *The consistency order p of an s-stage collocation method satisfies*

$$s \le p \le 2s.$$

Proof. The polynomial $\pi(t)$ in (3.43) is of degree s. If $\pi = p_s$, the Legendre polynomial of degree s on $[0,1]$, then π is orthogonal to all polynomials of degree $s-1$ by construction (cf. Numerik 0 for details). Thus, it follows from theorem 3.4.11 that there exists an s-stage collocation method of order $p = 2s$.

On the other hand, we know from Numerik 0 that there exists no quadrature rule such that $B(2s+1)$ is satisfied. Hence, it is clear the conistency order of an s-stage collocation method satisfies $p \le 2s$.

The lower bound has already been proved in lemma 3.4.10. □

In particular, if we consider the scalar model equation $u' = \lambda u$ with exact solution $u(t) = e^{\lambda t} = \sum_{j=0}^{\infty} (\lambda t)^j / j!$, the best we can hope for is that the collocation polynomial $y(t)$ matches the first $2s - 1$ terms in this infinite sum, such that

$$|u_1 - y_1| = \mathcal{O}(h^{2s}).$$

Definition 3.4.13: An s-stage **Gauss collocation method** is a collocation method, where the collocation points are the set of s Gauss points in the interval $[0, 1]$, namely the roots of the Legendre polynomial of degree s.

Example 3.4.14: (2- and 3-stage Gauss collocation methods)

$$\begin{array}{c|cc}
\frac{3-\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\
\frac{3+\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\qquad
\begin{array}{c|ccc}
\frac{5-\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\
\frac{5+\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\
\hline
 & \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array}$$

Theorem 3.4.15: The s-stage Gauss collocation method is consistent of order $2s$ and thus of optimal order.

Proof. Follows immediately from the proof of Corollary 3.4.12. □

Theorem 3.4.16: Gauss collocation methods are B-stable. The stability region of Gauss collocation is exactly the left half-plane of $\mathbb{C}$.

Proof. Let f be monotonic and let $y(t)$ and $z(t)$ be the collocation polynomials according to (3.39) with respect to initial values y_0 or z_0. Analogous to the proof of theorem 3.1.6 we introduce the auxiliary function $m(t) = |z(t) - y(t)|^2$. In the collocation points $\xi_i = t_0 + c_i h$ we have

$$\begin{aligned} m'(\xi_i) &= 2\mathrm{Re}\left\langle z'(\xi_i) - y'(\xi_i), z(\xi_i) - y(\xi_i)\right\rangle \\ &= 2\mathrm{Re}\left\langle f(\xi_i, z(\xi_i)) - f(\xi_i, y(\xi_i)), z(\xi_i) - y(\xi_i)\right\rangle \le 0. \end{aligned} \tag{3.44}$$

Since Gauss quadrature is exact for polynomials of degree $2s - 1$ and m' is a polynomial of degree $2s - 1$, we have:

$$\begin{aligned} |z_1 - y_1|^2 = m(t_0 + h) &= m(t_0) + \int_{t_0}^{t_0+h} m'(t)\,\mathrm{d}t \\ &= m_0 + h \sum_{i=1}^{s} b_i m'(\xi_i) \ \le \ m(t_0) = |z_0 - y_0|^2, \end{aligned}$$

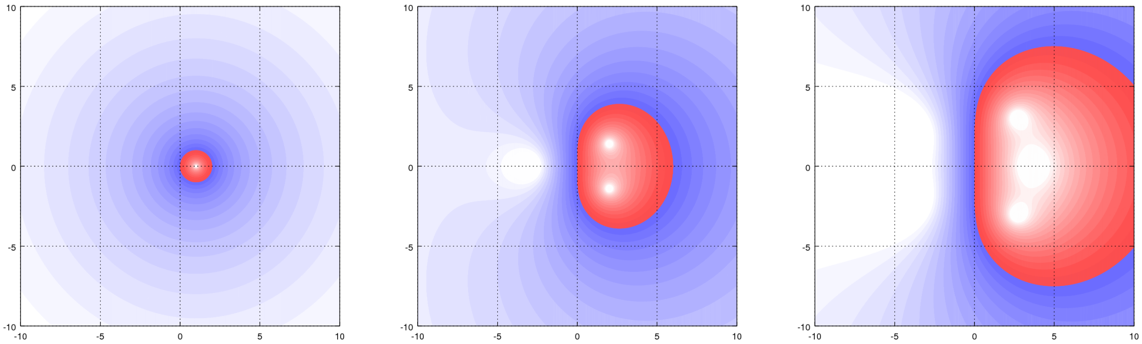


Figure 3.4: Stability domains of right Radau-collocation methods with one (implicit Euler), two, and three collocation points (left to right). Note the different scaling of coordinate axes in comparison with previous figures.

which establishes B-stability.

To show that the stability region is exactly the left half-plane of $\mathbb{C}$ we refer to the problem sheet. □

Remark 3.4.17. Similarly, we can construct collocation rules based on Radau- and Lobatto-quadrature. As in the proof of theorem 3.4.15, it can be shown that the s-stage Radau- and Lobatto-collocation methods are of orders $2s-1$ and $2s-2$, respectively.

Also as in the case of Gauss-quadrature it can be shown that collocation methods based on Radau- and Lobatto quadrature are B-stable (cf. [HW10]). In fact, Radau-collocation methods with right end point of the interval $[0,1]$ included in the quadrature set are L-stable.

The first right Radau collocation method with $s=1$ is simply the implicit Euler method. The definitions of the next two are given in example 3.4.18. The stability regions of the first three are shown in Figure 3.4.

Observe that the stability domains are shrinking with order of the method. Also, observe that the computation of y_1 coincides with that of g_s, such that we can save a few operations.

Example 3.4.18 (2- and 3-stage right Radau collocation methods):

$$\begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array} \qquad \begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

3.5 Symplectic integration of Hamiltonian Systems

So far we have mainly been concerned with stability of numerical integrators for systems with longterm behaviour $u(t) \to 0$ as $t \to \infty$. In this section, we will now consider a very important class of ODEs where certain integrals of the solution are preserved over time.

The following section is based on [Lub].

3.5.1 Hamiltonian Systems

Hamilton's equations appeared first, among thousands of other formulas, and inspired by previous research in optics, in [Hamilton, 1834]. The next mile-stones in the exposition of the theory were the monumental three volumes of Poincaré (1892,1893,1899) on celestial mechanics. Beyond that, Hamiltonian systems became fundamental in many branches of physics, e.g. in the context of particle accelerators.

Linking back to Examples 1.1.4 and 1.1.5, we start following [Lagrange, 1760 & 1788] and suppose that the position of a mechanical system with d degrees of freedom is described by $q = (q_1, \ldots, q_d)^\top$ as **generalized coordinates** (this can be for example Cartesian coordinates, angles, arc lengths along a curve, etc). The theory is then built upon two pillars, namely an expression

$$T = T(q, q')$$

which represents the **kinetic energy** (and which is often of the form $\frac{1}{2}(q')^\top M(q) q'$ where $M(q)$ is symmetric and positive definite), and by a function

$$U = U(q)$$

representing the **potential energy**. Then, after denoting by

$$L = T - U \tag{3.45}$$

the corresponding **Lagrangian**, the coordinates $q_1(t), \ldots, q_d(t)$ obey the differential equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial q'}\right) = \frac{\partial L}{\partial q}\ , \tag{3.46}$$

which constitute the Lagrange equations of the system. A numerical (or analytical) integration of these equations allows one to predict the motion of any such system from given initial values.

Example 3.5.1. For a point of mass m in $\mathbb{R}^3$ with Cartesian coordinates $x = (x_1, x_2, x_3)^\top$ we have $T(x') = \frac{1}{2}m|x'|^2$. If the point is assumed to move in a conservative force field $F(x) = -\nabla U(x)$, the Lagrange equations (3.46) become $mx'' = F(x)$, which is Newton's second law. The equations (1.2) for the planetary motion in Example 1.1.5 are precisely of this form.

Example 3.5.2 (Pendulum)**.** Another simple and common example is the mathematical pendulum of a mass m swinging (friction-less) on a (weight-less) string of length ℓ, subject only to gravitational acceleration g. Here, the coordinate is the angle α and the position of the mass at time t is uniquely determined by the initial release position at $t = 0$. The kinetic and potential energies are given by

$$T(\alpha') = \frac{1}{2}m|x'|^2 = \frac{1}{2}m\ell^2(\alpha')^2 \quad \text{and} \quad U(\alpha) = mgx_2 = -mg\ell\cos\alpha,$$

respectively. The Lagrange equations reduce to

$$-mg\ell\sin\alpha - m\ell^2\alpha'' = 0 \quad \Leftrightarrow \quad \alpha'' = -\frac{g}{\ell}\sin\alpha.$$

Hamilton (1834) simplified the structure of Lagrange's equations and turned them into a form that has remarkable symmetry, by

- introducing Poisson's variables, the conjugate momenta
$$p_k = \frac{\partial L}{\partial q_k'}(q, q'), \quad \text{for } k = 1, \ldots, d, \tag{3.47}$$
- and considering the **Hamiltonian**
$$H := p^\top q' - L(q, q') \tag{3.48}$$
as a function of p and q, i.e., taking $H = H(p, q)$ obtained by expressing q' as a function of p and q via (3.47).

This requires of course that (3.47) defines, for every q, a continuously differentiable bijection $q' \leftrightarrow p$. This map is called the **Legendre transform**.

Theorem 3.5.3: Lagrange's equations (3.46) are equivalent to Hamilton's equations

$$p_k' = -\frac{\partial H}{\partial q_k}(p, q), \qquad q_k' = \frac{\partial H}{\partial p_k}(p, q), \qquad k = 1, \ldots, d. \tag{3.49}$$

Proof. The definitions (3.47) and (3.48) for the momenta p and for the Hamiltonian H imply that, for all $k = 1, \ldots, d$,

$$\begin{aligned}\frac{\partial H}{\partial p_k} &= q_k' + p^\top \frac{\partial q'}{\partial p_k} - \frac{\partial L}{\partial q'}\frac{\partial q'}{\partial p_k} = q_k'\ ,\\ \frac{\partial H}{\partial q_k} &= p^\top \frac{\partial q'}{\partial q_k} - \frac{\partial L}{\partial q_k} - \frac{\partial L}{\partial q'}\frac{\partial q'}{\partial q_k} = -\frac{\partial L}{\partial q_k}\ .\end{aligned}$$

The Lagrange equations (3.46) are therefore equivalent to (3.49). □

One crucial property of a Hamiltonian system (3.49) is that any solution of the system preserves the Hamiltonian.

Theorem 3.5.4: Let $\big(p(t), q(t)\big)$ be the solution of the Hamiltonian system (3.49) with initial condition (p_0, q_0) at $t = 0$. Then

$$H\big(p(t), q(t)\big) = H\big(p_0, q_0\big), \quad \text{for all } t \geq 0.$$

In other words, $H(p, q)$ is a so-called **first integral** (or a **conserved quantity**) of the ODE system (3.49).

Proof. Using (3.49), we have

$$\frac{\mathrm{d}}{\mathrm{d}t} H\big(p(t), q(t)\big) = \sum_{k=1}^{d} \frac{\partial H}{\partial p_k} p_k' + \frac{\partial H}{\partial q_k} q_k' = 0.$$

□

Example 3.5.5 (The case of quadratic T)**.** Let $T = \frac{1}{2}(q')^\top M(q) q'$, where $M(q)$ is a symmetric and positive definite matrix. For a fixed q, we have $p = M(q)q'$ so that the existence of the Legendre transform is established. Further, by replacing the variable q' by $M(q)^{-1}p$ in the definition of $H(p, q)$ in (3.48), we obtain

$$\begin{aligned} H(p,q) &= p^\top M(q)^{-1} p - L\Big(q, M(q)^{-1}p\Big) \\ &= p^\top M(q)^{-1}p - \frac{1}{2} p^\top M(q)^{-1} p + U(q) = \frac{1}{2} p^\top M(q)^{-1} p + U(q) \end{aligned}$$

and the Hamiltonian is $H = T + U$, which is the total energy of the mechanical system. Thus, the total energy of the system is conserved.

We have now seen several examples of Hamiltonian systems, e.g., the Kepler problem in Example 1.1.4, the problem of celestial mechanics in Example 1.1.5 or the pendulum in Example 3.5.2. In the following, we consider Hamiltonian systems (3.49) where the Hamiltonian $H(p, q)$ is arbitrary and not necessarily related to a mechanical problem.

3.5.2 Symplectic transformations

In addition to the property in Theorem 3.5.4 that the Hamiltonian is preserved, Hamiltonian systems have another important property, namely the **symplecticity** of its flow.

The basic objects to be studied are two-dimensional parallelograms lying in $\mathbb{R}^{2d}$. We consider all parallelograms in the (p, q)-space spanned by the two vectors

$$\xi = \begin{pmatrix} \xi^p \\ \xi^q \end{pmatrix}, \qquad \eta = \begin{pmatrix} \eta^p \\ \eta^q \end{pmatrix}$$

with $\xi^p, \xi^q, \eta^p, \eta^q \in \mathbb{R}^d$, i.e. the set

$$P = \{s\xi + r\eta : 0 \le s \le 1,\ 0 \le r \le 1\}.$$

For $d = 1$, we consider the **oriented area**

$$\operatorname{area}(P) = \det \begin{pmatrix} \xi^p & \eta^p \\ \xi^q & \eta^q \end{pmatrix} = \xi^p \eta^q - \xi^q \eta^p \tag{3.50}$$

(see left picture of Fig. 3.5). In higher dimensions, we replace this by the sum of the oriented areas of the projections of P onto the coordinate planes (p_i, q_i), i.e., by

$$\omega(\xi, \eta) = \sum_{i=1}^{d} \det \begin{pmatrix} \xi_i^p & \eta_i^p \\ \xi_i^q & \eta_i^q \end{pmatrix} = \sum_{i=1}^{d} \big(\xi_i^p \eta_i^q - \xi_i^q \eta^p\big). \tag{3.51}$$

This defines a bilinear map acting on vectors of $\mathbb{R}^{2d}$, which will play a central role for Hamiltonian systems. In matrix notation, this map has the form

$$\omega(\xi,\eta) = \xi^\top J\eta \quad \text{with} \quad J = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \tag{3.52}$$

where I is the identity matrix of dimension d.

Definition 3.5.6: A linear mapping $A : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is called **symplectic** if

$$A^\top JA = J$$

or, equivalently, if $\omega(A\xi, A\eta) = \omega(\xi,\eta)$ for all $\xi,\eta \in \mathbb{R}^{2d}$.

In the case $d = 1$, where the expression $\omega(\xi,\eta)$ represents the area of the parallelogram P, symplecticity of a linear mapping A is therefore equivalent to the area preservation of A (see Fig. 3.5).

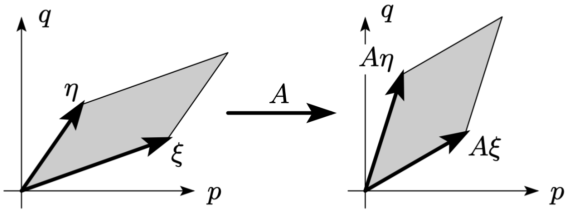


Figure 3.5: Symplecticity (area preservation) of a linear mapping.

In the general case $d > 1$, symplecticity means that the sum of the oriented areas of the projections of P onto (p_i, q_i) is the same as that for the transformed parallelograms $A(P)$.

We now turn our attention to nonlinear mappings. Differentiable functions can be locally approximated by linear mappings. This justifies the following definition.

Definition 3.5.7: A differentiable map $g : U \to \mathbb{R}^{2d}$ (where $U \subset \mathbb{R}^{2d}$ is open) is called **symplectic** if the Jacobian matrix $Dg(p,q)$ is everywhere symplectic, i.e., if

$$Dg(p,q)^T JDg(p,q) = J \qquad \text{or} \qquad \omega\big(Dg(p,q)\xi, Dg(p,q)\eta\big) = \omega(\xi,\eta),$$

for all $p,q \in U$ and all $\xi,\eta \in \mathbb{R}^{2d}$.

Let us give a geometric interpretation of symplecticity for nonlinear mappings. Consider a 2-dimensional sub-manifold M of the $2d$-dimensional set U, and suppose that it is given as the image $M = \psi(K)$ of a compact set $K \subset \mathbb{R}^2$, where $\psi(r,s)$ is a continuously differentiable function. The manifold M can then be considered as the limit of a union of small parallelograms spanned by the vectors

$$\frac{\partial\psi}{\partial r}\mathrm{d}r \qquad \text{and} \qquad \frac{\partial\psi}{\partial s}\mathrm{d}s\,.$$

For one such parallelogram we consider (as above) the sum over the oriented areas of its projections onto the (p_i, q_i)-planes. Summing over all parallelograms and taking the limit gives the expression

$$\Omega(M) := \iint_K \omega\left(\frac{\partial\psi}{\partial r}(r,s), \frac{\partial\psi}{\partial s}(r,s)\right) \mathrm{d}r\mathrm{d}s\ . \tag{3.53}$$

The transformation formula for double integrals implies that $\Omega(M)$ is independent of the parametrization ψ of M.

Lemma 3.5.8: If the mapping $g : U \to \mathbb{R}^{2d}$ is symplectic on U, then it preserves the expression $\Omega(M)$, i.e.,

$$\Omega\big(g(M)\big) = \Omega(M)$$

holds for all 2-dimensional manifolds M that can be represented as the image of a continuously differentiable function ψ.

Proof. The manifold $g(M)$ can be parametrized by $g \circ \psi$. We have

$$\Omega(g(M)) = \iint_K \omega\left(\frac{\partial g \circ \psi}{\partial r}(r,s), \frac{\partial g \circ \psi}{\partial s}(r,s)\right) \mathrm{d}r\mathrm{d}s = \Omega(M)$$

because $D(g \circ \psi)(r,s) = Dg(\psi(r,s))D\psi(r,s)$ and g is a symplectic transformation. □

For $d = 1$, M is already a subset of $\mathbb{R}^2$ and we choose $K = M$ with ψ the identity map. In this case, $\Omega(M) = \iint_M \mathrm{d}r\mathrm{d}s$ simply represents the area of M. Hence, Lemma 3.5.8 states that all symplectic mappings (also nonlinear ones) are area preserving.

We are now able to prove the main result of this section. We use the notation $y = (p, q)$, and we write the Hamiltonian system (3.49) in the form

$$y' = J^{-1}\nabla H(y), \tag{3.54}$$

where J is the matrix defined in (3.52) and $\nabla H(y) = DH(y)^\top$.

Definition 3.5.9: The **flow** $\varphi_t : U \subset \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ of a first-order ODE (in particular, of a Hamiltonian system) is the mapping that advances the initial condition y_0 by time $t \geq 0$ along the solution of the ODE , i.e.,

$$\varphi_t(y_0) = y(t),$$

where $y(t)$ is the solution of the system (3.54) with initial value $y(0) = y_0$.

Definition 3.5.10: The **variational equation** of the autonomous IVP $y' = f(y)$ with initial condition $y(0) = y_0$ is the linear differential equation

$$\Psi'(t) = \nabla_{y_0} f\big(\varphi_t(y_0)\big)\Psi(t),$$

whose solution Ψ is the derivative with respect to the initial condition y_0 of the solution $y(t)$ of the IVP.

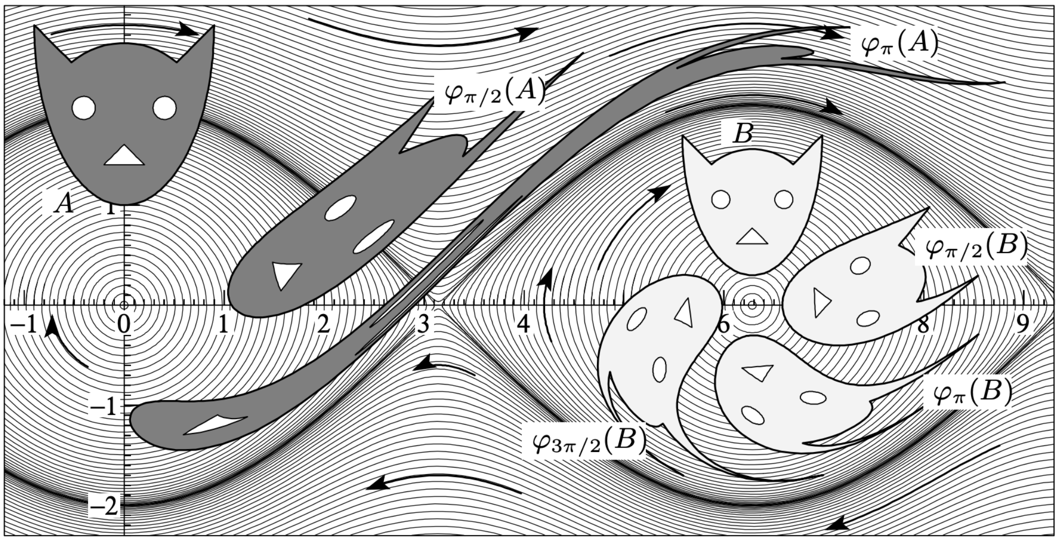


Figure 3.6: Area preservation of the flow of Hamiltonian systems.

Theorem 3.5.11: (Poincaré, 1899): Let $H(p, q)$ be a twice continuously differentiable function on $U \subset \mathbb{R}^{2d}$. Then, for each fixed t, the flow φ_t is a symplectic map wherever it is defined.

Proof. The derivative $D\varphi_t = D_{y_0}\varphi_t$ of the flow φ_t with respect to $y_0 = (p_0, q_0)$ is a solution of the variational equation of the Hamiltonian system (3.54). This is of the form $\Psi' = J^{-1}\nabla^2 H(\varphi_t(y_0))\Psi$, where $\nabla^2 H(p, q)$ is the Hessian of $H(p, q)$, which is symmetric by definition. Therefore, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(D\varphi_t^\top J D\varphi_t\right) &= \left(\frac{\mathrm{d}}{\mathrm{d}t}D\varphi_t\right)^\top J D\varphi_t + D\varphi_t^\top J\left(\frac{\mathrm{d}}{\mathrm{d}t}D\varphi_t\right)\\ &= D\varphi_t^\top \nabla^2 H(\varphi_t(y_0))J^{-\top}J^{-1}D\varphi_t + D\varphi_t^\top \nabla^2 H(\varphi_t(y_0))D\varphi_t = 0,\end{aligned}$$

because $J^\top = -J$ and thus $J^{-\top}J = -I$. Since $\varphi_0(y_0) = y_0$ is the identity map, the relation

$$D\varphi_t^\top J D\varphi_t = J \tag{3.55}$$

is satisfied for $t = 0$. Thus, it follows from above that (3.55) is satisfied for all t and for all $y_0 = (p_0, q_0)$ and φ_t is symplectic, as long as the solution remains in the domain of definition of H. □

Example 3.5.12. We illustrate this theorem for the pendulum problem in Example 3.5.2 with $m = \ell = g = 1$. We have $q = \alpha$, $p = \alpha'$ and the Hamiltonian is given by

$$H(p, q) = p^2/2 - \cos q.$$

Fig. 3.6 shows level curves of this function, and it also illustrates the area preservation of the flow φ_t. Indeed, by Theorem 3.5.11 and Lemma 3.5.8, the areas of A and $\varphi_t(A)$, as well as those of B and $\varphi_t(B)$ are the same, although their appearance is completely different.

Symplecticity of the flow is in fact a characteristic property for Hamiltonian systems.

Definition 3.5.13: We call a differential equation $y' = f(y)$ **locally Hamiltonian**, if for every $y_0 \in U$ there exists a neighbourhood where $f(y) = J^{-1}\nabla H(y)$ for some function H.

Theorem 3.5.14: Let $f : U \to \mathbb{R}^{2d}$ be continuously differentiable. Then, $y' = f(y)$ is locally Hamiltonian if and only if its flow $\varphi_t(y)$ is symplectic for all $y \in U$ and for all sufficiently small.

Proof. The necessity follows from Theorem 3.5.11. See [Lub] for a proof of the converse. □

An important property of symplectic transformations, which goes back to Jacobi (1836), is that they preserve the Hamiltonian character of the differential equation. Such transformations have been termed **canonical** since the 19th century. The next theorem shows that canonical and symplectic transformations are the same.

Theorem 3.5.15: Let $\psi : U \to V$ be a change of coordinates such that ψ and ψ^{-1} are continuously differentiable functions. Let ψ be symplectic and $y' = J^{-1}\nabla H(y)$ be a Hamiltonian system. Then, the transformed coordinates $z = \psi(y)$ satisfy the Hamiltonian system

$$z' = J^{-1}\nabla K(z) \quad \text{with} \quad K(\psi(y)) = H(y). \tag{3.56}$$

Conversely, if ψ transforms every Hamiltonian system to another Hamiltonian system via (3.56), then ψ is symplectic.

Proof. Since $z' = D\psi(y)y'$ and $D\psi(y)^\top \nabla K(\psi(y)) = \nabla H(y)$, the Hamiltonian system $y' = J^{-1}\nabla H(y)$ becomes

$$z' = D\psi(y)J^{-1}D\psi(y)^\top \nabla K(z) \tag{3.57}$$

in the transformed variables. It is equivalent to (3.56) if

$$D\psi(y)J^{-1}D\psi(y)^\top = J^{-1}. \tag{3.58}$$

Multiplying this relation from the right by $D\psi(y)^{-\top}$ and from the left by $D\psi(y)^{-1}$, and then taking its inverse yields $J = D\psi(y)^\top J D\psi(y)$, which shows that (3.58) is equivalent to the symplecticity of ψ.

For the inverse relation, we note that (3.57) is Hamiltonian for all $K(z)$ if and only if (3.58) holds. □

3.5.3 Examples of symplectic integrators

Since symplecticity is a characteristic property of Hamiltonian systems (cf. Theorem 3.5.14), it is natural to search for numerical methods that share this property. Pioneering work on symplectic integration is due to de Vogelaere (1956), Ruth (1983), and Feng Kang (1985). A good book on the subject is Leimkuhler & Reich [LR05].

Definition 3.5.16: A numerical one-step method is called **symplectic** if the one-step map

$$y_1 = \Phi_h(y_0),$$

the so-called **numerical flow**, is symplectic whenever the method is applied to a smooth Hamiltonian system.

Remark 3.5.17. It follows directly from Definition 3.5.16 that the composition of (two or more) symplectic transformations is again symplectic.

In the following, we show the symplecticity of various numerical methods when applied to the Hamiltonian system

$$p' = -H_q(p, q) \quad \text{and} \quad q' = H_p(p, q),$$

where H_p and H_q denote the column vectors of partial derivatives of the Hamiltonian $H(p, q)$ with respect to p and q, respectively. As above, we can also write equivalently $y' = J^{-1}\nabla H(y)$ for $y = (p, q)$.

Theorem 3.5.18: (de Vogelaere, 1956): The so-called **symplectic Euler** methods

$$\begin{aligned} p_{n+1} &= p_n - hH_q(p_{n+1}, q_n) \\ q_{n+1} &= q_n + hH_p(p_{n+1}, q_n) \end{aligned} \quad \text{and} \quad \begin{aligned} p_{n+1} &= p_n - hH_q(p_n, q_{n+1}) \\ q_{n+1} &= q_n + hH_p(p_n, q_{n+1}) \end{aligned} \tag{3.59}$$

are symplectic methods of order 1.

Proof. See Exercise Sheet. $\Big[$*Hint: Using the Implicit Function Theorem, we can compute* $D_{y_0}(\Phi_h(y_0)) = \partial y_1/\partial y_0$ *for the two methods in* (3.59). *It then suffices to verify the symplecticity condition* $D_{y_0}(\Phi_h(y_0))^\top J D_{y_0}(\Phi_h(y_0)) = J$.$\Big]$ □

Next, we consider the most popular symplectic numerical integrator, the Störmer-Verlet scheme (also called **Leapfrog** method).

Theorem 3.5.19: The **Störmer–Verlet** schemes

$$\begin{aligned} p_{n+1/2} &= p_n - \tfrac{h}{2} H_q(p_{n+1/2}, q_n) \\ q_{n+1} &= q_n + \tfrac{h}{2}\Big(H_p(p_{n+1/2}, q_n) + H_p(p_{n+1/2}, q_{n+1})\Big) \\ p_{n+1} &= p_{n+1/2} - \tfrac{h}{2} H_q(p_{n+1/2}, q_{n+1}) \end{aligned} \tag{3.60}$$

and

$$\begin{aligned} q_{n+1/2} &= q_n + \tfrac{h}{2} H_q(p_n, q_{n+1/2}) \\ p_{n+1} &= p_n - \tfrac{h}{2}\Big(H_p(p_n, q_{n+1/2}) + H_p(p_{n+1}, q_{n+1/2})\Big) \\ q_{n+1} &= q_{n+1/2} + \tfrac{h}{2} H_q(p_{n+1}, q_{n+1/2}) \end{aligned} \tag{3.61}$$

are symplectic methods of order 2.

Proof. This is an immediate consequence of the fact that the Störmer–Verlet scheme is the composition of the two symplectic Euler methods in (3.59). The order of the method can be verified as usual by Taylor expansion of the exact solution. □

The symplectic Euler methods in (3.59) and the Störmer-Verlet schemes in (3.60) and in (3.61) are in general implicit. However, for separable $H(p, q) = T(p) + U(q)$, all four variants turn out to be explicit.

Theorem 3.5.20: The **implicit midpoint rule**

$$y_{n+1} = y_n + hJ^{-1}\nabla H\left(\frac{y_{n+1} + y_n}{2}\right) \tag{3.62}$$

is a symplectic method of order 2.

Proof. The order follows from Theorem 3.4.15, since (3.62) is the 1-stage Gauss collocation method, and thus of order $2s = 2$. To show symplecticity we consider

$$F(y_0, \Phi_h(y_0)) := \Phi_h(y_0) - y_0 - hJ^{-1}\nabla H\left(\frac{\Phi_h(y_0) + y_0}{2}\right)$$

It follows from (3.62) that $F(y_0, \Phi_h(y_0)) = 0$ and thus, using the Implicit Function Theorem, that

$$D_{y_0}\Phi_h(y_0) = \left(I - J^{-1}B\right)^{-1}\left(I + J^{-1}B\right), \quad \text{where} \quad B := \frac{h}{2}\nabla^2 H\left(\frac{y_1 - y_0}{2}\right).$$

It remains to verify that $D_{y_0}\Phi_h(y_0)^\top J D_{y_0}\Phi_h(y_0) = J$. Recall that $J^{-1} = J^T = -J$. Thus, multiplying from left and right with $(I - JB)^{-T}$ and $(I - JB)^{-1}$, respectively, and inverting both sides we have

$$\begin{aligned} & \left(I + J^{-1}B\right)^T\left(I - J^{-1}B\right)^{-T}J\left(I - J^{-1}B\right)^{-1}\left(I + J^{-1}B\right) = J \\ \Leftrightarrow \quad & (I + JB)^{-T}J(I + JB)^{-1} = (I - JB)^{-T}J(I - JB)^{-1} \\ \Leftrightarrow \quad & (I + JB)J(I + JB)^T = (I - JB)J(I - JB)^T \\ \Leftrightarrow \quad & (I + JB)(J - JBJ) = (I - JB)(J + JBJ), \end{aligned}$$

which holds true. This completes the proof. □

Example 3.5.21. To demonstrate the effects of symplecticity of a numerical method, we consider again the pendulum problem of Example 3.5.12 with the same initial sets as in Fig. 3.6. We apply six different numerical methods to this problem: the explicit Euler method (2.2), the symplectic Euler method (3.59), and the implicit Euler method (3.10), which are all of first order, as well as a second-order explicit method by Runge, the Störmer–Verlet scheme (3.60), and the implicit midpoint rule (3.62). For two sets of initial values (p_0, q_0) we compute several steps with step size $h = \pi/4$ for the first order methods, and $h = \pi/3$ for the second order methods.

One clearly observes in Fig. 3.7 that the explicit Euler, the implicit Euler and the second order explicit method of Runge are not symplectic (not area preserving). As shown above, the other methods are symplectic. They are not exact, so the sets are slightly perturbed, but the area of the sets is preserved for all three methods.

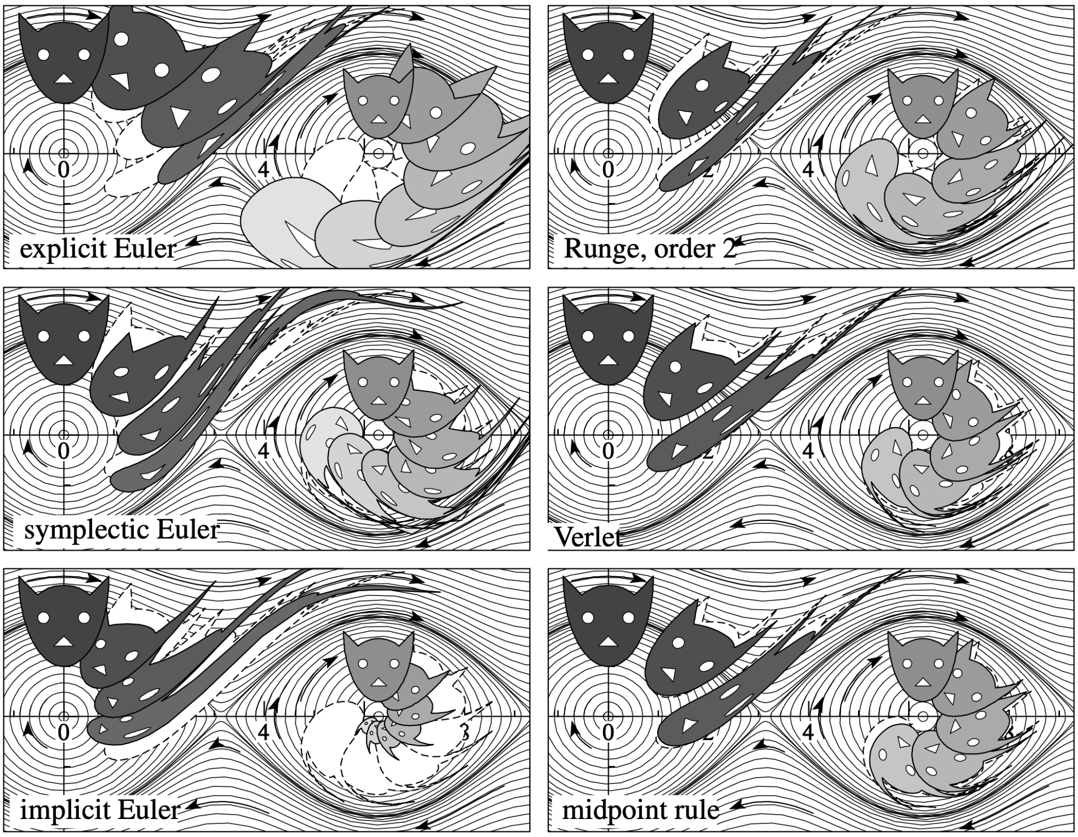


Figure 3.7: Area preservation of numerical methods for the pendulum problem using the initial sets of Fig. 3.6: first order methods (left column) with $h = \pi/4$; second order methods (right column) with $h = \pi/3$. The exact flow is shown with dashed lines.

We finish this section by stating two important results about higher-order symplectic Runge-Kutta methods, which we will not prove (see [Lub] for details).

Theorem 3.5.22: The Gauss collocation methods of Section 3.4.2 are symplectic.

Theorem 3.5.23: If the coefficients of a Runge–Kutta method satisfy

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \,, \quad \text{for all } \; i, j = 1, ..., s, \tag{3.63}$$

then it is symplectic.

Example 3.5.24. We can verify that indeed the symplectic Runge-Kutta methods we have already discussed all satisfy conditon (3.63). Clearly the explicit and implicit Euler methods do not satisfy (3.63) and, as we have seen in Example 3.5.21, they are not symplectic. Another method that is **not** symplectic is the Crank-Nicolson method, which has $a_{11} = a_{12} = 0$ and $a_{21} = a_{22} = b_1 = b_2 = 1/2$ and thus does not satisfy (3.63).

Symplectic integration is a rich and active research area, especially in the context of imaging – see, e.g., the research in Christoph Schnörr's and Stefania Petra's groups in Heidelberg. For more details see [Lub] or [LR05].

Chapter 4

Newton and quasi-Newton methods

4.1 Basics of nonlinear iterations

4.1.1. The efficient solution of nonlinear problems is an important ingredient to implicit timestepping schemes. Without attempting completeness, we present some important facts about iterative methods for this problem. We introduce the two generic schemes, Newton and gradient methods, discuss their respective pros and cons and combine their features in order to obtain better methods.

Consider the problem of finding $x \in \mathbb{R}^d$ such that

$$f(x) = 0, \qquad \text{for } \; f : \mathbb{R}^d \to \mathbb{R}^d. \tag{4.1}$$

Definition 4.1.2: An iteration

$$x^{(k+1)} = \Psi\bigl(x^{(k)}\bigr)$$

to find a fixpoint $x^* = \Psi(x^*)$ is said to be **convergent of order** $p \geq 1$ if

$$\|x^{(k+1)} - x^*\| \leq q \, \|x^{(k)} - x^*\|^p \, .$$

For $p = 1$, in addition we require that $q < 1$. In that case, q is called the **convergence rate**.

We have already seen in the proof of Lemma 3.3.9 that the fixpoint iteration, e.g. for the implicit Euler method:

$$y_1{}^{(m)} = \Psi\bigl(y_1{}^{(m-1)}\bigr) := y_0 + hf(t_1, y_1{}^{(m-1)}), \quad \text{with } \; y_1{}^{(0)} = y_0 \, ,$$

converges to y_1 provided $hL < 1$, but the convergence is only of order $p = 1$ (linear) and the convergence rate is $q := Lh$, which may be close to 1. Moreover, it may fail if $hL \geq 1$.

In Numerik 0, we have already seen a faster converging algorithm, the Newton method, and proved there that it converges with order $p = 2$, for sufficiently good initial guess.

Definition 4.1.3: The **Newton method** for finding the root of the nonlinear equation $f(x) = 0$ with $f : \mathbb{R}^d \to \mathbb{R}^d$ reads: given an initial value $x^{(0)} \in \mathbb{R}^d$, compute iterates $x^{(k)} \in \mathbb{R}^d$, $k = 1, 2, \dots$ as follows:

$$\begin{aligned} G\big(x^{(k)}\big) &= Df\big(x^{(k)}\big), \\ s^{(k)} &= -G\big(x^{(k)}\big)^{-1} f\big(x^{(k)}\big), \\ x^{(k+1)} &= x^{(k)} + s^{(k)}. \end{aligned} \tag{4.2}$$

We denote by the term **quasi-Newton method** any modification of this scheme employing an approximation $\widetilde{G}_k$ of the Jacobian $G\big(x^{(k)}\big)$.

Theorem 4.1.4: Let $U \subset \mathbb{R}^d$ and $f : U \to \mathbb{R}^d$ be continuously differentiable with

$$\|Df(x) - Df(y)\| \le L\|x - y\|, \quad \text{for all} \ \ x, y \in U \tag{4.3}$$

and for some $L > 0$. If there exists a $x^* \in U$ such that $f(x^*) = 0$ and

$$\|(Df\,(x^*))^{-1}\| \le M, \tag{4.4}$$

then there exists a $0 < R \le \frac{1}{2LM}$ such that for all $x^{(0)} \in \{x \in U : \|x^* - x\| \le R\}$, we have $x^{(k)} \to x^*$ with order $p = 2$.

Remark 4.1.5. The proof of this theorem can be found in the lecture notes for Numerik 0. There are also versions that do not require the existence of the root a priori, such as the Newton-Kantorovich Theorem [Ran17a, Satz 5.5], but we will only discuss some of the main assumptions and features.

The Lipschitz condition on Df can be seen as the deviation of f from being linear. Indeed, if f were linear, then $L = 0$ and provided $M \neq 0$ the method converges in a single step for any initial value.

The larger the constant M, the smaller one of the eigenvalues of the Jacobian $G = Df$. Therefore, the function becomes flat in the direction of the corresponding eigenvector and the root finding problem becomes unstable. In the limit as $M \to \infty$, the convergence reduces to linear order ($p = 1$).

Most importantly, for an arbitrary initial guess, the method may fail to converge entirely, but close enough to the solution the convergence is very fast, much faster than the fixpoint iteration above.

4.2 Descent methods

Nonlinear root finding of a vector-valued functions $f : \mathbb{R}^d \to \mathbb{R}^d$ – as required in implicit timestepping schemes – is closely related to optimisation of scalar functions $F : \mathbb{R}^d \to \mathbb{R}$, and the following problem is equivalent to (4.1) whenever $f = \nabla F$:

$$x = \operatorname*{arg\,min}_{y \in \mathbb{R}^d} F(y), \qquad \text{for} \ \ F : \mathbb{R}^d \to \mathbb{R}. \tag{4.5}$$

While we assume for most of this discussion that F is known, we will see at the end that the Newton method with line search does not require it.

Obvisously, by choosing $f = \nabla F$, Newton's method also solves the optimsation problem (4.5). An alternative family of methods for (4.5) are the following:

Definition 4.2.1: A **descent method** is an iterative method for finding minimizers of the functional $F : \mathbb{R}^d \to \mathbb{R}$ that, starting from an initial guess $x^{(0)} \in \mathbb{R}^d$, computes iterates $x^{(k)}$, $k = 1, 2 \dots$, by the following steps:

1. If $\nabla F(x^{(k)}) \neq 0$, choose a descent direction
$$s^{(k)} \in \mathbb{R}^d \quad \text{such that} \quad \left\langle \nabla F(x^{(k)}), s^{(k)} \right\rangle < 0 \tag{4.6}$$
and a positive parameter $\alpha^{(k)} > 0$; otherwise terminate.
2. Update: $\quad x^{(k+1)} = x^{(k)} + \alpha^{(k)} s^{(k)}$.

Lemma 4.2.2: Let $F : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable. For a given point x, assume $\nabla F(x) \neq 0$. Then, there is a constant $\vartheta > 0$ such that for any descent direction s satisfying (4.6) and for any stepsize $0 \leq \alpha \leq \vartheta$ there holds

$$F(x + \alpha s) \leq F(x) - \frac{\vartheta \alpha}{2} |\nabla F(x)|. \tag{4.7}$$

In particular, a positive scaling factor α for the descent method, and thus a strict decrease in the function value, can always be found.

Proof. Skipped. (See, e.g., [NW06].) □

The most prominent member of this family of methods is the following.

Definition 4.2.3: The **gradient method** for finding minimizers of $F(x)$ reads: Given an initial value $x^{(0)} \in \mathbb{R}^d$, compute iterates $x^{(k)}$, $k = 1, 2, \dots$ by the rule

$$\begin{aligned} s^{(k)} &= -\nabla F(x^{(k)}), \\ \alpha^{(k)} &= \operatorname*{argmin}_{\gamma > 0} F\left(x^{(k)} + \gamma s^{(k)}\right) \\ x^{(k+1)} &= x^{(k)} + \alpha^{(k)} s^{(k)}. \end{aligned} \tag{4.8}$$

It is also called the method of **steepest descent**.
The minimization process, called **line search**, to compute α_k is one-dimensional and therefore simple. It is sufficient to find an approximate minimum $\tilde{\alpha}^{(k)}$.

Theorem 4.2.4: Let $F(x) : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable and let $x^{(0)} \in \mathbb{R}^d$ be chosen such that the set

$$K = \{x \in \mathbb{R}^d \big| F(x) \le F(x^{(0)})\}$$

is compact. Then, each sequence defined by the gradient method has at least one accumulation point and each accumulation point is a stationary point of $F(x)$.

Proof. First, we observe that in any point $x^{(k)}$ with $\nabla F(x^{(k)}) \neq 0$, it follows from Lemma 4.2.2 that there exists $\gamma > 0$ such that

$$F(x^{(k)}) > F(x^{(k)} + \gamma s^{(k)}).$$

We conclude, that for such $x^{(k)}$, the line search obtains a positive value of $\alpha^{(k)}$. Thus, the sequence of the gradient iteration is monotonically decreasing and stays within the set K. Since K was assumed to be compact the sequence $x^{(k)}$ has at least one accumulation point x^*. However, the preceding discussion implies that $\nabla F(x^*) = 0$. □

Remark 4.2.5. However, we can also choose $s^{(k)} = -B_k^{-1}\nabla F(x^{(k)})$ in (4.8), for any positive definite matrix B_k, leading to **generalised steepest descent methods** that minimise the descent direction in (4.6) in the weighted inner product $\langle x, y\rangle_{B_k} = x^T B_k y$ instead of the Euclidean inner product $\langle x, y\rangle = x^T y$.

In particular, if the Hessian $\nabla^2 F(x^{(k)})$ is positive definite we can choose $B_k = \nabla^2 F(x^{(k)})$ and $\alpha^{(k)} = 1$, which reduces to the Newton method. This link is derived ina different way in the next section.

In fact, we also have the following result.

Lemma 4.2.6: Under the assumptions of Theorem 4.1.4, Newton's Method applied to the root finding problem $f(x) = 0$ for $f : \mathbb{R}^d \to \mathbb{R}^d$ is a descent method applied to the functional $F(x) = |f(x)|^2$.

Proof. The (multivariate) product rule gives

$$\nabla F(x) = 2\big(Df(x)\big)^{\top} f(x)$$

The search direction of the Newton method applied to $f(x)$ is

$$s^{(k)} = -\big(Df(x^{(k)})\big)^{-1} f(x^{(k)})$$

Now, assuming that $f(x^{(k)}) \neq 0$, we have

$$\Big\langle \nabla F, s^{(k)} \Big\rangle = -2f(x^{(k)})^T Df(x^{(k)})\big(Df(x^{(k)})\big)^{-1} f(x^{(k)}) = -2|f(x^{(k)})|^2 < 0,$$

and thus $s^{(k)}$ is a descent direction that satisfies (4.6). Here, we used the fact that for $x^{(k)}$ sufficiently close to x^* we have $\|\big(\nabla f(x^{(k)})\big)^{-1}\| < \infty$. (It follows from the Perturbation Theorem in Numerik 0.) Finally, choosing $\alpha^{(k)} := 1$ we have established the equivalence. □

4.3 Globalization of Newton

4.3.1. The convergence of the Newton method is only local, and it is the faster, the closer to the solution we start. Thus, finding good initial guesses is an important task.

A reasonable initial guess for finding y_1 in a one-step method seems to be y_0, but on closer inspection, this is true only if the time step is small. The convergence requirements of Newton's method would insert a new time step restriction, which we want to avoid. Therefore, we present methods which guarantee global convergence while still converging locally quadratically.

As a rule, Newton's method should never be used without some globalization strategy!

Definition 4.3.2: The **Newton method with line search** for finding the root of the nonlinear equation $f(x) = 0$ reads: given an initial value $x^{(0)} \in \mathbb{R}^d$, compute iterates $x^{(k)} \in \mathbb{R}^d$, $k = 1, 2, \dots$ by the rule

$$\begin{aligned} G\big(x^{(k)}\big) &= Df\big(x^{(k)}\big), \\ s^{(k)} &= -G\big(x^{(k)}\big)^{-1} f\big(x^{(k)}\big), \\ \alpha^{(k)} &= \operatorname*{argmin}_{\gamma>0} \left| f(x^{(k)} + \gamma s^{(k)}) \right|^2 \\ x^{(k+1)} &= x^{(k)} + \alpha^{(k)} s^{(k)}. \end{aligned} \tag{4.9}$$

Definition 4.3.3: A practically most often used variant is the **Newton method with step size control (backtracking line search)**: given an initial value $x^{(0)} \in \mathbb{R}^d$, compute iterates $x^{(k)} \in \mathbb{R}^d$, $k = 1, 2, \dots$ by the rule

$$\begin{aligned} G\big(x^{(k)}\big) &= Df\big(x^{(k)}\big), \\ s^{(k)} &= -G\big(x^{(k)}\big)^{-1} f\big(x^{(k)}\big), \\ x^{(k+1)} &= x^{(k)} + 2^{-j} s^{(k)}. \end{aligned} \tag{4.10}$$

Here, $j \in \mathbb{N}_0$ is the smallest non-negative integer, such that

$$|f(x^{(k)} + 2^{-j} s^{(k)})| < |f(x^{(k)})| . \tag{4.11}$$

Remark 4.3.4. The step size control algorithm can be implemented with very low overhead. In fact, in each Newton step we only have to monitor the norm of the residual $|f(x^{(k)} + s^{(k)})|$, which is typically needed for the stopping criterion anyway. If the residual grows, i.e. $|f(x^{(k)} + s^{(k)})| \geq |f(x^{(k)}|$, we halve the stepsize, recompute the residual norm and check again. A modification to the plain Newton method and additional work are only needed, when the original method was likely to fail anyway.

Under certain assumptions on f it can be shown [NW06] that this backtracking line search algorithm terminates after a finite (typically very small) number of steps. Also, the step size control typically only triggers within the first few steps. After that, condition (4.11) holds for $j = 0$ and the quadratic convergence of the Newton method kicks in.

4.4 Practical considerations – quasi-Newton methods

4.4.1. Quadratic convergence is an asymptotic statement, which for any practical purpose can be replaced by "fast" convergence. Most of the effort spent in a single Newton step consists in setting up the Jacobian $G = Df$ and solving the linear system in the second line of (4.2). Therefore, we will consider techniques here, which avoid some of this work. We will have to consider two cases

1. Small systems with $d \lesssim 1000$. For such systems, a direct method like LU- or QR-decomposition is advisable in order to solve the linear system. To this end, we compute the whole Jacobian and compute its decomposition, an effort of order d^3 operations. Comparing to d^2 operations for applying the inverse and order d for all other tasks, this must be avoided as much as possible.
2. Large systems, where the Jacobian is typically sparse (most of its entries are zero). For such a system, factorising the matrix at a cost of order d^3 is typically not affordable. Therefore, the linear problem is solved by an iterative method and we avoid the computation of the Jacobian when possible.

Remark 4.4.2. In order to save numerical effort constructing and inverting Jacobians, the following strategies have been successful:

(i) Fix a threshold $0 < \eta < 1$ which will be used as a bound for error reduction. In each Newton step, first compute the update vector $d^{(k)}$ using the approximation $\widetilde{G}_{k-1}$ of the Jacobian from the previous step. This yields the modified method

$$\begin{aligned}
&& \widetilde{G}_k &= \widetilde{G}_{k-1} \\
&& \widehat{x} &= x^{(k)} - \widetilde{G}_k^{-1} f(x^{(k)}) \\
\text{If } |f(\widehat{x})| \le \eta |f(x^{(k)})| && x^{(k+1)} &= \widehat{x} \\
\text{Else } \widetilde{G}_k = Df\big(x^{(k)}\big) && x^{(k+1)} &= x^{(k)} - \widetilde{G}_k^{-1} f(x^{(k)}).
\end{aligned} \tag{4.12}$$

Thus, an old Jacobian and its inverse are used until convergence rates deteriorate. This method is a quasi-Newton method which will not converge quadratically. However, we can obtain linear convergence at any rate η.

(ii) If Newton's method is used within a time stepping scheme, the Jacobian of the last Newton step in the previous time step is often a good approximation for the Jacobian of the first Newton step in the new time step. This holds in particular for small time steps and constant extrapolation. Therefore, the previous method should also be extended over the bounds of time steps.

(iii) An improvement of the method above can be achieved by so called low rank updates, e.g. for the rank-1 update: Let $\widetilde{G}_0 = Df(x^{(0)})$ or $\widetilde{G}_0 = I$. Then, at the kth step, given $x^{(k)}$ and $x^{(k-1)}$, compute

$$\begin{aligned}
s &= x^{(k)} - x^{(k-1)} \\
q &= f(x^{(k)}) - f(x^{(k-1)}) \\
\widetilde{G}_k &= \widetilde{G}_{k-1} + \frac{1}{|s|^2}\left(q - \widetilde{G}_{k-1} s\right) s^T
\end{aligned} \tag{4.13}$$

The fact that the rank of $\widetilde{G}_k - \widetilde{G}_{k-1}$ is at most one can be used to avoid computing and storing matrices at all. The inverse of such a matrix can be computed via the Sherman-Morrison formula.

The practically most efficient and used methods use rank-2 updates, such as the Broyden methods [NW06].

Remark 4.4.3. For problems leading to large, sparse Jacobians (e.g., space discretizations of PDEs), it is often too costly or infeasible to compute inverses or LU-decompositions. These matrices typically only feature a few nonzero elements per row, while the inverse and the LU-decomposition are fully populated, thus increasing the amount of memory from $O(d)$ to $O(d^2)$.

Linear systems like this are often solved by iterative methods, leading for instance to so called Newton-Krylov methods. In an iterative method, the solution of a linear system

$$Gs = b$$

is approximated using only multiplications of a vector with the matrix G. On the other hand, for any vector $v \in \mathbb{R}^d$, the term Gv denotes the directional derivative of f in direction Gv. Thus, it can be approximated easily by

$$Gv \approx \frac{f\left(x^{(k)} + \varepsilon v\right) - f\left(x^{(k)}\right)}{\varepsilon}.$$

The term $f\left(x^{(k)}\right)$ must be calculated anyway as it is the current Newton residual. Thus, each step of the iterative linear solver requires one evaluation of the nonlinear function, and no derivatives are computed.

The efficiency of such a method depends on the number of linear iteration steps which is determined by two factors: the gain in accuracy and the contraction speed. It turns out that typically gaining two digits in accuracy is sufficient to ensure fast convergence of the Newton iteration. The contraction number is a more difficult issue and typically requires preconditioning, which is problem-dependent and as such must be discussed when needed.

Chapter 5

Linear Multistep Methods

Instead of using only the *one* initial value at the beginning of the current time interval to the next time step, possibly with the help of intermediate steps, we can also use the values from several previous time steps. Intuitively, this could be more efficient, since function values at these points have been computed already.

Such methods that use values of several time steps in the past in order to achieve a higher order are called **multistep methods**. We will begin this chapter by introducing some of the common formulae, before studying their stability and convergence properties.

5.1 Examples of LMMs

Basically, there are two construction principles for the multistep methods: Quadrature and numerical differentiation.

Example 5.1.1 (Adams-Moulton formulae)**.** Here, the integral from point t_{k-1} to point t_k is approximated by an interpolatory quadrature rule based on the points t_{k-s} to t_k, i.e.,

$$y_k = y_{k-1} + \sum_{r=0}^{s} f_{k-r} \int_{t_{k-1}}^{t_k} L_r^{(s)}(t)\,\mathrm{d}t, \tag{5.1}$$

where f_j denotes the function value $f(t_j, y_j)$ and $L_r^{(s)}(t)$, $r = 0, \dots, s$, the Lagrange interpolation polynomials associated with the $s+1$ points t_{k-r}, $r = 0, \dots, s$.

Since the integral involves the function evaluated at the time step that is being computed, these methods are implicit. Here are the first four in this family:

$$\begin{aligned}
y_k &= y_{k-1} + hf_k && \text{(implicit Euler)}\\
y_k &= y_{k-1} + \frac{1}{2}h\big(f_k + f_{k-1}\big) && \text{(trapezoidal rule)}\\
y_k &= y_{k-1} + \frac{1}{12}h\big(5f_k + 8f_{k-1} - f_{k-2}\big)\\
y_k &= y_{k-1} + \frac{1}{24}h\big(9f_k + 19f_{k-1} - 5f_{k-2} + f_{k-3}\big)
\end{aligned}$$

Example 5.1.2 (Adams-Bashforth formulae)**.** With the same principle we obtain explicit methods by omitting the point in time t_k in the definition of the interpolation polynomial. This yields quadrature formulae of the form

$$y_k = y_{k-1} + \sum_{r=1}^{s} f_{k-r} \int_{t_{k-1}}^{t_k} L_r^{(s-1)}(t)\,\mathrm{d}t. \tag{5.2}$$

Here, $L_r^{(s-1)}(t)$ are the Lagrange interpolation polynomials associated with the s points t_{k-r}, $r = 1, \dots, s$. Again, we list the first few:

$$\begin{aligned}
y_k &= y_{k-1} + hf_{k-1} && \text{(explicit Euler)}\\
y_k &= y_{k-1} + \frac{1}{2}h\big(3f_{k-1} - 1f_{k-2}\big)\\
y_k &= y_{k-1} + \frac{1}{12}h\big(23f_{k-1} - 16f_{k-2} + 5f_{k-3}\big)\\
y_k &= y_{k-1} + \frac{1}{24}h\big(55f_{k-1} - 59f_{k-2} + 37f_{k-3} - 9f_{k-4}\big)
\end{aligned}$$

Example 5.1.3 (BDF methods)**.** Backward differencing formulas (BDF) are also based on Lagrange interpolation at the points t_{k-s} to t_k. However, in contrast to Adams formulae they do not use quadrature for the right hand side, but rather the derivative of the interpolation polynomial in the point t_k for the left hand side.

Using the Lagrange interpolation polynomials $L_{k-r}^{(s)}(t)$, we let

$$y(t) = \sum_{r=0}^{s} y_{k-r} L_{k-r}^{(s)}(t),$$

where y_k is yet to be determined. Now we assume that $y(t)$ satisfies the ODE at t_k. Thus,

$$\sum_{r=0}^{s} y_{k-r} \frac{\mathrm{d}L_{k-r}^{(s)}}{\mathrm{d}t}(t_k) = y'(t_k) = f(t_k, y_k)$$

leading to the following schemes:

$$\begin{aligned}
y_k - y_{k-1} &= hf_k && \text{(implicit Euler)}\\
y_k - \tfrac{4}{3}y_{k-1} + \tfrac{1}{3}y_{k-2} &= \tfrac{2}{3}hf_k\\
y_k - \tfrac{18}{11}y_{k-1} + \tfrac{9}{11}y_{k-2} - \tfrac{2}{11}y_{k-3} &= \tfrac{6}{11}hf_k\\
y_k - \tfrac{48}{25}y_{k-1} + \tfrac{36}{25}y_{k-2} - \tfrac{16}{25}y_{k-3} + \tfrac{3}{25}y_{k-4} &= \tfrac{12}{25}hf_k
\end{aligned}$$

For an example on how to derive these schemes see the appendix.

Remark 5.1.4. Recall from Numerik 0 (or any other introductory course in numerical analysis) that numerical differentiation and extrapolation of interpolation polynomials (i.e. the evaluation outside the interval which is spanned through the interpolation points) are both unstable numerically. Therefore, we expect stability problems for all these methods.

Secondly, recall that Lagrange interpolation with equidistant support points is unstable for higher degree polynomials. Therefore, we also expect all of the above methods to perform well only at moderate order.

5.2 General definition and consistency of LMMs

Definition 5.2.1: A **linear multistep method** (LMM) with s steps is a method of the form

$$\sum_{r=0}^{s} \alpha_{s-r} y_{k-r} = h \sum_{r=0}^{s} \beta_{s-r} f_{k-r}, \tag{5.3}$$

where $f_k = f(t_k, y_k)$ and $t_k = t_0 + hk$, and where we assume $|\alpha_0| + |\beta_0| \neq 0$ and $\alpha_s = 1$. There are explicit ($\beta_s = 0$) and implicit ($\beta_s \neq 0$) methods.
It is convenient to define the **generating polynomials**

$$\varrho(x) = \sum_{r=0}^{s} \alpha_{s-r} x^{s-r} = \sum_{j=0}^{s} \alpha_j x^j \qquad \sigma(x) = \sum_{r=0}^{s} \beta_{s-r} x^{s-r} = \sum_{j=0}^{s} \beta_j x^j. \tag{5.4}$$

for each of these methods.

Remark 5.2.2. The LMM was defined for constant step size h. In principle it is possible to implement the method with a variable step size but we restrict ourselves to the constant case. Notes to the step size control can be found later on in this chapter.

Definition 5.2.3: As for one-step methods, we use the abbreviation $u_k := u(t_k)$, where $u(t)$ denotes the exact solution of $u' = f(t, u),\ u(t_0) = u_0$.
The **local error** of a linear multistep method (LMM) at the kth timestep is again defined by

$$u_k - y_k\,,$$

where y_k is the numerical solution obtained from (5.3) using the exact initial values $y_{k-r} = u_{k-r}$ for $r = 1, ..., s$.
The **truncation error** of an LMM, on the other hand, is defined as

$$\tau_k(u) := h^{-1}(L_h u)(t_k), \tag{5.5}$$

using the linear **difference operator**

$$(L_h u)(t_k) := \sum_{r=0}^{s} \Big(\alpha_{s-r} u_{k-r} - h \beta_{s-r} f\big(t_{k-r}, u_{k-r}\big) \Big). \tag{5.6}$$

Lemma 5.2.4. *For h sufficiently small, the two local errors satisfy the following relation*

$$u_k - y_k = \left(\mathbb{I} - h\beta_s \overline{Df}_k\right)^{-1} (L_h u)(t_k), \tag{5.7}$$

where

$$\overline{Df}_k := \int_0^1 Df\big(t_k, u_k + \vartheta(y_k - u_k)\big)\, d\vartheta.$$

and $Df(t, y)$ is the Jacobian of f with respect to the second argument.

Proof. Sinc we assumed $y_{k-r} = u_{k-r}$, for $r = 1, ..., s$, in the definition of the local error and $\alpha_s = 1$, (5.3) is equivalent to

$$0 = y_k - hf(t_k, y_k) + \sum_{r=1}^{s} \alpha_{s-r} u_{k-r} - h \sum_{r=1}^{s} \beta_{s-r} f(t_{k-r}, u_{k-r}).$$

Subtracting this from (5.6), we obtain

$$(L_h u)(t_k) = (u_k - y_k) - h\beta_s \Big(f(t_k, u_k) - f(t_k, y_k) \Big).$$

Finally, the result follows by applying the Integral Mean Value Theorem (see, e.g., [Numerik 0, Hilfssatz 5.8]) and the fact that for h sufficiently small $\mathbb{I} - h\beta_s \overline{\mathrm{D}f}_k$ is invertible. □

Remark 5.2.5. Note that it follows from lemma 5.2.4 that

$$u_k - y_k = \Big(h + \mathcal{O}(h^2) \Big) \tau_k(u)$$

and that the higher-order term is exactly zero for explicit LMMs.

Definition 5.2.6: An LMM is **consistent of order** p, if for all sufficiently regular functions f and for all relevant k there holds

$$|\tau_k(u)| = \mathcal{O}(h^p), \tag{5.8}$$

or equivalently, that the local error is $\mathcal{O}(h^{p+1})$.

Theorem 5.2.7: A LMM with constant step size h is consistent of order p if and only if

$$\sum_{r=0}^{s} \alpha_{s-r} = 0 \quad \text{and} \quad \sum_{r=0}^{s} \left(\alpha_{s-r} r^q + q\beta_{s-r} r^{q-1} \right) = 0, \qquad q = 1, \ldots, p \tag{5.9}$$

Proof. We start with the Taylor expansion of the ODE solution u around t_k:

$$u(t) = \sum_{q=0}^{p} \frac{u^{(q)}(t_k)}{q!} (t - t_k)^q + \underbrace{\frac{u^{(p+1)}(\xi)}{(p+1)!} (t - t_k)^{p+1}}_{=: R_u(t)},$$

where ξ is a point between t and t_k that depends on t. It follows from $f(t, u) = u'$ that the corresponding right hand side can be expanded as

$$f\big(t, u(t)\big) = \sum_{q=1}^{p} \frac{u^{(q)}(t_k)}{(q-1)!} (t - t_k)^{q-1} + \underbrace{\frac{u^{(p+1)}(\eta)}{p!} (t - t_k)^p}_{=: R_f(t)}.$$

with η again a point between t and t_k that depends on t.

Substituting the two expansions into (5.6) we get:

$$\begin{aligned}
L_h u(t_k) = {} & \sum_{r=0}^{s} \alpha_{s-r} \left(\sum_{q=0}^{p} \frac{u^{(q)}(t_k)}{q!} (-rh)^q + R_u(t_{k-r}) \right) - \\
& - \beta_{s-r} h \left(\sum_{q=1}^{p} \frac{u^{(q)}(t_k)}{(q-1)!} (-rh)^{q-1} + R_f(t_{k-r}) \right) \\
= {} & u(t_k) \left(\sum_{r=0}^{s} \alpha_{s-r} \right) + \sum_{q=1}^{p} \frac{u^{(q)}(t_k)}{q!} (-1)^q \left(\sum_{r=0}^{s} \alpha_{s-r} r^q + q \beta_{s-r} r^{q-1} \right) h^q \; + \; C h^{p+1},
\end{aligned}$$

where

$$C := \sum_{r=0}^{s} \frac{(-1)^{p+1} r^p}{(p+1)!} \Big(\alpha_{s-r} r \, u^{(p+1)}(\xi_r) + (p+1) \beta_{s-r} \, u^{(p+1)}(\eta_r) \Big)$$

and $\xi_r, \eta_r \in [t_{k-r}, t_k]$, for $r = 0, \dots, s$, which in general may all be different.

Since the right hand side f was arbitrary, $L_h u(t_k) = \mathcal{O}(h^{p+1})$ if and only if the conditions in (5.9) hold. In that case we have

$$|L_h u(t_k)| \le \left(\frac{\|u^{(p+1)}\|_\infty}{(p+1)!} \left(\sum_{r=0}^{s} \alpha_{s-r} r^{p+1} + (p+1) \beta_{s-r} r^p \right) \right) h^{p+1}.$$

□

Remark 5.2.8. A consistent LMM is not necessarily convergent. To understand this and to develop criteria for convergence we diverge into the theory of difference equations.

5.3 Properties of difference equations

5.3.1. The stability of LMM can be understood by employing the fairly old theory of difference equations. In order to keep the presentation simple in this section, we use a different notation for numbering indices in the equations. Nevertheless, the coefficients of the characteristic polynomial are the same as for LMM.

Definition 5.3.2: An equation of the form

$$\sum_{r=0}^{s} \alpha_r y_{n+r} = 0 \tag{5.10}$$

is called a homogeneous **difference equation**. Assume that $\alpha_s \alpha_0 \neq 0$ (such that (5.10) does not reduce to a lower order difference equation). A sequence $(y_n)_{n \ge 0}$ is solution of the difference equation, if the equation holds true for all $n \ge s$. The values y_n may be from any of the spaces $\mathbb{R}$, $\mathbb{C}$, $\mathbb{R}^d$ or $\mathbb{C}^d$.
The **generating polynomial** of this difference equation is

$$\chi(x) = \sum_{r=0}^{s} \alpha_r x^r. \tag{5.11}$$

Lemma 5.3.3: The solutions of equation (5.10) form a vector space of dimension s.

Proof. Since the equation (5.10) is linear and homogeneous, it is obvious that if two sequences of solutions $(y_n)_{n\geq 0}$ and $(\tilde{y}_n)_{n\geq 0}$ satisfy the equation, then $(\alpha y_n + \tilde{y}_n)_{n\geq 0}$ also satisfies (5.10), for any $\alpha \in \mathbb{R}$ (or $\mathbb{C}$).

As soon as the initial values y_0 to y_{s-1} are chosen, all other sequence members are uniquely defined. Moreover it holds

$$y_0 = y_1 = \cdots = y_{s-1} = 0 \quad \Longrightarrow \quad y_n = 0,\ n \geq 0.$$

Therefore it is sufficient to consider the first s elements. Since those can be chosen arbitrarily, they span an s-simensional vector space. □

Lemma 5.3.4: For each root ξ of the generating polynomial $\chi(x)$ the sequence $y_n = \xi^n$ is a solution of the difference equation (5.10).

Proof. Inserting the solution $y_n = \xi^n$ into the difference equation results in

$$\sum_{r=0}^{s} \alpha_r \xi^{n+r} = \xi^n \sum_{r=0}^{s} \alpha_r \xi^r = \xi^n \chi(\xi) = 0.$$

□

Theorem 5.3.5: Let $\{\xi_j\}_{j=1,\dots,J}$ be the roots of the generating polynomial χ with multiplicity μ_j. Then, the sequences of the form

$$y_n^{(j,k)} = n^{k-1}\xi_j^n \quad j = 1,\dots,J; \quad k = 1,\dots,\mu_j \tag{5.12}$$

form a basis of the solution space of the difference equation (5.10).

Proof. First we observe that the sum of the multiplicities of the roots has to result in the degree of the polynomial:

$$s = \sum_{j=1}^{J} \mu_j\,.$$

Moreover, we know from Lemma 5.3.3, that s is the dimension of the solution space. However, the sequences $\big(y_n^{(j,k)}\big)_{n\geq 0}$ are also linearly independent. This is clear for sequences of different index j. It can also be easily shown for multiple roots. (Please check this yourself.)

It remains to show that the sequences $\big(y_n^{(j,k)}\big)_{n\geq 0}$ are in fact solutions of the difference equations. For $k = 1$ we have proven this already in lemma 5.3.4. We prove the fact here for $k = 2$ and for a double zero ξ_j; the principle for higher order roots should be clear then. Equation (5.10) applied to the sequence $(n\xi_j^n)_{n\geq 0}$ results in

$$\begin{aligned}\sum_{r=0}^{s} \alpha_r (n+r)\xi_j^{n+r} &= n\xi_j^n \sum_{r=0}^{s} \alpha_r \xi_j^r + \xi_j^{n+1} \sum_{r=1}^{s} \alpha_r r \xi_j^{r-1} \\ &= n\xi_j^n \chi(\xi_j) + \xi_j^{n+1}\chi'(\xi_j) = 0.\end{aligned}$$

Here, the term with α_0 vanishes, because it is multiplied with $r = 0$. Moreover, $\chi(\xi_j) = \chi'(\xi_j) = 0$, because ξ_j is a double root. □

Corollary 5.3.6 (Root test): All solutions $\big(y_n\big)_{n\geq 0}$ of the difference equation (5.10) are bounded for $n \to \infty$ if and only if:

- all roots of the generating polynomial $\chi(x)$ lie in $\big\{z \in \mathbb{C} \;\big|\; |z| \leq 1\big\}$ (closed unit circle) and
- all roots on the boundary of the unit circle are simple.

Proof. According to theorem 5.3.5 we can write all solutions as linear combinations of the sequences $(y^{(j,k)})$ in equation (5.12). Therefore, for $n \to \infty$,

1. all solutions with $|\xi_i| < 1$ converge to zero
2. all solutions with $|\xi_i| > 1$ diverge to infinity
3. all solutions with $|\xi_i| = 1$ stay bounded if and only if ξ_i is simple.

This proves the statement of the theorem. □

5.4 Stability and convergence

In contrast to one-step methods, the Lipschitz condition (1.24) for the RHS f of the differential equation is not sufficient to ensure that consistency of a multistep method implies convergence. As for A-stability, stability conditions will again be deduced by means of a simple model problem.

Remark 5.4.1. In the following we investigate the solution to a fixed point problem in time t with a shrinking step size h. Therefore we choose n steps of step size $h = t/n$ and let n go towards infinity.

Definition 5.4.2: An LMM is **zero-stable** (or simply **stable**) if, applied to the trivial ODE

$$u' = 0 \tag{5.13}$$

with arbitrary initial values $y_0 = u_0$ to $y_{s-1} = u_{s-1}$, it generates solutions y_k which stay bounded at each point in time $t > 0$, if the step size h converges to zero.

Theorem 5.4.3: A LMM is zero-stable if and only if all roots of the first generating polynomial $\varrho(x)$ of equation (5.4) satisfy the root test in corollary 5.3.6.

Proof. The application of the LMM to the ODE (5.13) results in the difference equation

$$\sum_{r=0}^{s} \alpha_{s-r} y_{k-r} = \sum_{j=0}^{s} \alpha_j y_{n+j} = 0 \tag{5.14}$$

with $n = k - s$. Thus, the generating polynomial $\varrho(x)$ is equivalent to the generating polynomial $\chi(x)$ of the difference equation in (5.10) which is independent of h.

If $\alpha_s\alpha_0 \neq 0$, the result than follows directly from corollary 5.3.6. Otherwise, (5.14) reduces to a finite difference equation with generating polynomial $\varrho_m(x)$ of order $s-m$, for some $1 \leq m \leq s-1$, and $\varrho(x) = x^m\varrho_m(x)$. Thus, ϱ satisfying the root test is equivalent to ϱ_m satisfying the root test and the result follows again from corollary 5.3.6. □

Corollary 5.4.4: Adams-Bashforth and Adams-Moulton methods are zero-stable.

Proof. For all of these methods the first generating polynomial is $\varrho(x) = x^s - x^{s-1}$. It has the simple root $\xi_1 = 1$ and the $s-1$-fold root 0. □

Theorem 5.4.5: The BDF methods are zero-stable for $s \leq 6$ and not zero-stable for $s \geq 7$.

Proof. See [HNW09, Theorem 3.4]. □

Definition 5.4.6: An LMM is **convergent of order** p, if for any IVP with sufficiently smooth right hand side f there exists a constant $h_0 > 0$ such that, for all $h \leq h_0$,

$$|u(t_n) - y(t_n)| \leq ch^p, \quad n \geq 0, \tag{5.15}$$

whenever the initial values satisfy

$$|u(t_i) - y(t_i)| \leq c_0h^p, \quad i = 0, \dots, s-1. \tag{5.16}$$

Here, u is the continuous solution of the IVP and y is the LMM approximation.

To prove convergence, we will for simplicity only consider the scalar case $d = 1$. The case $d > 1$ can be proved similarly.

Lemma 5.4.7: Let $d = 1$. Every LMM can be recast as a one-step method

$$Y_k = AY_{k-1} + hF_h(t_{k-1}, Y_{k-1}) \tag{5.17}$$

where

$$Y_k = \begin{pmatrix} y_k \\ \vdots \\ y_{k-s+1} \end{pmatrix} \in \mathbb{R}^s, \quad A = \begin{pmatrix} -\alpha_{s-1} & -\alpha_{s-2} & \cdots & -\alpha_0 \\ 1 & 0 & \cdots & 0 \\ & \ddots & \cdots & 0 \\ & & 1 & 0 \end{pmatrix} \in \mathbb{R}^{s\times s}, \tag{5.18}$$

and $F_h(t_k, Y_k) = (\psi_k, 0, \dots, 0)^T \in \mathbb{R}^s$ with ψ_k implicitly defined as the solution of

$$\psi_k = \sum_{r=1}^{s} \beta_{s-r} f(t_{k-r}, y_{k-r}) + \beta_s f\left(t_k, h\psi_k - \sum_{r=1}^{s} \alpha_{s-r} y_{k-r}\right). \tag{5.19}$$

Proof. From the general form of LMM we obtain

$$\sum_{r=0}^{s} \alpha_{s-r} y_{k-r} = h \sum_{r=0}^{s-1} \beta_{s-r} f(t_{k-r}, y_{k-r}) + h\beta_s f(t_k, y_k).$$

We rewrite this to

$$y_k = -\sum_{r=1}^{s} \alpha_{s-r} y_{k-r} + h\psi_k,$$

where this formula is also entered implicitly as the value for y_k in the computation of $f(t_k, y_k)$. This is the first equation in (5.17). The remaining equations are simply shifting the entries in the vector Y_{k-1}, i.e. $(Y_k)_{i+1} = (Y_{k-1})_i = y_{k-i}$, for $i = 1, \dots, s-1$. □

Lemma 5.4.8: Let $d = 1$ and let $u(t)$ be the exact solution of the IVP. Suppose $\widehat{Y}_k$ is the solution of a single step

$$\widehat{Y}_k = AU_{k-1} + hF_h(t_{k-1}U_{k-1}),$$

with correct initial values $U_{k-1} = (u_{k-1}, u_{k-2}, \dots, u_{k-s})^T$.
If the multistep method is consistent of order p and f is sufficiently smooth, then there exist constants $h_0 > 0$ and $M \ge 0$ such that for $h \le h_0$ there holds

$$|U_k - \widehat{Y}_k| \le Mh^{p+1}. \tag{5.20}$$

Proof. The first component of $U_k - \widehat{Y}_k$ is the local error $u_k - y_k$ of step k, as defined in definition 5.2.3, which is of order h^{p+1} by the assumption. The other components vanish by the definition of the method. □

Lemma 5.4.9: Assume that an LMM is zero-stable. Then, there exists a vector norm $\|\cdot\|$ on $\mathbb{C}^s$ such that the induced operator norm of the matrix A satisfies

$$\|A\| \le 1. \tag{5.21}$$

Proof. We notice that $\varrho(x) = \sum \alpha_{s-r} x^r$ is the characteristic polynomial of the matrix A.

By the root test we know that simple roots, which correspond to irreducible blocks of dimension one have maximal modulus one. Furthermore, every Jordan block of dimension greater than one corresponds to a multiple root, which by assumption has modulus strictly less than one. Let ξ_i be such a multiple root with multiplicity μ_i. Such a block admits a modified canonical form

$$J_i = \begin{pmatrix} \lambda_i & 1-|\lambda_i| & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1-|\lambda_i| \\ & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{\mu_i \times \mu_i}.$$

Thus, the canonical form $J = T^{-1}AT$ has norm $\|J\|_\infty \le 1$. If we define the vector norm

$$\|x\| = \|T^{-1}x\|_\infty,$$

it follows that

$$\|Ax\| = \|T^{-1}Ax\|_\infty = \|JT^{-1}x\|_\infty \le \|T^{-1}x\|_\infty = \|x\|.$$

□

Theorem 5.4.10: Let f be sufficiently smooth. If a linear multi-step method is zero-stable and consistent of order p, then it is convergent of order p.

Proof. As already stated, we only prove the case $d = 1$ explicitly. See the original notes by Guido Kanschat for the general proof. Since f was assumed to be sufficiently smooth, F_h satsifies a uniform Lipschitz condition with Lipschitz constant L.

We reduce the proof to the convergence of the one-step method

$$Y_k = AY_{k-1} + hF_h(t_{k-1}, Y_{k-1}) =: G(Y_{k-1}). \tag{5.22}$$

Let Y_{k-1} and Z_{k-1} be two initial values for the interval I_k. By the previous lemma, we have in the norm defined there, for sufficiently small h, that

$$\|G(Y_{k-1}) - G(Z_{k-1})\| \le (1 + hL)\|Y_{k-1} - Z_{k-1}\|. \tag{5.23}$$

By lemma 5.4.8, the local error $\eta_k = \|U_k - \widehat{Y}_k\|$ at step k is bounded by Mh^{p+1} (where M also contains the equivalence constant γ between the Euclidean norm and the norm defined in the previous lemma). Thus:

$$\begin{aligned}
\|U_1 - Y_1\| &\le (1 + hL)\|U_0 - Y_0\| + Mh^{p+1} \\
\|U_2 - Y_2\| &\le (1 + hL)^2\|U_0 - Y_0\| + Mh^{p+1}\big(1 + (1 + hL)\big) \\
\|U_3 - Y_3\| &\le (1 + hL)^3\|U_0 - Y_0\| + Mh^{p+1}\Big(1 + (1 + hL) + (1 + hL)^2\Big) \\
&\vdots \\
\|U_n - Y_n\| &\le (1 + hL)^n\|U_0 - Y_0\| + Mh^{p+1}\Big(1 + (1 + hL) + \ldots + (1 + hL)^n\Big) \\
&\le e^{nhL}\|U_0 - Y_0\| + \frac{Mh^p}{L}\big(e^{nhL} - 1\big) \le Ch^p
\end{aligned}$$

where we recall that $U_n = u(t_n)$ and $t_n = t_0 + T$ where $T = nh$ and where

$$C := c_0\gamma e^{TL} + \frac{M}{L}(e^{TL} - 1)$$

with c_0 as defined in (5.16). □

5.5 Starting procedures

5.5.1. In contrast to one-step methods, where the numerical solution is obtained solely from the differential equation and the initial value, multistep methods require more than one start value. An LMM with s steps requires s known start values $y_{k-s}, \dots, y_{k-1}$. Mostly, they are not provided by the IVP itself. Thus, general LMM decompose into two parts:

- a *starting phase* where the start values are computed in a suitable way and
- a *run phase* where the LMM is executed.

It is crucial that the starting procedure provides a suitable order corresponding to the LMM of the run phase, recall condition (5.16) in Definition 5.4.6. Possible choices for the starting phase include multistep methods with variable order and one-step methods.

Example 5.5.2 (Self starter)**.** A 2-step BDF method requires y_0 and y_1 to be known. y_0 is given by the initial value while y_1 is unknown so far. To guarantee that the method has order 2, y_1 needs to be at least locally of order 2, i.e.,

$$|u(t_1) - y_1| \le c_0 h^2. \tag{5.24}$$

This is ensured, for example, by one step of the 1-step BDF method (implicit Euler).

However, starting an LMM with $s > 2$ steps by a first-order method and then successively increasing the order until s is reached does not provide the desired global order. That is due to the fact that a one-step method cannot have more than order 2, limiting the overall convergence order to 2. Nevertheless, self starters are often used in practice.

Example 5.5.3 (Runge-Kutta starter)**.** One can use Runge-Kutta methods to start LMMs. Since only a fixed number of starting steps are performed, the local order of the Runge-Kutta approximation is crucial. For an implicit LMM with convergence order p and stepsize h one could use an RK method with consistency order $p-1$ with the same step size h.

Consider a 3-step BDF method. Thus, apart from y_0, we need start values y_1, y_2 with errors less than $c_0 h^3$. They can be computed by RK methods of consistency order 2, for example by two steps of the 1-stage Gauß collocation method with step size h since it has consistency order $2s = 2$, see theorem 3.4.15.

Remark 5.5.4. In practice not the order of a procedure is crucial but rather the fact that the errors of all approximations (the start values and all approximations of the run phase) are bounded by the user-given tolerance, compare Section 2.4. Generally, LMMs are applied with variable step sizes and orders in practice.

Thus, the step sizes of all steps are in practice controlled using local error estimates. Hence, self starting procedures usually start with very small step sizes and increase them successively. Due to their higher orders RK starters usually are allowed to use moderate step sizes in the beginning.

5.6 LMM and stiff problems

To study A-stability of LMMs we consider again the model equation $u' = \lambda u$. Applying a general LMM (5.3) to this model equation leads to the linear model difference equation

$$\sum_{r=0}^{s} (\alpha_{s-r} - z\beta_{s-r}) y_{k-r} = 0. \tag{5.25}$$

with $z = \lambda h$.

Definition 5.6.1 (A-stability of LMM): The **stability region** of an LMM is the set of points $z \in \mathbb{C}$, for which all sequences $(y_n)_{n=0}^{\infty}$ of solutions of the equation (5.25) stay bounded for $n \to \infty$. An LMM is called **A-stable**, if the stability region contains the left half-plane of $\mathbb{C}$.

Note that this definition is equivalent to the definition of A-stability for one-step methods in definition 3.2.5.

Definition 5.6.2: The so-called **stability polynomial** of an LMM is obtained by inserting $y_n = x^n$ to obtain

$$R_z(x) = \sum_{r=0}^{s} (\alpha_{s-r} - z\beta_{s-r}) x^{s-r}. \tag{5.26}$$

Remark 5.6.3. Instead of the simple amplification function $R(z)$ of the one-step methods, we get here a function of two variables: the point z for which we want to show stability and the artificial variable x from the analysis of the method.

Lemma 5.6.4: Let $\{\xi_1(z), \dots, \xi_s(z)\}$ be the set of roots of the stability polynomial $R_z(x)$ as functions of z. A point $z \in \mathbb{C}$ is in the stability region of a LMM, if these roots satisfy the root test in corollary 5.3.6.

Proof. The proof is analog to that of theorem 5.4.3. □

Theorem 5.6.5 (2nd Dahlquist barrier): There is no A-stable LMM of order $p > 2$. Among the A-stable LMM of order 2, the trapezoidal rule (Crank-Nicolson) has the smallest error constant.

Proof. See [HW10, Theorem V.1.4]. □

5.6.1 A(α)-stability

5.6.6. Motivated by the fact that there are no higher order A-stable LMMs people have introduced relaxed concepts of A-stability.

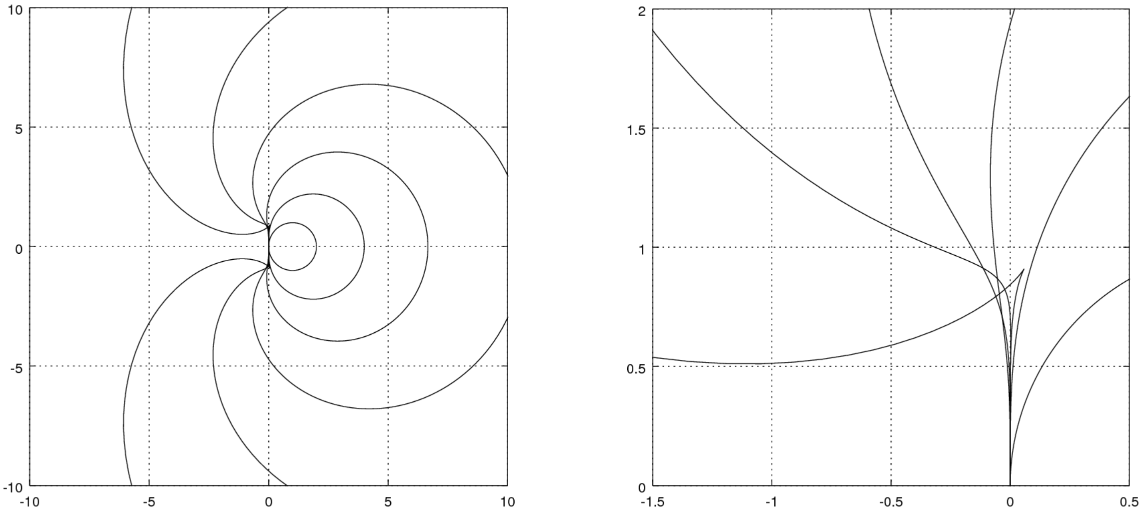


Figure 5.1: Boundaries of the stability regions for BDF(1) to BDF(6); the unstable region is right of the origin. The right figure shows a zoom near the origin.

Definition 5.6.7: A LMM is called **A(α)-stable**, for $\alpha \in [0°, 90°]$, if its stability region contains the sector

$$\left\{ z \in \mathbb{C} \,\middle|\, \operatorname{Re}(z) < 0 \quad \text{and} \quad \left|\frac{\operatorname{Im}(z)}{\operatorname{Re}(z)}\right| \le \tan\alpha \right\}.$$

It is called **A(0)-stable**, if the stability region contains the negative real axis.

It is called **stiffly stable**, if it contains the set $\{\operatorname{Re}(z) < -D\}$.

Remark 5.6.8. The introduction of A(0)-stability is motivated by linear systems of the form $u' = -Au$ with symmetric positive definite matrix A. Only stability on the real axis is required in that case, since all eigenvalues are real. Any non-negative angle α is sufficient.

Similarly, A(α)-stable LMM are suitable for linear problems in which high-frequency vibrations (large Imλ) decay fast (large $-$Reλ).

LMMs behave similarly for nonlinear problems if the Jacobian matrix $\mathrm{D}_y f$ satisfies corresponding properties.

Example 5.6.9. The stability regions of the stable BDF methods are in Figure 5.1. The corresponding values for A(α)-stability and stiff stability are in Table 5.1. (Recall from theorem 5.4.5 that BDF(7) is not even zero-stable.)

k	1	2	3	4	5	6
α	90°	90°	86.03°	73.35°	51.84°	17.84°
D	0	0	0.083	0.667	2.327	6.075

Table 5.1: Values for A(α)- and stiff stability for BDF methods of order k.

Chapter 6

Boundary Value Problems

This chapter deals with problems of a fundamentally different type than the problems we examined in Chapter 1, namely boundary value problems. Here, we have prescribed values at the beginning and at the end of an interval of interest. They will require the design of different numerical methods.

6.1 General boundary value problems

Due to Lemma 1.2.4 we know that every ODE can be written as a system of first-order ODEs. Thus, we make the following definition (restricting our attention to explicit ODEs).

Definition 6.1.1: A **boundary value problem** (BVP) is a differential equation problem of the form: Find $u : [a, b] \subset \mathbb{R} \to \mathbb{R}^d$, such that

$$u'(t) = f\big(t, u(t)\big) \qquad t \in (a, b) \tag{6.1a}$$

$$r\big(u(a), u(b)\big) = 0. \tag{6.1b}$$

Definition 6.1.2: A BVP (6.1) is called linear, if the right hand side f as well as the boundary conditions are linear in u, i.e.: Find $u : [a, b] \to \mathbb{R}^d$ such that

$$u'(t) = A(t)u(t) + c(t) \qquad \forall t \in (a, b) \tag{6.2a}$$

$$B_a u(a) + B_b u(b) = g. \tag{6.2b}$$

with $A : [a, b] \to \mathbb{R}^{d\times d},\ c : [a, b] \to \mathbb{R}^d,\ B_a, B_b \in \mathbb{R}^{d\times d}$ and $g \in \mathbb{R}^d$.

Remark 6.1.3. Since boundary values are imposed at two different points in time, the concept of local solutions from Definition 1.2.8 is not applicable. Thus, tricks, such as going forward from interval to interval, as is done in the proof of Péano's theorem using Euler's method, are here not applicable. For this reason, nothing can be concluded from the local properties of the solution and from the right hand side f. In fact, it is hard in general even to establish that a solution exists.

Example 6.1.4. Consider the linear BVP

$$\begin{bmatrix} u_1'(t) \\ u_2'(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{with}$$

(i) $u_1(0) = 0,\ u_2(1) = 0,$

(ii) $u_1(0) = 0,\ u_1(1) = 0,$

(iii) $u_2(0) = 0,\ u_2(1) = 0.$

By substitution, we can easily see that this first-order system of ODEs is in fact equivalent to the second-order ODE $u_1'' = 1$, which can be explicitly solved by integration to give

$$u_2(t) = u_1'(t) = t + c_1 \quad \text{and} \quad u_1(t) = \frac{t^2}{2} + c_1 t + c_2.$$

But the BVP is **not** solvable for each of the three choices of boundary conditions (BCs). In fact, we get

(i) $c_2 = 0$ and $c_1 = -1$, i.e., $u(t) = \big(t(\frac{t}{2} - 1), t - 1\big)^T$,

(ii) $c_2 = 0$ and $c_1 = -1/2$, i.e., $u(t) = \big(\frac{t}{2}(t-1), t - \frac{1}{2}\big)^T$,

(iii) but here the two BCs lead to $c_1 = 0$ and $c_1 = -1$, respectively, which cannot be satisfied simultaneously.

Remark 6.1.5. Note that due to Lemma 1.3.13 we know that the solution space of the linear ODE in (6.2a) is d-dimensional. Hence, we need d additional pieces of information to determine the solution uniquely. However, whether the d boundary conditions in (6.2b) are sufficient is more subtle than in the case of an IVP, as we have just seen.

However, for linear BVPs we have the following existence an uniqueness result.

Theorem 6.1.6: Let $\Psi(t) \in \mathbb{R}^{d \times d}$ be the solution of the variational equation (cf. Definition 3.5.10)

$$\Psi'(t) = A(t)\Psi(t),$$

of the ODE (6.2a) with initial condition $\Psi(a) = \mathbb{I}$. The linear BVP (6.2) has an unique solution $u(t)$ if and only if the $d \times d$ matrix

$$B_a + B_b \Psi(b)$$

is regular.

Proof. See Rannacher [Ran17b, Satz 8.2]. □

We will not discuss this further and instead consider an important subclass of linear BVPs, as well as the most important numerical methods for them. For more details on the general solution theory, see chapter 6 in the original notes by G. Kanschat or [Ran17b, Chap. 8].

6.2 Second-order, scalar two-point boundary value problems

Let us consider following linear, second-order BVP of finding $u : [a, b] \to \mathbb{R}$ such that

$$-u''(x) + \beta(x)u'(x) + \gamma(x)u(x) = f(x), \qquad u(a) = u_a, \quad u(b) = u_b. \tag{6.3}$$

for some functions $\beta, \gamma, f : [a, b] \to \mathbb{R}$ and two boundary values $u_a, u_b \in \mathbb{R}$. (As is common in practice, we use x instead of t to denote the independent variable here.)

We introduce the set

$$\mathcal{B} = \left\{ u \in C^2(a, b) \cap C[a, b] \;\middle|\; u(a) = u_a \text{ and } u(b) = u_b \right\}.$$

Then, we can see the LHS of (6.3) as a differential operator applied to u, mapping $\mathcal{B}$ to the set of continuous functions. Namely, we define

$$\begin{aligned} L : \mathcal{B} &\to C[a, b] \\ u &\mapsto -u'' + \beta u' + \gamma u. \end{aligned} \tag{6.4}$$

To simplify our life we can (without loss of generality) get rid of the inhomogeneous boundary values u_a and u_b. To this end, let

$$\psi(x) = \frac{b - x}{b - a}\, u_a + \frac{x - a}{b - a}\, u_b\,,$$

and introduce the new function $u_0 := u - \psi$. Then, u_0 solves the BVP

$$\begin{aligned} -u_0''(x) + \beta(x)u_0'(x) + \gamma(x)u_0(x) &= \underbrace{f(x) - \beta(x)\frac{u_b - u_a}{b - a} - \gamma(x)\psi(x)}_{=:\, f_0(x)}, \\ u_0(a) = u_0(b) &= 0. \end{aligned}$$

Thus, it is sufficient to consider the boundary value problem:

Definition 6.2.1: Given an interval $I = [a, b]$, find a function

$$u \in V = \left\{ u \in C^2(a, b) \cap C[a, b] \;\middle|\; u(a) = u(b) = 0 \right\}, \tag{6.5}$$

such that for a differential operator of second order as defined in (6.4) above and a right hand side $f \in C[a, b]$ there holds

$$Lu = f. \tag{6.6}$$

6.3 Finite difference methods

We subdivide the interval $I = [a, b]$ again into subintervals, as in Definition 2.1.2, and consider the solution only at the partitioning points $a = x_0 \le x_1 \ldots \le x_n = b$. As with one-step and multistep timestepping methods, we denote the approximate solution values at those partitioning points by y_k, $k = 0, \ldots, n$.

While one-step methods directly discretize the Volterra integral equation in order to compute the solution at every new step, **finite difference methods** discretize the differential equation on the whole interval at once and then solve the resulting discrete (finite-dimensional) system of equations.

Thus, we have accomplished the first step and decided that instead of function values $u(x)$ in every point x of the interval I, we only approximate $u(x_k)$ in the points of the partition by y_k, $k = 0, \dots, n$. What is left is the definition of the discrete operator that approximates the differential operator L in (6.4).

Definition 6.3.1 (Finite differences): To approximate **first** derivatives of a function u, we introduce the operators

$$\text{Forward difference} \qquad D_h^+ u(x) = \frac{u(x+h) - u(x)}{h}, \tag{6.7}$$

$$\text{Backward difference} \qquad D_h^- u(x) = \frac{u(x) - u(x-h)}{h}, \tag{6.8}$$

$$\text{Central difference} \qquad D_h^c u(x) = \frac{u(x+h) - u(x-h)}{2h}. \tag{6.9}$$

For **second** derivatives we introduce the

$$\text{3-point stencil} \qquad D_h^2 u(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2}. \tag{6.10}$$

Remark 6.3.2. Note that the 3-point stencil is the product of the forward and backward difference operators:

$$D_h^2 u(x) = D_h^+ \Big(D_h^- u(x) \Big) = D_h^- \Big(D_h^+ u(x) \Big).$$

For simplicity, we only present finite differences of uniform subdivisions. Nevertheless, the definition of the operators can be extended easily to h changing between intervals.

Definition 6.3.3: A finite difference operator D_h^α is **consistent of order** p with the αth derivative, if there exists a constant $c > 0$ independent of h and a subset $\tilde{I} \subset [a, b]$, such that for any $u \in C^{\alpha+p}(a, b)$ and for any $x \in \tilde{I}$:

$$|D_h^\alpha u(x) - u^{(\alpha)}(x)| \le ch^p \tag{6.11}$$

Lemma 6.3.4: The forward and backward difference operators D_h^+ and D_h^- in definition 6.3.1 are consistent of order 1 with the first derivative, i.e. for any $x \in [a, b-h]$ (resp. $x \in [a+h, b]$),

$$|D_h^+ u(x) - u'(x)| \le ch \quad \text{and} \quad |D_h^- u(x) - u'(x)| \le ch. \tag{6.12}$$

The central difference operator D_h^c and the 3-point stencil D_h^2 are consistent of order 2 with the first and second derivative, respectively, i.e. for any $x \in [a+h, b-h]$,

$$|D_h^c u(x) - u'(x)| \le ch^2 \quad \text{and} \quad |D_h^2 u(x) - u''(x)| \le ch^2. \tag{6.13}$$

Proof. Taylor expansion: Let $x \in [a, b-h]$. Then there exists $\xi \in (x, x+h)$ such that

$$\begin{aligned} D_h^+ u(x) - u'(x) &= \frac{u(x+h) - u(x)}{h} - u'(x) \\ &= \frac{u(x) + hu'(x) + \frac{h^2}{2}u''(\xi) - u(x)}{h} - u'(x) \;=\; \tfrac{h}{2}u''(\xi). \end{aligned}$$

Thus, the result for D_h^+ holds with $c := \frac{1}{2}\max_{x\in(a,b)}|u''(x)|$. The same computation can be applied to $D_h^- u(x)$.

For D_h^c, let $x \in [a+h, b-h]$. Then, there exist $\xi^-, \xi^+ \in (x-h, x+h)$ such that

$$\begin{aligned} &D_h^c u(x) - u'(x) = \\ &\quad \frac{u(x) + hu'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(\xi^+) - \big(u(x) - hu'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(\xi^-)\big)}{2h} - u'(x) \\ &\qquad = \tfrac{h^2}{12}\left(u'''(\xi^-) + u'''(\xi^+)\right). \end{aligned}$$

The final result for the 3-point stencil D_h^2 follows in a similar way $\boxed{\text{DIY}}$. □

Remark 6.3.5. When applied to the equation $u' = f(t, u)$ the solutions obtained by forward and backward differences correspond to the explicit and implicit Euler methods, respectively.

Definition 6.3.6: The **finite difference method** (with uniform subdivisions) for the discretization of the BVP $Lu = f$ on the interval $I = [a, b]$ with homogeneous boundary conditions, i.e., for $u \in V$ as in definition 6.2.1, is defined by

1. choosing a partition $a = x_0 < x_1 < \cdots < x_n = b$ with $n \in \mathbb{N}$,

$$h := (b-a)/n \quad \text{and} \quad x_k = a+kh, \quad k = 0, \dots, n,$$

2. replacing all differential operators in L by finite differences, evaluated at x_k,
3. considering and computing the approximations y_k of the solution $u(x_k)$ at the discrete points $x_0, \dots, x_n$.

Example 6.3.7. Using the 3-point stencil for u'' and central differences for u', the BVP

$$-u''(x) + \beta(x)u'(x) + \gamma(x)u(x) = f(x), \qquad u(a) = u(b) = 0,$$

and the abbreviations $\beta_k = \beta(x_k)$, $\gamma_k = \gamma(x_k)$ $f_k = f(x_k)$, we obtain the discrete system of equations

$$\frac{-y_{k+1} + 2y_k - y_{k-1}}{h^2} + \beta_k \frac{y_{k+1} - y_{k-1}}{2h} + \gamma_k y_k \;=\; f_k\,, \quad \text{for} \;\; k = 1, \dots, n-1, \tag{6.14}$$

with $y_0 = y_n = 0$, or in matrix notation

$$L_h\, y = \begin{pmatrix} \lambda_1 & \nu_1 & & \\ \mu_2 & \lambda_2 & \ddots & \\ & \ddots & \ddots & \nu_{n-2} \\ & & \mu_{n-1} & \lambda_{n-1} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_{n-1} \end{pmatrix} = f_h. \tag{6.15}$$

where

$$\lambda_k = \frac{2}{h^2} + \gamma_k\,, \quad \mu_k = -\frac{1}{h^2} - \frac{\beta_k}{2h}\,, \quad \text{and} \quad \nu_k = -\frac{1}{h^2} + \frac{\beta_k}{2h}\,. \tag{6.16}$$

Remark 6.3.8. Like our view to the continuous BVP has changed w.r.t. IVPs, the discrete problem is now a fully coupled linear system which has to be solved by methods of linear algebra, rather than time stepping. In fact, we have $n-1$ unknown variables $y_1,\dots,y_{n-1}$ and $n-1$ equations, such that existence and uniqueness of solutions are equivalent.

6.3.1 Existence, stability and convergence

6.3.9. Since the solution of the discretized boundary value problem is a problem in linear algebra, we have to study properties of the matrix L_h. The shortest and most elegant way to prove stability is through the properties of M-matrices, which we present here very shortly. We are not dwelling on this approach too long, since it is sufficient for stability, but by far not necessary and particular to low order methods.

The fact that L_h is an M-matrix requires some knowledge of irreducible weakly diagonal dominant matrices, which we have already come across in the last chapter in Numerik 0, in the context of stationary iterative methods.

Definition 6.3.10: A quadratic $n\times n$-matrix A is called an **M-matrix** if it satisfies the following properties:

$$a_{ii} > 0, \quad a_{ij} \le 0, \qquad i,j = 1,\dots,n, \quad j \neq i. \tag{6.17}$$

The entries of $A^{-1} = (c_{ij})_{i,j=1}^n$ satisfy

$$c_{ij} \ge 0, \qquad i,j = 1,\dots,n. \tag{6.18}$$

Lemma 6.3.11: The matrix L_h defined in (6.5.4) above is an M-matrix provided that

$$\gamma_k \ge 0 \quad \text{and} \quad |\beta_k| < \frac{2}{h}. \tag{6.19}$$

Proof. It is easy to verify that the two conditions in (6.19) are sufficient for the first M-matrix property. The proof of positivity of the inverse is based on irreducible diagonal dominance, which is too long and too specialized at this point and thus we will omit it. See, e.g., [Ran17b, Hilfssatz 10.2]. □

Remark 6.3.12. The finite element method, intorduced in Section 6.4 below and discussed in more detail next semester in “Numerik 2 – Finite Elements”, provides a much more powerful theory to deduce solvability and stability of the discrete problem.

Lemma 6.3.13: Let A be an M-matrix. If there is a vector w such that for the vector $v = Aw$ there holds

$$v_i \geq 1, \quad i = 1, \dots, n,$$

then

$$\|A^{-1}\|_\infty \leq \|w\|_\infty. \tag{6.20}$$

Proof. Let $x \in \mathbb{R}^n$ and $y = A^{-1}x$. Then,

$$|y_i| = \left|\sum_{j=1}^n c_{ij}x_j\right| \leq \sum_{j=1}^n c_{ij}|x_j| \leq \|x\|_\infty \sum_{j=1}^n c_{ij}v_j.$$

Thus,

$$|y_i| \leq \|x\|_\infty \big(A^{-1}v\big)_i = \|x\|_\infty \big(A^{-1}Aw\big)_i \leq \|x\|_\infty |w_i|.$$

Taking the maximum over all i and dividing by $\|x\|_\infty$, we obtain

$$\|A^{-1}\|_\infty = \sup_{x\in\mathbb{R}^n} \frac{\|A^{-1}x\|_\infty}{\|x\|_\infty} \leq \|w\|_\infty.$$

□

Theorem 6.3.14: Assume that (6.19) holds and that there exists a constant $\delta < 2$ such that

$$|\beta_k| \leq \frac{\delta}{b-a}. \tag{6.21}$$

Then, the matrix L_h defined in (6.5.4) is invertible and

$$\|L_h^{-1}\|_\infty \leq \frac{(b-a)^2}{8-4\delta}. \tag{6.22}$$

Proof. Consider the function

$$p(x) = (x-a)(b-x) = -x^2 + (a+b)x - ab,$$

with derivatives $p'(x) = a + b - 2x$ and $p''(x) = -2$, and a maximum of $(b-a)^2/4$ at $x = (a+b)/2$. Choose the values $p_k = p(x_k)$. Due to the consistency results in Lemma 6.3.4, we know that $D_h^2 p \equiv p''$ and $D_h^c p \equiv p'$ are exact, such that, for all $k = 1, \dots, n-1$,

$$(L_h p)_k = 2 + \beta_k p'(x_k) + \gamma_k p(x_k) \geq 2 - |\beta_k|(b-a) \geq 2 - \delta.$$

Since L_h is a M-matrix, the vector $w = \frac{1}{2-\delta}p$ can then be used to bound the inverse of L_h using Lemma 6.3.13. □

Remark 6.3.15. The assumptions of the previous theorem involve two sets of conditions on the parameters β_k and γ_k. Since

$$\frac{\delta}{b-a} < \frac{2}{b-a} \le \frac{2n}{b-a} = \frac{2}{h},$$

condition (6.21) actually implies the second condition in (6.19). It is in fact not necessary in this form, but a better estimate requires more advanced analysis.

The condition on γ_k in (6.19) is indeed necessary, as will be seen when we study partial differential equations. The second condition in (6.19), on the other hand, relates the coefficients β_k to the mesh size and can be avoided, as seen in the next example.

Example 6.3.16. By changing the discretization of the first order term to an **upwind** finite difference method, we obtain an M-matrix independent of the relation of β_k and h. To this end define

$$\beta(x)D_h^{\uparrow}u(x) = \begin{cases} \beta(x)D_h^- u(x) & \text{if } \beta(x) > 0 \\ \beta(x)D_h^+ u(x) & \text{if } \beta(x) < 0 \end{cases}. \tag{6.23}$$

This changes the matrix L_h to a matrix $L_h^{\uparrow}$ with entries

$$\lambda_k = \frac{2}{h^2} + \frac{|\beta_k|}{h} + \gamma_k\,, \quad \mu_k = -\frac{1}{h^2} - \frac{\max\{0,\beta_k\}}{h}\,, \quad \nu_k = -\frac{1}{h^2} + \frac{\min\{0,\beta_k\}}{h}\,. \tag{6.24}$$

As a consequence, the off-diagonal elements always remain non-positive and the diagonal elements remain positive provided only that $\gamma_k \ge 0$, for all k. Thus, $L_h^{\uparrow}$ is an M-matrix with a bounded inverse, independent of the values of β_k. However, crucially, the consistency order is reduced from two to one.

Theorem 6.3.17: Consider the boundary value problem defined in definition 6.2.1 with $\beta, \gamma, f \in C^4(a,b)$ and $\gamma(x) \ge 0$ for all $x \in [a,b]$. Let $y \in \mathbb{R}^{n-1}$ be the finite difference approximation for this problem in Example 6.3.7. If there exists a $\delta < 2$ such that $\max_{x\in[a,b]}|\beta(x)| \le \delta/(b-a)$, then there exists a constant c independent of h such that

$$\max_{0\le k\le n} |u_k - y_k| \le ch^2. \tag{6.25}$$

For the solution $y^{\uparrow} \in \mathbb{R}^{n-1}$ of the upwind finite difference approximation in Example 6.3.16 there exists a constant c independent of h such that

$$\max_{0\le k\le n} |u_k - y_k^{\uparrow}| \le ch. \tag{6.26}$$

without any additional assumptions on the function β.

Proof. Let $n \in \mathbb{N}$ (and thus $h > 0$) be arbitrary but fixed and let $U = (u_k)_{k=1}^{n-1}$ be the vector containing the values of the exact solution at $x_1, \dots, x_{n-1}$. Considering first the discretisation in Example 6.3.7 and denoting by

$$\tau_k := (L_hU)_k - (Lu)(x_k), \quad k = 1, \dots, n-1,$$

the consistency errors at the interior grid points. Then

$$\big(L_h(U-y)\big)_k = (L_hU)_k - (Lu)(x_k) + (Lu)(x_k) - (L_hy)_k = \tau_k + f_k - f_k = \tau_k$$

and it follows from (6.13) that

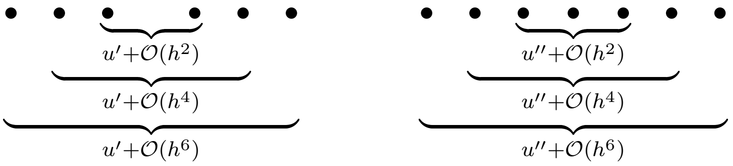
$$\|L_h(U-y)\|_\infty = \|\tau\|_\infty \le ch^2$$

with c independent of h. Since β, γ satisfy the assumptions of theorem 6.3.14 (for arbitrary $h > 0$), we can conclude that there exists a $c' > 0$ independent of h such that

$$\|U-y\|_\infty = \|L_h^{-1}L_h(U-y)\|_\infty \le \|L_h^{-1}\|_\infty \|\tau\|_\infty \le c'h^2.$$

The proof for the upwind discretisation in Example 6.3.16 is identical, but as discussed does not require any boundedness of β to guarantee stability and due to the use of the forward/backward difference quotients, the consistency error is only of $\mathcal{O}(h)$. □

Remark 6.3.18. Finite differences can be generalized to higher order by extending the stencils by more than one point to the left and right of the current point. Whenever we add two points to the symmetric difference formulas, we can gain two orders of consistency.



Similarly, we can define one-sided difference formulas, which get us close to multistep methods. The matrices generated by these formulas are not M-matrices anymore, although you can show for the 4th order formula for the second derivative that it yields a product of two M-matrices. While this rescues the theory in a particular instance, M-matrices do not provide a theoretical framework for general high order finite differences anymore.

Very much like the starting procedures for high order multistep methods, high order finite differences can lead to difficulties at the boundaries. Here, the formulas must be truncated and for instance be replaced by one-sided formulas of equal order.

All these issues motivate the study of different discretization methods in the next course.

6.4 Finite element methods

To describe this alternative numerical method for BVPs, we onsider the problem

$$-(ku')'(x) = f(x), \quad \text{for all } \ x \in [0,1]\ , \quad u(0) = u(1) = 0\ , \tag{6.27}$$

where $f \in C[0,1]$ and $k \in C^1[0,1]$ are given functions and k is assumed to be uniformly positive, i.e.

$$k(x) \ge k_0 > 0, \quad x \in [0,1].$$

Remark 6.4.1. Note that (6.27) may also be written in the form

$$-ku'' - k'u' = f, \tag{6.28}$$

i.e. a second order ODE of the form (6.3) with $\beta = -k'/k$, but the so-called **divergence-form** of the ODE in (6.27) is preferable. As for finite difference methods, we could also add a nonzero zero-order term and the analysis would remain the same provided $\gamma(x) \geq 0$, for all $x \in (0,1)$, but we choose not to for simplicity.

A solution of (6.27) can be found by integrating the equation twice and by building in the boundary conditions. This requires integrals of k and f which often can not be found explicitly. So numerical methods are needed even in this simple case.

The idea of the **finite element method (FEM)** now is:

Step 1. Rewrite (6.27) in a *'weak form'* only involving first derivatives of u.

Step 2. Approximate the weak form.

6.4.1 Weak form of the differential equation

To find the weak form of (6.27) we start by defining the following set of functions on $[0,1]$:

$$V \;:=\; \big\{ v \in C[0,1] : v' \text{ is bounded \& piecewise continuous on } [0,1] \;\text{ and }\; v(0) = 0 = v(1) \big\}$$

Notation. A function v is called piecewise continuous on $[a,b]$, if there exist $a = a_0 < a_1 < \cdots < a_m = b$ such that $v \in C(a_{i-1}, a_i)$ for all $i = 1, \ldots, m$.

Now suppose u solves (6.27). Then multiplying (6.27) by an arbitrary function $v \in V$ and integrating over $[0,1]$ we obtain

$$-\int_0^1 \big(ku'\big)'\, v\, dx \;=\; \int_0^1 fv\, dx$$

Integrating by parts and applying the boundary conditions $v(0) = 0 = v(1)$ leads to

$$\int_0^1 ku'v'\, dx \;-\; \underbrace{\big[ku'v\big]_0^1}_{=0} \;=\; \int_0^1 fv\, dx$$

Since $v \in V$ was arbitrary, u is a solution of the problem: Find $u \in V$ such that

$$a(u,v) = \ell(v), \qquad \forall\, v \in V \tag{6.29}$$

where

$$a(u,v) \;:=\; \int_0^1 ku'v'\, dx \quad \text{and} \quad \ell(v) \;:=\; \int_0^1 fv\, dx \tag{6.30}$$

The problem (6.29) is called the **weak form** of (6.27).

Lemma 6.4.2: The problem (6.29) has a unique solution in V.

Proof. There exists a solution u of (6.29), namely, the solution u of (6.27) and $u \in V$. $\boxed{\text{DIY}}$

Before proving uniqueness, note that for all $v \in V$

$$a(v,v) = \int_0^1 k|v'|^2 \,\mathrm{d}x \geq k_0 \int_0^1 |v'|^2 \,\mathrm{d}x$$

since $k_0 > 0$. Hence, $a(v,v) = 0$ implies $\int_0^1 |v'|^2 \,\mathrm{d}x = 0$ and thus $v'(x) = 0$ for all $x \in [0,1]$. Using the boundary conditions that implies

$$a(v,v) = 0 \quad \Longrightarrow \quad v \equiv 0 \quad \text{on} \quad [0,1]. \tag{6.31}$$

Now if $u_1, u_2 \in V$ are two solutions of (6.29) then for all $v \in V$,

$$0 = \ell(v) - \ell(v) = a(u_1, v) - a(u_2, v) = a(u_1 - u_2, v) \tag{6.32}$$

by linearity of (6.30) with respect to its first argument. Since V is a vector space $\boxed{\text{DIY}}$ then $u_1 - u_2 \in V$, so putting $v = u_1 - u_2$ in (6.32) and using (6.31) it follows that $u_1 \equiv u_2$. □

Let us now proceed to **Step 2**, i.e., to approximate (6.29). We choose any finite dimensional space $V_h \subset V$ and introduce the approximate weak form: Find $u_h \in V_h$ such that

$$a(u_h, v_h) \;=\; \ell(v_h) \qquad \forall\, v_h \in V_h\,. \tag{6.33}$$

The notation 'subscript' h typically indicates the stepsize of a mesh on $[0,1]$ (see below).

Let $\{\varphi_1, \ldots, \varphi_n\}$ be a basis for V_h. Then we can write

$$u_h \;=\; \sum_{j=1}^{n} y_j \varphi_j \tag{6.34}$$

for some unknown coefficients $y_j \in \mathbb{R}$. Since a is linear in its first argument, (6.33) is equivalent to

$$\sum_{j=1}^{n} a(\varphi_j, v_h)\, y_j \;=\; \ell(v_h) \qquad \forall\, v_h \in V_h\,. \tag{6.35}$$

Furthermore, the linearity of ℓ and of a (in its second argument) imply that (6.35) is in fact equivalent to

$$\sum_{j=1}^{n} a(\varphi_j, \varphi_i)\, y_j \;=\; \ell(\varphi_i), \qquad i = 1, \ldots, n. \tag{6.36}$$

Thus, the vector of coefficients $y = (y_j)_{j=1}^n \in \mathbb{R}^n$ is the solution of the (algebraic) system

$$Ay = b \tag{6.37}$$

where

$$A_{i,j} \;:=\; a(\varphi_j, \varphi_i) \qquad \text{and} \qquad b_i \;:=\; \ell(\varphi_i)\,.$$

Hence, by solving (6.37) for $y \in \mathbb{R}^n$ and substituting the solution into (6.34) we obtain the solution of (6.33).

Lemma 6.4.3: The matrix A in (6.37) is non-singular and so (6.37) has a unique solution.

Proof. Since A is a square $n \times n$ matrix we only have to check whether the following property holds

$$Az = 0 \quad \Longrightarrow \quad z = 0 \tag{6.38}$$

Indeed, if (6.38) is true, A is non-singular and the columns of A are linearly independent, which ensures that $Ay = b$ has a solution for all right hand sides $b \in \mathbb{R}^n$.

So suppose $Az = 0$. Then

$$0 = \sum_{i=1}^{n} z_i \, (Az)_i = \sum_{i=1}^{n} \sum_{j=1}^{n} z_i \underbrace{A_{ij}}_{= a(\varphi_j, \varphi_i)} z_j$$

So by linearity of a with respect to both arguments

$$0 = a\left(\sum_{j=1}^{n} z_j \varphi_j, \sum_{i=1}^{n} z_i \varphi_i \right) = a(v_h, v_h) \quad \text{where} \quad v_h := \sum_{i=1}^{n} z_i \varphi_i \,.$$

Hence, $v_h = 0$ by (6.31) and thus $z = 0$, since $\{\varphi_i\}$ is a basis and thus linearly independent.

□

6.4.2 Piecewise linear finite elements

This is a practical method which corresponds to a special choice of V_h in (6.33).

Definition 6.4.4: Let

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1$$

be a mesh on $[0, 1]$ with $h := \max_{i=1,\ldots,n+1} (x_i - x_{i-1})$. Then the **piecewise linear finite element space** is defined by

$$V_h \;:=\; \left\{ v \in C[0,1] : v|_{(x_{i-1},x_i)} \text{ is linear and } \; v(0) = 0 = v(1) \right\}. \tag{6.39}$$

A basis for V_h is given by the **hat functions** $\{\varphi_i : i = 1, \ldots, n\}$, defined by

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i]\,, \\[2ex] \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in [x_i, x_{i+1}]\,, \\[2ex] 0, & \text{for all other } x\,. \end{cases}$$

The graphs of two hat functions φ_i and φ_j $(j \neq i)$ are shown in Figure 6.1.

In the special case of a *uniform mesh*, where $x_i = ih$ and $h = \frac{1}{n+1}$, the hat functions are given by

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h}, & x \in [x_{i-1}, x_i]\,, \\[2ex] \dfrac{x_{i+1} - x}{h}, & x \in [x_i, x_{i+1}]\,, \\[2ex] 0, & \text{for all other } x\,. \end{cases}$$

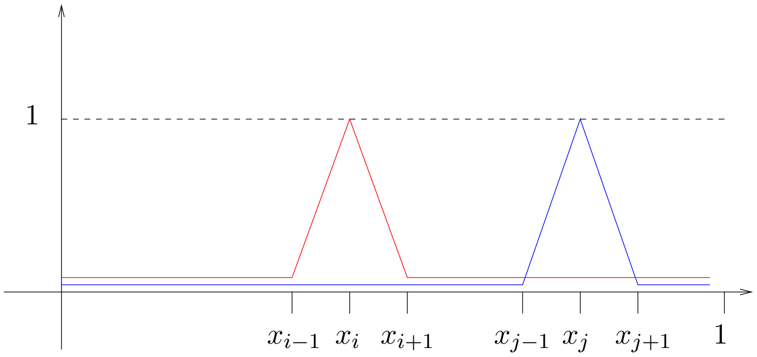


Figure 6.1: Two hat functions φ_i and φ_j $(j \neq i)$.

Note. Clearly, $\varphi_i \in V_h$ for all $i = 1, \ldots, n$. We do not include φ_0 and φ_{n+1}, since they do not satisfy the boundary conditions in V_h.

Exercise. It can be shown that $\{\varphi_1, \ldots, \varphi_n\}$ forms a basis for V_h. $\boxed{\text{DIY}}$

Now considering the problem (6.33) with V_h as defined in (6.39), or equivalently

$$V_h \;=\; \text{span}\;\{\varphi_i : i = 1, \ldots, n\}\,.$$

Then, by the definition of φ_i and $a(.,.)$, the matrix in (6.37) has entries

$$A_{i,j} = \begin{cases} 0, & \text{if } \; |i-j| > 1 \\ \displaystyle\int_{x_{i-1}}^{x_i} k\varphi'_{i-1}\varphi'_i \, dx\,, & \text{if } \; j = i-1, \\ \displaystyle\int_{x_i}^{x_{i+1}} k\varphi'_{i+1}\varphi'_i \, dx\,, & \text{if } \; j = i+1, \\ \displaystyle\int_{x_{i-1}}^{x_{i+1}} k\varphi'_i\varphi'_i \, dx\,, & \text{if } \; j = i. \end{cases} \tag{6.40}$$

If k constant or if the integral of k is known explicitly, all matrix entries can be computed exactly $\boxed{\text{DIY}}$. Otherwise, we may need numerical integration.

We observe that the so-called **stiffness matrix** A is *tridiagonal*:

$$A = \begin{bmatrix} * & * & 0 & 0 & \ldots & \ldots & 0 \\ * & * & * & 0 & \ldots & \ldots & 0 \\ 0 & * & * & * & 0 & \ldots & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & \ldots & 0 & * & * & * \\ 0 & 0 & \ldots & \ldots & 0 & * & * \end{bmatrix}$$

where $*$ denotes a non-zero entry. It can be solved very efficiently, e.g. using the *Thomas Algorithm* (Gaussian elimination for tridiagonal systems, see **Numerik 0**).

6.4.3 Abstract error analysis

Definition 6.4.5: Let V be a real vector space.

(a) A function $a : V \times V \to \mathbb{R}$ that is linear in each argument is called a **bilinear form**.

(b) A bilinear form is called an **inner product** on V if it is **symmetric**,

$$a(v, w) \;=\; a(w, v) \qquad \forall v, w \in V,$$

and if

$$a(v, v) \;\geq\; 0 \qquad \forall v \in V$$

with equality iff $v = 0$.

If a is an inner product, we can define the (induced) **energy norm**

$$\|v\|_a \;:=\; a(v, v)^{\frac{1}{2}}\,,$$

which satisfies the Cauchy-Schwarz inequality

$$|a(v, w)| \;\leq\; \|v\|_a \|w\|_a \qquad \forall v, w \in V\,. \tag{6.41}$$

Now, consider the abstract problem (6.29) of finding $u \in V$ such that

$$a(u, v) \;=\; \ell(v) \qquad \forall v \in V$$

and its approximation (6.33) of finding $u_h \in V_h$ such that

$$a(u_h, v_h) \;=\; \ell(v_h) \qquad \forall\, v_h \in V_h$$

(By "abstract" we mean that $a(.,.)$ and $\ell(.)$ are not neccessarily as defined in (6.30).)

Theorem 6.4.6: Assume a is a bilinear form and ℓ is linear. Let u solve problem (6.29) and let u_h solve problem (6.33). Then

$$a(u - u_h, v_h) \;=\; 0 \qquad \forall v_h \in V_h, \tag{6.42}$$

i.e. the error $u - u_h$ is orthogonal to the approximating space V_h with respect to the bilinear form $a(\cdot, \cdot)$. If in addition a is an inner product on V, then

$$\|u - u_h\|_a \;\leq\; \|u - v_h\|_a \qquad \forall v_h \in V_h, \tag{6.43}$$

i.e. u_h is the best approximation to u contained in V_h.

Proof. Since $V_h \subset V$, (6.29) implies

$$a(u, v_h) = \ell(v_h) \quad \forall v_h \in V_h\,.$$

Subtracting (6.33) from this and using the linearity of a, it follows that

$$a(u-u_h,v_h)=0 \quad \forall v_h \in V_h\,.$$

Hence, if a is an inner product it follows, for all $v_h \in V_h$, that

$$\begin{aligned}\|u-u_h\|_a^2 &= a(u-u_h,u-u_h)\\ &= a(u-u_h,u-v_h)+\underbrace{a(u-u_h,\overbrace{v_h-u_h}^{\in V_h})}_{=0} \ \le\ \|u-u_h\|_a\|u-v_h\|_a\end{aligned}$$

using the Cauchy-Schwarz inequality (6.41). Now either $\|u-u_h\|_a=0$ and (6.43) holds trivially, or we can divide by $\|u-u_h\|_a$ to obtain the result. □

6.4.4 Application

Returning to the differential equation (6.27) and its weak form (6.29) and (6.30) on the space

$$V \ := \ \big\{v\in C[0,1] : v' \text{ bounded \& piecewise continuous on } [0,1] \text{ and } v(0)=0=v(1)\big\}$$

with

$$a(u,v) \ := \ \int_0^1 ku'v'\,dx \quad\text{and}\quad \ell(v) \ := \ \int_0^1 fv\,dx\,.$$

It is easy to show (similarly to (6.31)) that a is an inner product on V. $\boxed{\text{DIY}}$

Let (6.33) be the approximation of (6.29) in the (finite-dimensional) space V_h of piecewise linear finite elements in (6.39). Then, Theorem 6.4.6 implies

$$\|u-u_h\|_a \ \le\ \|u-v_h\|_a \qquad \forall v_h\in V_h\,,$$

that is

$$\left\{\int_0^1 k|(u-u_h)'|^2\right\}^{1/2} \ \le\ \left\{\int_0^1 k|(u-v_h)'|^2\right\}^{1/2} \qquad \forall v_h\in V_h\ . \tag{6.44}$$

The LHS of (6.44) is a measure of the error $u-u_h$, where u is the unknown exact solution. To prove $u_h\to u$ as $h\to 0$ we need to pick a clever v_h so that the RHS can be estimated in terms of some power of h. For this we typically make use of the piecewise linear *interpolant* of u. For any $v\in C[0,1]$ we introduce

$$(\Pi_h v)=\sum_{i=1}^n v(x_i)\,\varphi_i$$

with φ_i the hat functions given in Section 6.4.2. Then $\Pi_h v\in V_h$ and $\Pi_h v(x_j)=v(x_j)$ for $j=1,\dots,n$, i.e. $\Pi_h v$ interpolates v at the points x_j, $j=1,\dots,n$.

Definition 6.4.7: For any $v\in C[0,1]$, we define the **uniform (or infinity) norm**

$$\|v\|_\infty \ :=\ \max_{x\in[0,1]} |v(x)|\,.$$

Lemma 6.4.8: Suppose u solves (6.29) and suppose $u \in C^2[0,1]$. Then

$$\|(u - \Pi_h u)'\|_\infty \ \leq \ h\,\|u''\|_\infty\,.$$

Before we prove this lemma, let us note the consequence:

Theorem 6.4.9: Let u solve problem (6.29) and let u_h solve problem (6.33) in the piecewise linear subspace V_h defined in (6.39). Then

$$\|u - u_h\|_a \ \leq \ h\,\|k\|_\infty^{1/2}\,\|u''\|_\infty\,.$$

Proof. Squaring (6.44) and choosing $v_h = \Pi_h u$, it follows from Lemma 6.4.8 that

$$\begin{aligned}\|u - u_h\|_a^2 \ &\leq \ \int_0^1 k(x)|(u - \Pi_h u)'(x)|^2 dx \\ &\leq \ \|k\|_\infty \|(u - \Pi_h u)'\|_\infty^2 \left\{\int_0^1 dx\right\} \ \leq \ h^2 \|k\|_\infty \|u''\|_\infty^2\end{aligned}$$

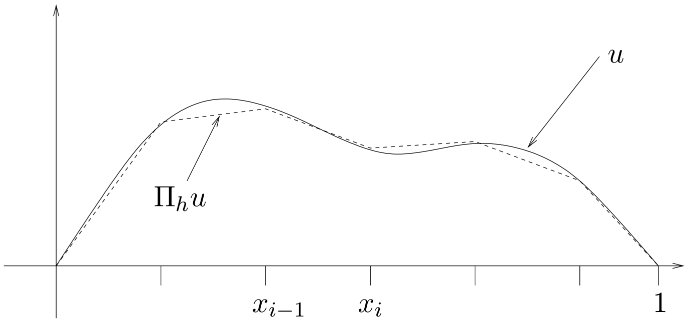
Taking the square root yields the result. □

Remark 6.4.10. Theorem 6.4.9 shows that if u has two continuous derivatives then $u_h \to u$ as $h \to 0$. It follows immediately by rearranging (6.28) that u'' is continuous under the assumptions on the data k and f made at the beginning of Section 6.4. In the norm induced by a, the convergence is linear. In the infinity-norm it can in fact be shown that

$$\|u - u_h\|_\infty = \mathcal{O}(h^2)\,.$$

(Note that $\Pi_h u$ is used for theoretical reasons only. In general we do not know u or $\Pi_h u$.)

Proof of Lemma 6.4.8.



Let $d_h = (u - \Pi_h u)$ and let $\tau_i := (x_{i-1}, x_i)$. Applying the Mean Value Theorem (MVT) on τ_i (see **Numerik 0**), there exists some $\xi_i \in (x_{i-1}, x_i)$ such that

$$d_h'(\xi_i) = \frac{\overbrace{d_h(x_i)}^{=0} - \overbrace{d_h(x_{i-1})}^{=0}}{x_i - x_{i-1}} = 0$$

Hence, for $x \in \tau_i \backslash \{\xi_i\}$, using the MVT again, it follows that there exists η_i between x and ξ_i such that

$$\frac{d_h'(x) - \overbrace{d_h'(\xi_i)}^{=0}}{x - \xi_i} = d_h''(\eta_i) = u''(\eta_i),$$

since $\Pi_h u$ is linear on τ_i and thus $(\Pi_h u)''(\eta_i) = 0$. So, we have shown that for any $x \in \tau_i$ there exists $\eta_i, \xi_i \in (x_{i-1}, x_i)$ such that

$$d_h'(x) = (x - \xi_i)u''(\eta_i).$$

In particular, this implies that

$$\max_{x\in\tau_i} |(u - \Pi_h u)'(x)| \; = \; \Big(\max_{x\in\tau_i} |x - \xi_i|\Big)|u''(\eta_i)| \; \le \; h \max_{x\in\tau_i} |u''(x)| \, .$$

Taking the maximum over $i = 1, \ldots, n$ leads to the desired result. □

Remark 6.4.11. (a) This completes the basic description and analysis of piecewise linear finite elements in one dimension.

(b) However, the real power of the finite element method is the fact that the analysis extends straightforwardly also to higher-order methods, by simply exchanging the space V_h of piecewise linear functions in (6.39) by a space of piecewise polynomial functions of higher order.

(c) Furthermore, it is also easier to extend the finite element method to boundary value problems for partial differential equations (PDEs) in two or three spatial dimensions. More about that in **Numerik 2 - Finite Elements** next semester.

6.5 Iterative methods for discretised linear BVPs

As we have just seen, the discretisation of linear BVPs with finite difference or finite element methods leads to a (sparse) linear equation system to be solved for $y \in \mathbb{R}^n$:

$$Ay = b. \tag{6.45}$$

To finish this section, let us discuss how to iteratively solve such a system using the gradient method in Definition 4.2.3, as well as the significantly more efficient **conjugate gradient method** (introduced already in **Numerik 0**, **Section 6.5**).

We restrict ourselves to symmetric positive definite (SPD) $A \in \mathbb{R}^{n\times n}$, i.e.

$$A = A^\top \quad \text{and} \quad \langle Ay, y\rangle > 0 \quad \text{for all} \;\; y \in \mathbb{R}^n\backslash\{0\}.$$

Recall (again from **Numerik 0**) that all eigenvalues $\lambda_i \in \sigma(A)$ of such a SPD matrix A are real and positive, i.e.

$$0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n < \infty, \tag{6.46}$$

and

$$\|x\|_A := \langle Ax, x\rangle^{1/2}$$

defines a norm on $\mathbb{R}^n$.

Now, consider the quadratic functional

$$F(x) := \tfrac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle, \quad \text{for all} \;\; x \in \mathbb{R}^n\,. \tag{6.47}$$

Theorem 6.5.1: The vector $y \in \mathbb{R}^n$ is the unique solution of (6.45) with SPD system matrix A if and only if

$$F(y) < F(x), \quad \text{for all} \;\; x \in \mathbb{R}^n, \;\; x \ne y. \tag{6.48}$$

Proof. Let $y \in \mathbb{R}^n$ be such that $Ay = b$. Since A was assumed to be SPD, it follows for any $x \neq y$ that

$$\begin{aligned} F(x) - F(y) &= \tfrac{1}{2}\Big(\langle Ax, x\rangle - 2\langle b, x\rangle - \langle Ay, y\rangle + 2\langle b, y\rangle\Big) \\ &= \tfrac{1}{2}\Big(\langle Ax, x\rangle - 2\langle Ay, x\rangle + \langle Ay, y\rangle\Big) = \tfrac{1}{2}\big\langle A(x-y), (x-y)\big\rangle > 0. \end{aligned}$$

Conversely, if $F(y) < F(x)$ for all $x \neq y$, i.e. y is a minimum of the quadratic function F on $\mathbb{R}^n$, it is necessary that $\nabla F(y) = 0$. Thus, since

$$\frac{\partial F}{\partial x_i}(x) = \frac{\partial}{\partial x_i}\left(\tfrac{1}{2}\sum_{j,k=1}^{n} x_j A_{jk} x_k - \sum_{j=1}^{n} b_j x_j\right) = \sum_{j=1}^{n} A_{ij} x_j - b_i, \quad \text{for all} \;\; i = 1, \dots, n,$$

this implies $Ay = b$. □

6.5.1 Gradient method

Let us now apply the gradient method in Definition 4.2.3 to the quadratic function in (6.47), i.e. a descent method with search direction $s^{(k)} = -\nabla F(x^{(k)}) =: -g^{(k)}$ and with line search parameter $\alpha^{(k)} = \operatorname{argmin}_{\gamma>0} F\left(x^{(k)} - \gamma g^{(k)}\right)$ in the kth iteration.

Note that since $\nabla F(x) = Ax - b$ we have

$$\nabla F(x^{(k+1)}) = A\big(x^{(k)} - \alpha^{(k)} g^{(k)}\big) - b = g^{(k)} - \alpha^{(k)} A g^{(k)}$$

and the minimum $\alpha^{(k)}$ of $F\left(x^{(k)} - \gamma g^{(k)}\right)$ is attained when γ satisfies

$$\begin{aligned} 0 &= \frac{\partial}{\partial \gamma} F\big(x^{(k)} - \gamma g^{(k)}\big) \\ &= -\Big\langle \nabla F\big(x^{(k)} - \gamma g^{(k)}\big), g^{(k)}\Big\rangle \\ &= \langle b - Ax^{(k)}, g^{(k)}\rangle + \gamma\langle Ag^{(k)}, g^{(k)}\rangle = \langle g^{(k)}, g^{(k)}\rangle + \gamma\langle Ag^{(k)}, g^{(k)}\rangle. \end{aligned}$$

This leads to the following iteration for finding the solution $y \in \mathbb{R}^n$ of $Ay = b$.

Definition 6.5.2 (Gradient method): Given an initial value $x^{(0)} \in \mathbb{R}^n$ and $g^{(0)} = Ax^{(0)} - b$, compute iterates $x^{(k)}$, $k = 1, 2, \dots$ such that

$$\begin{aligned} \alpha^{(k)} &= \frac{\langle g^{(k)}, g^{(k)}\rangle}{\langle Ag^{(k)}, g^{(k)}\rangle} \\ x^{(k+1)} &= x^{(k)} - \alpha^{(k)} g^{(k)}, \\ g^{(k+1)} &= g^{(k)} - \alpha^{(k)} A g^{(k)}. \end{aligned} \tag{6.49}$$

Theorem 6.5.3: Suppose A is SPD. Then, the gradient method in Definition 6.5.2 converges for any $x^{(0)} \in \mathbb{R}^n$ to the solution of (6.45) and the error in the A-norm satisfies the following bound:

$$\|x^{(k)} - y\|_A \le \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^k \|x^{(0)} - y\|_A\,, \quad \text{for all} \;\; k \in \mathbb{N}. \tag{6.50}$$

Proof. Let $e^{(k)} := x^{(k)} - y$. Then, it follows that

$$\begin{aligned}
\frac{\|e^{(k)}\|_A^2 - \|e^{(k+1)}\|_A^2}{\|e^{(k)}\|_A^2} &= \frac{\big\langle Ae^{(k)}, e^{(k)}\big\rangle - \big\langle A\big(e^{(k)} - \alpha^{(k)} g^{(k)}\big), e^{(k)} - \alpha^{(k)} g^{(k)}\big\rangle}{\big\langle Ae^{(k)}, e^{(k)}\big\rangle} \\
&= \frac{2\alpha^{(k)}\big\langle Ag^{(k)}, e^{(k)}\big\rangle - \big(\alpha^{(k)}\big)^2\big\langle Ag^{(k)}, g^{(k)}\big\rangle}{\big\langle Ae^{(k)}, e^{(k)}\big\rangle} \\
&= \frac{2\alpha^{(k)}\big\langle g^{(k)}, g^{(k)}\big\rangle - \big(\alpha^{(k)}\big)^2\big\langle Ag^{(k)}, g^{(k)}\big\rangle}{\big\langle A^{-1}g^{(k)}, g^{(k)}\big\rangle} \\
&= \frac{\big\langle g^{(k)}, g^{(k)}\big\rangle^2}{\big\langle Ag^{(k)}, g^{(k)}\big\rangle\big\langle A^{-1}g^{(k)}, g^{(k)}\big\rangle} \;=:\; \varrho\big(g^{(k)}\big),
\end{aligned}$$

where we have used the fact that $Ae^{(k)} = Ax^{(k)} - b = g^{(k)}$. Rearranging this, we can conclude that

$$\|e^{(k+1)}\|_A^2 = \Big(1 - \varrho\big(g^{(k)}\big)\Big)\|e^{(k)}\|_A^2 \tag{6.51}$$

Since

$$\lambda_1\|x\|^2 \le \langle Ax, x\rangle \le \lambda_n\|x\|^2 \quad \text{and} \quad \lambda_n^{-1}\|x\|^2 \le \langle A^{-1}x, x\rangle \le \lambda_1^{-1}\|x\|^2,$$

it follows immediately that

$$\varrho(x) \ge \lambda_1/\lambda_n \quad \text{for all} \;\; 0 \ne x \in \mathbb{R}^n.$$

Together with (6.51) this is in fact sufficient, to prove convergence for any $x^{(0)} \in \mathbb{R}^n$, since $0 < \lambda_1/\lambda_n \le 1$. However, to obtain the error bound (6.50) we need the sharper lower bound

$$\varrho(x) \ge \frac{4\lambda_n\lambda_1}{(\lambda_n + \lambda_1)^2} \quad \text{for all} \;\; 0 \ne x \in \mathbb{R}^n, \tag{6.52}$$

which can be proved by first expanding the vector x in the eigenbasis. It is referred to as the *Kantorovich Inequality*, but we will skip its proof (see e.g. [Ran17a, Hilfssatz 6.4]).

Now, using the bound (6.52) in (6.51) we get

$$\|e^{(k+1)}\|_A^2 \le \left(1 - \frac{4\lambda_n\lambda_1}{(\lambda_n + \lambda_1)^2}\right)\|e^{(k)}\|_A^2 = \left(\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}\right)^2\|e^{(k)}\|_A^2,$$

which implies (6.50). □

Example 6.5.4. Let us consider the finite difference system from Example 6.3.7 with $[a, b] = [0, 1]$ and $\beta_k = \gamma_k = 0$, i.e.

$$L_h\, y = h^{-2}\begin{pmatrix} 2 & -1 & & \\ -1 & 2 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ \vdots \\ f_{n-1} \end{pmatrix} = f_h.$$

with $h = 1/n$. Note that the finite element stiffness matrix A in (6.40), on the same uniform mesh as in Example 6.3.7 and with $k \equiv 1$, satisfies $A = hL_h$.

Using the trigonometric identities, it can be shown that the eigenvalues and eigenvectors of L_h are given by $\boxed{\text{DIY}}$

$$\lambda_k = 4h^{-2}\sin^2\Big(\frac{\pi}{2}kh\Big) \quad \text{and} \quad v_k = \Big(\sin(\pi kh), \sin(2\pi kh), \dots, \sin((n-1)\pi kh)\Big)^\top,$$

for $k = 1, \dots, n-1$.

Using the identity $2\sin^2(\vartheta) = 1 - \cos(2\vartheta)$, we can then compute the convergence rate for the gradient method in (6.50) as

$$\begin{aligned}\frac{\lambda_{n-1} - \lambda_1}{\lambda_{n-1} + \lambda_1} &= \frac{2\sin^2\big(\frac{\pi}{2}(n-1)h\big) - 2\sin^2\big(\frac{\pi}{2}h\big)}{2\sin^2\big(\frac{\pi}{2}(n-1)h\big) + 2\sin^2\big(\frac{\pi}{2}h\big)} \\ &= \frac{\cos(\pi h) - \cos(\pi - \pi h)}{2 - \cos(\pi h) - \cos(\pi - \pi h)} = \frac{2\cos(\pi h)}{2 - \cos(\pi h) + \cos(\pi h)} = \cos(\pi h).\end{aligned}$$

Taylor expanding $\cos(\pi h)$ around 0, we can finally deduce that

$$\begin{aligned}\frac{\lambda_{n-1} - \lambda_1}{\lambda_{n-1} + \lambda_1} &= \underbrace{\cos(0)}_{=1} - \underbrace{\sin(0)}_{=0}\pi h - \underbrace{\cos(0)}_{=1}\frac{\pi^2 h^2}{2} + \underbrace{\sin(0)}_{=0}\frac{\pi^3 h^3}{6} + \mathcal{O}(h^4) \\ &= 1 - \frac{\pi^2}{2}h^2 + \mathcal{O}(h^4)\end{aligned}$$

Thus, the convergence rate of the gradient method applied to the finite difference system becomes quickly very poor: For $h = 1/10$ the convergence rate is 0.951, but for $h = 1/40$ the convergence rate is already 0.997 and over 700 iterations are necessary to reduce the error just by a factor 1/10 (indeed $0.997^{745} \approx 0.1$).

6.5.2 Conjugate gradient method

The gradient method only uses the structure of the quadratic functional $F(\cdot)$ (i.e. the distribution of the eigenvalues of A) only locally from one iteration to the next. It would be better to incorporate the already gained information by orthogonalising the descent directions with respect to previous directions. This is the basic idea of the conjugate gradient (CG) method by Hestenes & Stiefel [HS52].

Note first that

$$\begin{aligned}2F(x) &= \langle Ax, x\rangle - \langle b, x\rangle - \langle Ay, x\rangle \\ &= \langle Ax - b, x - y\rangle - \langle b, y\rangle = \|x - y\|_A^2 - \|y\|_A^2,\end{aligned}$$

so that minimising F is equivalent to minimising $\|x - y\|_A$.

To derive the CG-method consider a general descent method, as in Definition 4.2.1, such that

$$x^{(k)} = x^{(0)} + \sum_{j=1}^{k-1} \alpha^{(j)} s^{(j)},$$

where the stepsizes $\alpha^{(j)}$, $j = 1, \dots, k-1$, are chosen such as to minimise $F(x)$ over the subspace

$$x^{(0)} + B_k := x^{(0)} + \operatorname{span}\{s^{(0)}, \dots, s^{(k-1)}\} \tag{6.53}$$

with yet to be specified, linearly independent search directions $s^{(j)}$, i.e.

$$F(x^{(k)}) = \min_{x \in x^{(0)} + B_k} F(x) \quad \Leftrightarrow \quad \|x^{(k)} - y\|_A = \min_{x \in x^{(0)} + B_k} \|x - y\|_A \,. \tag{6.54}$$

By setting the derivatives of F with respect to the parameters $\alpha^{(j)}$ to 0, we can derive the following (necessary and sufficient) orthogonality conditions for the gradient $g^{(k)} = Ax^{(k)} - b$ at the kth iteration:

$$\left\langle g^{(k)}, s^{(j)} \right\rangle = \left\langle Ax^{(k)} - b, s^{(j)} \right\rangle = \frac{\partial F}{\partial \alpha^{(j)}}(x^{(k)}) = 0, \quad j = 0, \ldots, k-1. \tag{6.55}$$

A natural choice for the space B_k is the so-called **Krylov-subspace**

$$B_k = K_k(s^{(0)}; A) := \operatorname{span}\{s^{(0)}, As^{(0)}, \ldots, A^{k-1}s^{(0)}\}, \quad \text{with} \quad s^{(0)} := b - Ax^{(0)}. \tag{6.56}$$

The motivation for this choice is the following lemma.

Lemma 6.5.5: $A^k s^{(0)} \in K_k(s^{(0)}; A)$ implies $g^{(k)} = 0$ (i.e. $x^{(k)} = y$).

Proof. If $A^k s^{(0)} \in K_k(s^{(0)}; A)$ then

$$g^{(k)} = A(x^{(k)} - x^{(0)}) - s^{(0)} \in -s^{(0)} + AK_k(s^{(0)}; A) \subset K_k(s^{(0)}; A).$$

However, since (6.55) implies $\langle g^{(k)}, s \rangle = 0$, for all $s \in K_k(s^{(0)}; A)$, we get $g^{(k)} = 0$. □

In the CG method, given a starting vector $x^{(0)}$, the search directions $s^{(j)}$ are now constructed by inductively building an A-orthogonal basis of $K_k(s^{(0)}; A)$:

$\boxed{k = 0}$ We set $s^{(0)} = b - Ax^{(0)}$.

$\boxed{k \to k+1}$ Let $\{s^{(0)}, \ldots, s^{(k-1)}\}$ be an A-orthogonal basis of $K_k(s^{(0)}; A)$ and let $x^{(k)} \neq y$, such that $g^{(k)} \neq 0$ (otherwise we are finished). Then we construct $s^{(k)} \in K_{k+1}(s^{(0)}; A)$ such that

$$\langle As^{(k)}, s^{(j)} \rangle = 0\,, \quad \text{for all } \; j = 0, \ldots, k-1, \tag{6.57}$$

using the following ansatz:

$$s^{(k)} = -g^{(k)} + \sum_{j=0}^{k-1} \beta_j s^{(j)} \in K_{k+1}(s^{(0)}; A)\,, \tag{6.58}$$

for some $\beta_0, , \ldots, \beta_{k-1}$. Note that $g^{(k)} \notin K_k(s^{(0)}; A)$, since otherwise $g^{(k)} = 0$ (as in the proof of Lemma 6.5.5). Substituting in (6.57) we get for all $i = 0, \ldots, k-1$ that

$$0 = \langle s^{(k)}, As^{(i)} \rangle = \left\langle -g^{(k)} + \sum_{j=0}^{k-1} \beta_j s^{(j)}, As^{(i)} \right\rangle = -\langle g^{(k)}, As^{(i)} \rangle + \beta_i \langle s^{(i)}, As^{(i)} \rangle\,.$$

From (6.55) it follows that $\langle g^{(k)}, As^{(i)} \rangle = 0$ and thus $\beta_i = 0$, for all $i = 0, \ldots, k-2$. For $i = k-1$ we get

$$\beta^{(k-1)} := \beta_{k-1} = \frac{\langle g^{(k)}, As^{(k-1)} \rangle}{\langle s^{(k-1)}, As^{(k-1)} \rangle} \quad \text{and} \quad s^{(k)} = -g^{(k)} + \beta^{(k-1)} s^{(k-1)}. \tag{6.59}$$

For the next stepsize $\alpha^{(k)}$ it follows from (6.55) that

$$0 = \Big\langle Ax^{(k+1)} - b, s^{(k)} \Big\rangle = \Big\langle \underbrace{Ax^{(k)} - b}_{=g^{(k)}} + \alpha^{(k)} As^{(k)}, s^{(k)} \Big\rangle = \Big\langle g^{(k)}, s^{(k)} \Big\rangle + \alpha^{(k)} \Big\langle As^{(k)}, s^{(k)} \Big\rangle$$

and thus $x^{(k+1)} = x^{(k)} + \alpha^{(k)} s^{(k)}$ and $g^{(k+1)} = g^{(k)} + \alpha^{(k)} As^{(k)}$ with

$$\alpha^{(k)} = -\frac{\langle g^{(k)}, s^{(k)} \rangle}{\langle s^{(k)}, As^{(k)} \rangle}\,. \tag{6.60}$$

Finally, we can exploit the orthogonality relations

$$\langle g^{(k+1)}, g^{(k)} \rangle = 0 \quad \text{and} \quad \langle g^{(k+1)}, s^{(j)} \rangle = \langle s^{(k+1)}, As^{(j)} \rangle = 0, \quad 0 \le j \le k,$$

to deduce

$$\alpha^{(k)} \Big\langle s^{(k)}, As^{(k)} \Big\rangle = \Big\langle s^{(k)}, \underbrace{\alpha^{(k)} As^{(k)} - g^{(k+1)}}_{=-g^{(k)}} \Big\rangle = \Big\langle \underbrace{s^{(k)} - \beta^{(k)} s^{(k-1)}}_{=-g^{(k)}}, -g^{(k)} \Big\rangle = \Big\langle g^{(k)}, g^{(k)} \Big\rangle$$

$$\alpha^{(k)} \Big\langle As^{(k)}, g^{(k+1)} \Big\rangle = \Big\langle g^{(k)} + \alpha^{(k)} As^{(k)}, g^{(k+1)} \Big\rangle = \Big\langle g^{(k+1)}, g^{(k+1)} \Big\rangle$$

which allows to simplify the formulae for $\alpha^{(k)}$ and for $\beta^{(k)}$ in (6.60) and (6.59) to arrive at the following algorithm.

Definition 6.5.6 (Conjugate Gradient (CG) method): Given an initial value $x^{(0)} \in \mathbb{R}^n$ and $s^{(0)} = -g^{(0)} = Ax^{(0)} - b$, compute iterates $x^{(k)}$, $k = 1, 2, \ldots$ such that

$$\begin{aligned} \alpha^{(k)} &= \frac{\langle g^{(k)}, g^{(k)} \rangle}{\langle As^{(k)}, s^{(k)} \rangle} \\ x^{(k+1)} &= x^{(k)} + \alpha^{(k)} s^{(k)}, \\ g^{(k+1)} &= g^{(k)} + \alpha^{(k)} As^{(k)}, \\ \beta^{(k)} &= \frac{\langle g^{(k+1)}, g^{(k+1)} \rangle}{\langle g^{(k)}, g^{(k)} \rangle} \\ s^{(k+1)} &= -g^{(k)} + \beta^{(k)} s^{(k)}. \end{aligned} \tag{6.61}$$

By construction, unless the algorithm finishes early with $x^{(k)} = y$ for some $k < n$, the directions $s^{(0)}, \ldots, s^{(n-1)}$ produced by the CG method are A-orthogonal and thus linearly independent. Hence, they form a basis of $\mathbb{R}^n$ and we can collect all the properties derived so far in the following theorem.

Theorem 6.5.7: Suppose $A \in \mathbb{R}^{n\times n}$ is SPD. In exact arithmetic, the CG method terminates for any $x^{(0)} \in \mathbb{R}^n$ after $K \le n$ iterations with $x^{(K)} = y$. In each iteration, we have

$$\|x^{(k)} - y\|_A = \min_{x \in x^{(0)} + K_k} \|x - y\|_A$$

with $K_k := \operatorname{span}\{s^{(0)}, \ldots, s^{(k-1)}\} = \operatorname{span}\{s^{(0)}, \ldots, A^{(k-1)} s^{(0)}\}$.

Remark 6.5.8. As we have just seen, the CG method is in fact a "direct method" and provides an explicit solution in at most n iterations. However, in practice it is always used as an iterative method, since the A-orthogonality may not be guaranteed due to rounding errors and the method leads to good approximations already after significantly fewer than n iterations for large matrices.

Finally we can also establish an error bound and a convergence rate in the A-norm as for the gradient method.

Theorem 6.5.9: Suppose $A \in \mathbb{R}^{n\times n}$ is SPD and $x^{(0)} \in \mathbb{R}^n$. Then, the iterates produced by the CG method in Definition 6.5.6 satisfy the following bound in the A-norm:

$$\|x^{(k)} - y\|_A \leq \left(\frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}}\right)^k \|x^{(0)} - y\|_A\,, \quad \text{for all} \;\; k \in \mathbb{N}. \tag{6.62}$$

Proof. See [Ran17a, Satz 6.10] (**not examinable**). □

As for the gradient method, the convergence rate becomes very poor when $\lambda_n \gg \lambda_1$. However, the number of iterations grows significantly slower than for the gradient method and in contrast to the bound for the gradient method in Theorem 6.5.3, which is attained in practice, the bound for the CG method in Theorem 6.5.9 is often rather pessimistic.

Example 6.5.10. As for the gradient method, we can also compute the convergence rate for the CG method in Theorem 6.5.9 for the finite difference system in Example 6.5.4 explicitly. Using the eigenvalues derived there and Taylor expanding, we get $\boxed{\text{DIY}}$

$$\begin{aligned}\frac{\sqrt{\lambda_{n-1}} - \sqrt{\lambda_1}}{\sqrt{\lambda_{n-1}} + \sqrt{\lambda_1}} &= \frac{\sin\left(\frac{\pi}{2} - \frac{\pi}{2}h\right) - \sin\left(\frac{\pi}{2}h\right)}{\sin\left(\frac{\pi}{2} - \frac{\pi}{2}h\right) + \sin\left(\frac{\pi}{2}h\right)} = \frac{\cos\left(\frac{\pi}{2}h\right) - \sin\left(\frac{\pi}{2}h\right)}{\cos\left(\frac{\pi}{2}h\right) + \sin\left(\frac{\pi}{2}h\right)} \\ &= 1 - \pi h + \frac{\pi^2}{2}h^2 + \mathcal{O}(h^3).\end{aligned}$$

Thus, the rate degenerates significantly slower with $h \to 0$.

Finally, in Table 6.1, let us compare the rates for the two iterative methods on the problem in Example 6.5.4, as well as the actual numbers of iterations to reduce the error in the A-norm by a factor of 10^{-4}, starting with $x^{(0)} = 0$. The right hand side is chosen to be $f(x) = \frac{\pi^2}{4}\sin(\frac{\pi}{2}x)$, so that the exact solution is given by $u(x) = \sin(\frac{\pi}{2}x) - x$.

In the second and third column in the table we present the discretisation errors in the energy norm and in the uniform norm, respectively. In Columns 4 and 5, we see the rate of convergence and the iterations for the gradient method. As predicted, the convergence is extremely slowly, even for the smaller systems. The CG method converges in exactly $n = 1/h - 1$ iterations for $h = 1/10, 1/20, 1/40$. For the larger two systems the iteration terminates for $K < n$.

	Error		Gradient Method		CG Method	
$1/h$	$\|u-u_h\|_a$	$\|u-u_h\|_\infty$	rate	# iterations	rate	# iterations
10	4.36×10^{-2}	2.98×10^{-3}	0.9511	174	0.727	9
20	2.18×10^{-2}	7.58×10^{-4}	0.9877	722	0.854	19
40	1.09×10^{-2}	1.91×10^{-4}	0.9969	2935	0.925	39
80	5.45×10^{-3}	4.80×10^{-5}	0.9992	11829	0.962	78
160	2.73×10^{-3}	1.20×10^{-5}	0.9998	47486	0.981	156

Table 6.1: Discretisation errors in energy and uniform norm for the Poisson problem in Example 6.5.4 (Columns 2–3). Comparison of the convergence rates and the numbers of iterations to reduce the error by a factor 10^{-4} for the gradient method (Columns 4–5) and for the CG method (Columns 6–7).

Remark 6.5.11. (a) In practice, the CG method is typically used in conjunction with a so-called **preconditioner**, a matrix M that is a "cheap" approximation of the inverse of A, such that the system matrix MA of the preconditioned system $MAy = Mb$ has a narrower spectrum, i.e.

$$\frac{\lambda_n(MA)}{\lambda_1(MA)} < \frac{\lambda_n(A)}{\lambda_1(A)}.$$

(b) The symmetry and the positive definiteness of A was not essential. There are other Krylov subspace methods, such as the **Generalised Minimal Residual (GMRES)** method, that are extensions of CG to general nonsymmetric systems with invertible, square matrices A.

Chapter 7

Outlook towards partial differential equations (not examinable)

Finite difference methods for two-point boundary value problems have a natural extension to higher dimensions. There, we deal with partial derivatives $\frac{\partial}{\partial x_1}$, $\frac{\partial}{\partial x_2}$, $\frac{\partial}{\partial x_3}$ and $\frac{\partial}{\partial t}$.

As an outlook towards topics in the numerical analysis of partial differential equations, we close these notes by a short introduction by means of some examples.

7.1 The Laplacian and harmonic functions

Definition 7.1.1: The **Laplacian** in two (three) space dimensions is the sum of the second partial derivatives

$$\Delta u = \frac{\partial^2}{\partial x_1^2}u + \frac{\partial^2}{\partial x_2^2}u \left(+ \frac{\partial^2}{\partial x_3^2}u \right) \tag{7.1}$$

The **Laplace equation** is the partial differential equation

$$-\Delta u = 0. \tag{7.2}$$

The **Poisson equation** is the partial differential equation

$$-\Delta u = f. \tag{7.3}$$

Solutions to the Laplace equations are called **harmonic functions**.

Theorem 7.1.2 (Mean-value formula for harmonic functions): Let $\Omega \subset \mathbb{R}^d$ and let $u \in C^2(\Omega)$ be a solution to the Laplace equation on Ω. Then, u has the mean value property

$$u(\mathbf{x}) = \frac{1}{r^{d-1}\omega(d)} \int_{\partial B_r(\mathbf{x})} u(\mathbf{y})\, ds, \tag{7.4}$$

where $\partial B_r(\mathbf{x}) \subset \Omega$ is the sphere of radius r around $\mathbf{x}$ and $\omega(d)$ is the volume of the unit sphere in $\mathbb{R}^d$.

Proof. First, we rescale the problem to

$$\Phi(r) = \frac{1}{r^{d-1}\omega(d)} \int_{\partial B_r(\mathbf{x})} u(\mathbf{y})\, ds = \frac{1}{\omega(d)} \int_{\partial B_1(0)} u(\mathbf{x} + r\mathbf{z})\, ds.$$

Then, it follows by the Gauß theorem for the vector valued function ∇u that

$$\begin{aligned}
\Phi'(r) &= \frac{1}{\omega(d)} \int_{\partial B_1(0)} \nabla u(\mathbf{x} + r\mathbf{z}) \cdot \mathbf{z}\, ds_z \\
&= \frac{1}{r^{d-1}\omega(d)} \int_{\partial B_r(\mathbf{x})} \nabla u(\mathbf{y}) \cdot \frac{\mathbf{y} - \mathbf{x}}{r}\, ds_y \\
&= \frac{1}{r^{d-1}\omega(d)} \int_{\partial B_r(\mathbf{x})} \frac{\partial}{\partial \mathbf{n}} u(\mathbf{y})\, ds_y \\
&= \frac{1}{r^{d-1}\omega(d)} \int_{B_r(\mathbf{x})} \Delta u(\mathbf{y})\, d\mathbf{y} = 0.
\end{aligned}$$

Therefore, $\Phi(r)$ is constant. Because of continuity, we have

$$\lim_{r\to 0} \Phi(r) = \lim_{r \to 0} \frac{1}{r^{d-1}\omega(d)} \int_{\partial B_r(\mathbf{x})} u(\mathbf{y})\, ds = u(\mathbf{x}),$$

which proves our theorem. □

Theorem 7.1.3 (Maximum principle): Let a function $u \in C^2(\Omega)$ be a solution to the Laplace equation on an open, bounded, connected domain $\Omega \subset \mathbb{R}^d$. Then, if there is an interior point $\mathbf{x}_0$ of Ω, such that for a neighborhood $U \subset \Omega$ of $\mathbf{x}_0$ there holds

$$u(\mathbf{x}_0) \ge u(\mathbf{x}) \qquad \forall \mathbf{x} \in U,$$

then the function is constant in Ω.

Proof. Let $\mathbf{x}_0$ be such a local maximum. Then, there exists a $R > 0$ such that $B_r(\mathbf{x}_0) \subset \Omega$ and $u(\mathbf{x}_0) \ge u(\mathbf{x})$ for all $\mathbf{x} \in B_r(\mathbf{x}_0)$ and for all $0 < r \le R$. Assume that there is a point $\mathbf{x}$ on $\partial B_r(\mathbf{x}_0)$, such that $u(\mathbf{x}) < u(\mathbf{x}_0)$. Then, this holds for points $\mathbf{y}$ in a neighborhood of $\mathbf{x}$. Thus, in order that the mean value property holds, there must be a subset of $\partial B_r(\mathbf{x}_0)$ where $u(\mathbf{y}) > u(\mathbf{x}_0)$, contradicting that $\mathbf{x}_0$ is a maximum. Thus, $u(\mathbf{x}) = u(\mathbf{x}_0)$ for all $\mathbf{x} \in \partial B_r(\mathbf{x}_0)$. Since $r \le R$ was arbitrary this implies $u(\mathbf{x}) = u(\mathbf{x}_0)$ for all $\mathbf{x} \in B_R(\mathbf{x}_0)$.

Let now $\mathbf{x} \in \Omega$ be arbitrary. Then, there is a (compact) path from $\mathbf{x}_0$ to $\mathbf{x}$ in Ω. Thus, the path can be covered by a finite set of overlapping balls inside Ω, and the argument above can be used iteratively to conclude $u(\mathbf{x}) = u(\mathbf{x}_0)$. □

Corollary 7.1.4. *Let $u \in C^2(\Omega)$ be a solution to the Laplace equation. Then, its maximum and its minimum lie on the boundary, that is, there are points $\underline{\mathbf{x}}, \overline{\mathbf{x}} \in \partial\Omega$, such that*

$$u(\underline{\mathbf{x}}) \le u(\mathbf{x}) \le u(\overline{\mathbf{x}}) \quad \forall \mathbf{x} \in \Omega.$$

Proof. If the maximum of u is attained in an interior point, the maximum principle yields a constant solution and the theorem holds trivially. On the other hand, theorem 7.1.3 does not make any prediction on points at the boundary, which therefore can be maxima. The same holds for the minimum, since $-u$ is also a solution to the Laplace equation. □

Corollary 7.1.5. *Let $u, v \in\in C^2(\Omega)$ be two solutions of the Poisson equation (7.3) with homogeneous boundary conditions $u = v \equiv 0$ on $\partial\Omega$. Then $u = v$ on Ω.*

Proof. Assume there are two functions $u, v \in C^2(\Omega)$ with $u = v = 0$ on $\partial\Omega$ such that

$$-\Delta u = -\Delta v = f.$$

Then $w = u - v$ solves the Laplace equation, i.e. $-\Delta w = -\Delta u + \Delta v = 4$, and $w = 0$ on $\partial\Omega$. Due to the maximum principle, $w \equiv 0$ and thus $u = v$. □

The proof of existence of a solution of the Poisson equation is more involved and we will not address it here.

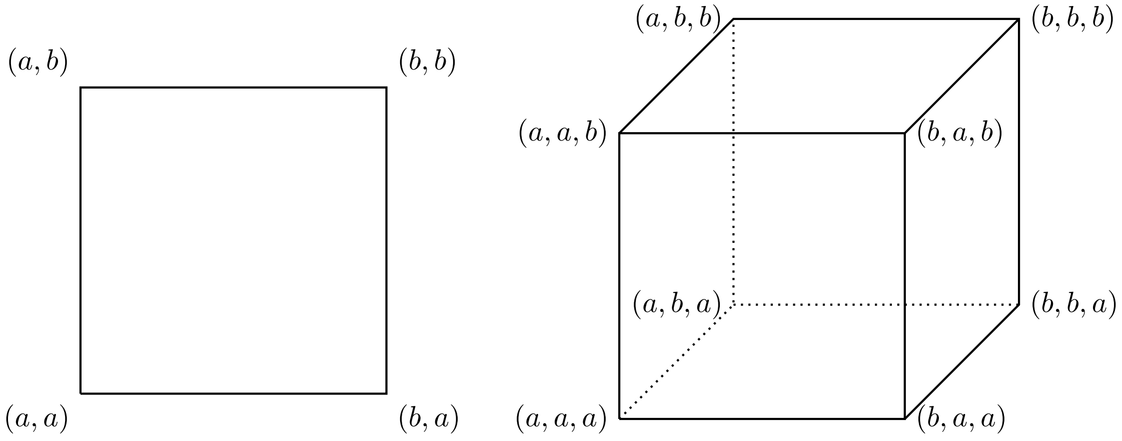
7.2 Finite difference methods in higher dimensions

7.2.1. Let us now study how to apply the finite difference method to solve the Poisson equation in (7.3). We consider Dirichlet boundary conditions

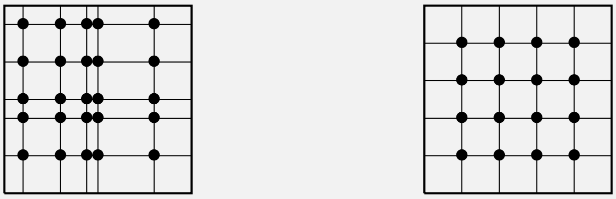
$$u(\mathbf{x}) = u_B(\mathbf{x}), \qquad \text{for } \mathbf{x} \in \partial\Omega. \tag{7.5}$$

As for two-point boundary value problems, we can reduce our considerations to homogeneous boundary conditions $u_B \equiv 0$ by changing the right hand side in the Poisson equation.

Example 7.2.2. The notion of an interval I can be extended to higher dimensions by a square $\Omega = I^2$ or a cube $\Omega = I^3$.



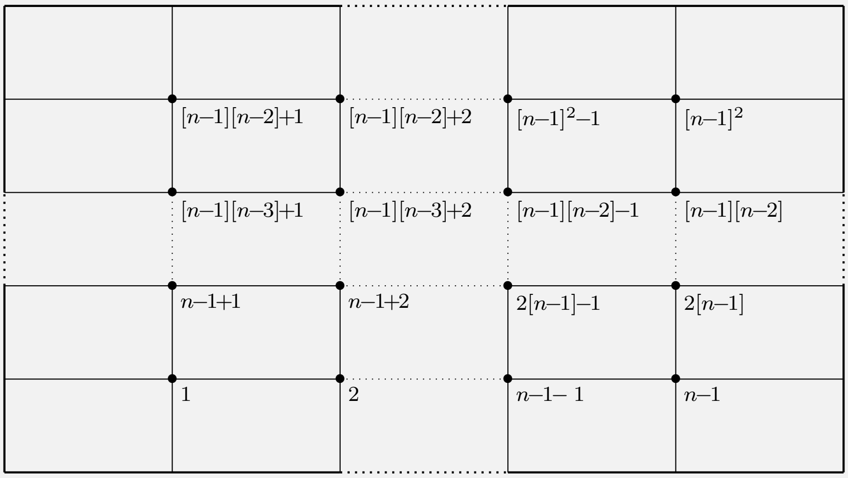
Definition 7.2.3: A **Cartesian grid** on a square (cube) domain Ω consists of the intersection points of lines (planes) parallel to the coordinate axes (planes).



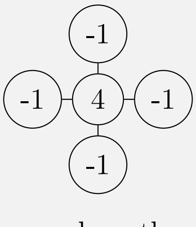
The grid is called **uniform**, if all lines (planes) are at equal distances.

For the remainder of this discussion let us restirct to the two-dimensional case, $d = 2$, and to uniform Cartesian grids.

Definition 7.2.4: The vector y of discrete values is defined in grid points which run in x_1- and x_2-direction. In order to obtain a single index for every entry of this vector in linear algebra, we use **lexicographic numbering**.



Definition 7.2.5: The **5-point stencil** consists of the sum of a 3-point stencil in x_1- and a 3-point stencil in x_2-direction. Its graphical representation is



For a generic row of the linear system, where the associated point is not neighboring the boundary, this leads to

$$D_h^2 u(\mathbf{x}^{(k)}) = \frac{-u(\mathbf{x}^{(k-n+1)}) - u(\mathbf{x}^{(k-1)}) + 4u(\mathbf{x}^{(k)}) - u(\mathbf{x}^{(k+1)}) - u(\mathbf{x}^{(k+n-1)})}{h^2} \tag{7.6}$$

If the point $\mathbf{x}^{(k)}$ is next to the boundary, the entry corresponding to the neighboring boundary point can be omitted, since the value is assumed to be zero there.

Example 7.2.6. The matrix L_h obtained for the Laplacian on $\Omega = [0,1]^2$ using the 5-point stencil on a uniform Cartesian mesh of mesh spacing $h = 1/n$ with lexicographic numbering is in $\mathbb{R}^{N\times N}$ with $N = (n-1)^2$ and has the structure

$$L_h = n^2 \begin{bmatrix} D & -I & & & \\ -I & D & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & D & -I \\ & & & -I & D \end{bmatrix}, \quad D = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}.$$

Theorem 7.2.7: The matrix L_h obtained by discretising the Laplace operator via the 5-point stencil formula is an M-matrix and the solution of the discrete problem

$$L_h y = f$$

is stable in the sense that there is a constant c independent of h such that

$$\|L_h^{-1}\|_\infty \le c.$$

Proof. The proof is identical to the proof for 2-point boundary value problems. To show boundedness of $\|L_h^{-1}\|_\infty$ we can use in a similar way the function

$$p(x_1, x_2) = x_1(1-x_1)x_2(1-x_2).$$

□

Theorem 7.2.8: The finite difference approximation in Example 7.2.6 for the Poisson equation in (7.3) on the unit square $\Omega = [0,1]^2$ with homogeneous Dirichlet conditions is convergent of second order, i.e.

$$\max_{k=1,\dots,(n-1)^2} |u(\mathbf{x}^{(k)}) - y_k| \le ch^2.$$

Proof. We apply the consistency bound in (6.13) in the x_1- and x_2-direction separately, obtaining

$$\left| \frac{\partial^2}{\partial x_1^2} u(\mathbf{x}) - \frac{u(x_1+h, x_2) - 2u(x_1,x_2) + u(x_1-h,x_2)}{h^2} \right| \le ch^2$$

$$\left| \frac{\partial^2}{\partial x_2^2} u(\mathbf{x}) - \frac{u(x_1, x_2+h) - 2u(x_1,x_2) + u(x_1,x_2-h)}{h^2} \right| \le ch^2,$$

and deduce the second-order consistency of the 5-point stencil by the triangle inequality. The remainder of the proof is identical to the proof of theorem 6.3.17. □

Theorem 7.2.9: Let y be the solution to the finite difference method for the Laplace equation with the 5-point stencil. Then, the maximum principle holds for y, namely, if there is $k \in \{1,\dots,(n-1)^2\}$ such that $y_k \ge y_j$ for all $j \ne k$ and $y_k \le y_B$ for any boundary value, then y is constant.

Proof. From equation (7.6), it is clear that a discrete mean value property holds, that is, y_k is the mean value of its four neighbors. Therefore, if $y_k \geq y_j$, for all neighboring indices j of k, we have $y_j = y_k$. We conclude by following a path through the grid points. □

7.3 Evolution equations

After an excursion to second order differential equations depending on more than one spatial variables, we are now returning to problems depending on time. But this time, on time *and* space. As for the nomenclature, we have encountered ordinary differential equations as equations or systems depending on a time variable only, then partial differential equations (PDE) with several, typically spatial, independent variables. While the problems considered here are covered by the definition of PDE, time and space are fundamentally different. Therefore, we introduce the concept of evolution equations.

While the problems in Definition 7.1.1 are PDEs of **elliptic** type. The following problems can be either **parabolic** or **hyperbolic**.

Definition 7.3.1:

(a) An equation of the form

$$\frac{\partial u}{\partial t}(t,x) = Lu(t,x), \tag{7.7}$$

where $u(t,.)$ is in a function space V on a domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$, for all time $t \in \mathbb{R}$, and $L : V \to C(\Omega)$ is a differential operator with respect to the spatial variables x only, is called a linear **evolution equation** of first order (in time).

(b) Considering for simplicity only the case $\Omega = (a,b) \subset \mathbb{R}$ of one spatial variable, the existence and uniqueness of solutions of such evolution equations can be guaranteed by specifying suitable conditions

$$u(0,x) = u_0(x) \qquad x \in \Omega \tag{7.8}$$

$$B_a u(t,a) + B_b u(t,b) = g(t) \qquad t > 0. \tag{7.9}$$

This is then referred to as an **initial boundary value problem** (**IBVP**).

7.3.1 Parabolic PDEs

Example 7.3.2. Consider the case of one spatial variable, i.e. $\Omega = (a,b) \subset \mathbb{R}$, and the differential operator $-L$ as defined in (6.4), i.e. a general, linear second order differential operator with respect to the spatial variable x, for simplicity with $\beta = \beta(x)$ and $\gamma = \gamma(x)$ independent of t. Furthermore, let $u(t,a) = u(t,b) = 0$.

This PDE is **parabolic** and for $\beta = \gamma = 0$ it is called the **heat equation**.

We can now discretise the right hand side of (7.7), for every fixed $t \geq 0$ on a spatial grid $x_0, \dots, x_n$, as in Example 6.3.7, to obtain a system of ODEs

$$y'(t) = -L_h y(t)$$

for the unknown (semi-discrete) vector $y(t) \in \mathbb{R}^{n-1}$ of approximations to the solution $u(t, \cdot)$ of (7.7) at time t. By choosing as the initial condition

$$y_k(0) = u_0(x_k), \quad k = 1, \dots, n-1$$

we obtain an autonomous linear IVP for $y : [0, T] \to \mathbb{R}^{n-1}$ that we can now solve with our favourite time stepping method.

For $\gamma_k \geq 0$ and $|\beta_k|$ sufficiently small, the eigenvalues of $-L_h$ have negative real part and vary strongly in size, e.g. for $\beta = \gamma = 0$ we have $\lambda_1 = -4n^2 \sin(\pi/2n) \approx -\pi^2$ and $\lambda_{n-1} = -4n^2 \sin(\pi(n-1)/2n) \approx -4n^2$. Thus, the problem is stiff, especially for n large, and we should use a stable time stepping method.

From Theorem 6.3.17 we know that the spatial discretisation is of second order. Thus, a common time stepping method to use is the Crank-Nicolson method (cf. Definition 3.3.7), which is the second order A-stable LMM with the smallest error constant. To distinguish between spatial grid points and time steps, choose $m \in \mathbb{N}$ and let $\Delta t = T/m$ be the time step size. We denote the approximation of $y(t_j)$ at the jth time step t_j, $j = 1, \dots, m$, by $Y^{(j)} \in \mathbb{R}^{n-1}$. Applying the Crank-Nicolson method we finally obtain the fully discrete system

$$Y^{(j)} = Y^{(j-1)} - \frac{\Delta t}{2}\left(L_h Y^{(j)} + L_h Y^{(j-1)}\right) \quad \Leftrightarrow \quad \left(I + \frac{\Delta t}{2} L_h\right) Y^{(j)} = \left(I - \frac{\Delta t}{2} L_h\right) Y^{(j-1)}$$

for the jth time step. Since the real part of the spectrum of L_h is positive, the matrix on the left hand side is SPD and thus invertible, so that we can solve this system uniquely, for any $\Delta t > 0$.

We finish by stating the convergence result for this example.

Theorem 7.3.3: Consider the problem in definition 7.3.1 in one space dimension, i.e. $\Omega = (a, b) \subset \mathbb{R}$, and the differential operator $-L$ as defined in (6.4) with $\beta = \beta(x)$ and $\gamma = \gamma(x)$ independent of t. Furthermore, let $u(t, a) = u(t, b) = 0$. Then, with central finite difference discretisation of L with mesh width h and applying the Crank-Nicolson method to discretise in time, as described in Example 7.3.2 with step size $\Delta t \leq h$, there exists a constant $c > 0$ independent of h such that

$$\max_{j=0,\dots,m} \max_{k=0,\dots,n} \left| Y_k^{(j)} - u(t_j, x_k) \right| \leq ch^2.$$

7.3.2 Hyperbolic PDEs

Example 7.3.4. Finally, consider (7.7) with $Lu = -a\frac{\partial u}{\partial x}$ and $a > 0$, i.e. the **linear advection equation**

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0, \tag{7.10}$$

which models (spatially) one-dimensional transport or advection of a substance (e.g., by a fluid) from left to right with constant **advection speed** $a > 0$. (The case $a < 0$ is similar, but the advection is from right to left.)

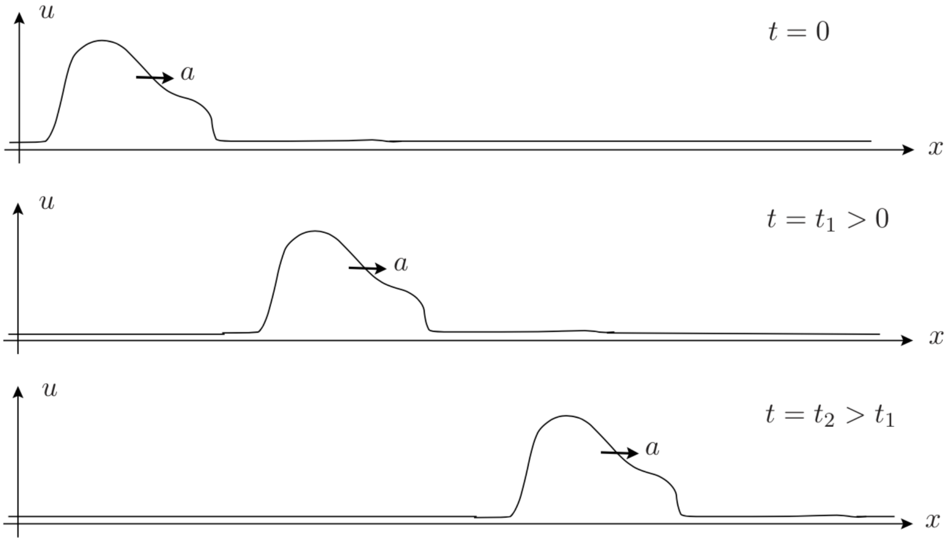


Figure 7.1: Traveling wave solution to the advection equation.

This PDE is **hyperbolic** and solutions behave very differently to those of parabolic evolution equations, such as the heat equation in Example 7.3.2 above, which approach a steady state (without further perturbations). As an important consequence, after a spatial discretisation of Lu the resulting system of ODEs is not stiff and explicit time stepping is typically sufficient.

Given an initial condition $u(0,x) = u_0(x)$, for all $x \in \mathbb{R}$ at $t = 0$, the solution of (7.10) can in fact very easily be seen (by inspection) to be

$$u(t,x) = u_0(x - at), \quad \text{for all} \;\; x \in \mathbb{R}, \;\; t \geq 0, \tag{7.11}$$

i.e. a traveling wave (cf. Fig. 7.1. The shape of the solution does not change over time.

We can easily translate this also into an initial boundary value problem, as in Definition 7.3.1. Consider (w.l.o.g. that) $\Omega = (0,1)$ and assume that the initial condition $u_0(x) = 0$ on Ω for simplicity. Then choosing the boundary condition $u(t,a) = g(t) := u_0(0 - at)$ for $t > 0$, leads again to the traveling wave solution (7.11).

Note that since the spatial operator L in (7.10) is of first order, we only need **one** boundary condition and it needs to be "upstream"; for $a < 0$ the boundary condition needs to be specified at $x = 1$. For any $t > 0$, the solution only depends on values of the solution upstream at earlier times. This gives an **analytical domain of dependence**. Thus, for any numerical approximation we have the following necessary condition found by Courant, Friedrichs and Lewy.

Definition 7.3.5 (CFL Condition): For any approximation of the solution of (7.10), the numerical domain of dependence must contain the analytical domain of dependence. In particular, given a spatial discretisation of (7.10) with mesh size h, we have the following time step restriction

$$\Delta t \leq \frac{h}{a}\,. \tag{7.12}$$

It is beyond the scope of this course to explain this condition in detail in terms of characteristics.

Let us now consider discretisation of Lu in (7.10) by a central difference and an explicit Euler time discretisation, i.e.

$$Y_k^{(j)} = Y_k^{(j-1)} - \frac{\Delta t}{2h} a \left(Y_{k+1}^{(j-1)} - Y_{k-1}^{(j-1)} \right). \tag{7.13}$$

Let $u_0(x)$ be 2π-periodic, so that it can be expanded in a Fourier series

$$u_0(x) = \sum_{\ell=-\infty}^{\infty} \beta_\ell e^{i\ell x}$$

and thus

$$Y_k^{(0)} = u_0(x_k) = \sum_{\ell=-\infty}^{\infty} \beta_\ell e^{i\ell kh}, \quad k = 0, \pm 1, \pm 2, \ldots$$

Applying (7.13) with $j = 1$ we get

$$Y_k^{(1)} = \sum_{\ell=-\infty}^{\infty} \beta_\ell e^{i\ell kh} \left(1 - \frac{a\Delta t}{2h} (e^{i\ell h} - e^{-i\ell h}) \right) = \sum_{\ell=-\infty}^{\infty} \gamma_\ell \beta_\ell e^{i\ell kh}$$

with

$$\gamma_\ell := 1 - \frac{a\Delta t}{h} i \sin(\ell h).$$

Thus, proceeding recursively, yields

$$Y_k^{(j)} = \sum_{\ell=-\infty}^{\infty} \gamma_\ell^j \beta_\ell e^{i\ell kh}.$$

The number γ_ℓ is called the **amplification coefficient** of the ℓth frequency at each time step. Since

$$|\gamma_\ell|^2 = 1 + \frac{a^2 \Delta t^2}{h^2} \sin^2(\ell h) > 1, \quad \text{if} \quad \ell h \neq m\pi, \quad m \in \mathbb{Z},$$

the nodal values $|Y_k^{(j)}|$ continue to grow as $j \to \infty$ and the numerical solution “blows up”, whereas the exact solution satisfies

$$|u(t,x)| = |u_0(x - at)| \leq \|u_0\|_\infty, \quad \text{for all} \;\; x \in \mathbb{R}, \quad t > 0.$$

Thus, the central discretisation scheme (7.13) is **unconditionally unstable**, i.e. unstable for any choice of $h > 0$ and $\Delta t > 0$.

A simple remedy is to use the upwind difference quotient again (as before at the expense of a lower consistency order), i.e. to use instead

$$Y_k^{(j)} = Y_k^{(j-1)} - \frac{\Delta t}{h} a \left(Y_k^{(j-1)} - Y_{k-1}^{(j-1)} \right) = \left(1 - \frac{\Delta t}{h} a \right) Y_k^{(j-1)} + \frac{\Delta t}{h} a Y_{k-1}^{(j-1)}. \tag{7.14}$$

Since the right hand side is just a convex combination of two neighbouring grid values, we can easily see that $\boxed{\text{DIY}}$

$$\inf_{m \in \mathbb{Z}} \{u_0(x_m)\} \leq Y_k^{(j)} \leq \sup_{m \in \mathbb{Z}} \{u_0(x_m)\}$$

and thus

$$\|Y^{(j)}\|_\infty \le \|Y^{(0)}\|_\infty\,.$$

The upwind difference scheme is stable and the time step is only constrained by the CFL condition in (7.12).

It is possible to also define stable schemes that are of higher-order in h and Δt, but we will not discuss this any further.

Appendix A

Appendix

A.1 Comments on uniqueness of an IVP

For a first order differential equation, Lipschitz continuity of f is only a sufficient and not, as one might think, a necessary condition for uniqueness of a first order differential equation. The following theorem and proof show that it is indeed possible to have uniqueness without assuming Lipschitz continuity.

Theorem A.1.1 (Non-necessity of L-continuity): Let f be a continous function satisfying $f(x) > 0$ for all $x \in \mathbb{R}$. Then, the solution to the (autonomous) IVP

$$u'(t) = f\big(u(t)\big) \tag{A.1a}$$

$$u(t_0) = u_0 \tag{A.1b}$$

is globally unique for all $(t_0, u_0) \in \mathbb{R}^2$.

Proof. Assume two solutions φ, $\psi \colon I \to \mathbb{R}$ on an open intervall I with $t_0 \in I$. Then,

$$1 = \frac{\varphi(t)'}{f(\varphi(t))} = \frac{\psi(t)'}{f(\psi(t))} \qquad \text{for all } t \in I. \tag{A.2}$$

Define the function $F \colon \mathbb{R} \to \mathbb{R}$ through

$$F(x) = \int_{u_0}^{x} \frac{\mathrm{d}s}{f(s)}.$$

F is continously differentiable since

$$\partial_x F(x) = \partial_x \left(\int_{u_0}^{x} \frac{\mathrm{d}s}{f(s)} \right) = \frac{1}{f(x)}.$$

Obviously, F is also stricly increasing, hence injective on $\mathbb{R}$: Take x, $y \in \mathbb{R}$ and assume without loss of generality that $x < y$. Then we have $F(x) < F(y)$ and thus $F(x) \neq F(y)$. Thus, F is an injection.

Also, for all $t \in I$, it follows from (A.2) that

$$F(\varphi(t)) = \int_{u_0}^{t} \frac{\varphi'(s)}{f(\varphi(s))} \,\mathrm{d}s = \int_{u_0}^{t} \frac{\psi'(s)}{f(\psi(s))} \,\mathrm{d}s = F(\psi(t)).$$

Thus, since F is injective, we have $\varphi(t) = \psi(t)$ for all $t \in I$. In conclusion, the IVP (A.1) has a unique solution. □

A.2 Properties of matrices

A.2.1 The matrix exponential

Definition A.2.1. The matrix exponential e^A of a matrix $A \in \mathbb{R}^{d\times d}$ is defined by its power series

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}. \tag{A.3}$$

Lemma A.2.2. *The power series* (A.3) *converges for each matrix A. It is therefore valid to write*

$$e^A = \lim_{m\to\infty} \sum_{k=0}^{m} \frac{A^k}{k!} = \sum_{k=0}^{\infty} \frac{A^k}{k!}. \tag{A.4}$$

Proof. Let $\|\cdot\|$ be a submultiplicative matrix norm on $\mathbb{R}^d$. We want to show that the sequence of partial sums $(S_n)_{n\in\mathbb{N}_0}$ with S_n given as $\lim_{m\to\infty} \sum_{k=n}^{m} \frac{A^k}{k!}$ converges to $S := e^A = \lim_{m\to\infty} \sum_{k=0}^{m} \frac{A^k}{k!}$. Consider therefore

$$\|S - S_n\| = \left\| \lim_{m\to\infty} \sum_{k=n+1}^{m} \frac{A^k}{k!} \right\| = \lim_{m\to\infty} \left\| \sum_{k=n+1}^{m} \frac{A^k}{k!} \right\|. \tag{A.5}$$

Using the triangle-inequality and the fact that $\|\cdot\|$ is submultiplicative yields

$$\lim_{m\to\infty} \sum_{k=n+1}^{m} \left\| \frac{A^k}{k!} \right\| \leq \lim_{m\to\infty} \sum_{k=n+1}^{m} \frac{1}{k!} \|A\|^k. \tag{A.6}$$

Considering the limit $n \to \infty$ concludes the proof. □

Lemma A.2.3 (Properties of the matrix exponential)**.** *The following relations hold true:*

$$e^0 = \mathbb{I} \tag{A.7}$$

$$e^{\alpha A} e^{\beta A} = e^{(\alpha+\beta)A}, \qquad \forall A \in \mathbb{R}^{d\times d}\ \forall \alpha, \beta \in \mathbb{R}, \tag{A.8}$$

$$e^{A} e^{-A} = \mathbb{I} \qquad \forall A \in \mathbb{R}^{d\times d}, \tag{A.9}$$

$$e^{T^{-1}AT} = T^{-1}\, e^A\, T \qquad \forall A, T \in \mathbb{R}^{d\times d} \text{ invertible}, \tag{A.10}$$

$$e^{\operatorname{diag}(\lambda_1,\ldots,\lambda_d)} = \operatorname{diag}(e^{\lambda_1}, \ldots, e^{\lambda_d}) \qquad \forall \lambda_i \in \mathbb{R},\ i = 1, \ldots, d. \tag{A.11}$$

Moreover, e^A is invertible for arbitrary quadratic matrices A with $(e^A)^{-1} = e^{-A}$.

Proof. The equality (A.7) follows directly from the definition.

For (A.8) consider the function $\varphi(\alpha)$ given by

$$\varphi(\alpha) = e^{\alpha A}e^{\beta A} - e^{(\alpha+\beta)A}.$$

Then

$$\varphi'(\alpha) = A\big(e^{\alpha A}e^{\beta A} - e^{(\alpha+\beta)A}\big) = A\varphi(\alpha) \quad \text{and} \quad \varphi(0) = \mathbb{I}e^{\beta A} - e^{\beta A} = 0,$$

giving us an IVP for $\varphi(\alpha)$ with unique solution $\varphi(\alpha) = e^{\alpha A}\varphi(0) = 0$, and the identity in (A.8) follows.

Equation (A.9) is a special case of (A.8) with parameters $\alpha = 1$ and $\beta = -1$, which in combination with (A.7) leads to the result.

For (A.10) note that $\mathbb{R}^{d\times d}$ forms a ring and is thus associative. Then, for $k \in \mathbb{N}_0$, we have

$$\begin{aligned}(T^{-1}AT)^k &= (T^{-1}AT)(T^{-1}AT)\cdots(T^{-1}AT)(T^{-1}AT)\\ &= T^{-1}A(TT^{-1})A(T\cdots T^{-1})A(TT^{-1})AT = T^{-1}A^kT\end{aligned}$$

and thus

$$e^{T^{-1}AT} = \sum_{k=0}^{\infty}\frac{1}{k!}(T^{-1}AT)^k = \sum_{k=0}^{\infty}\frac{1}{k!}T^{-1}A^kT = \ T^{-1}\cdot\left(\sum_{k=0}^{\infty}\frac{1}{k!}A^k\right)\cdot T \ = \ T^{-1}e^AT.$$

To prove (A.11), let $D = \operatorname{diag}(\lambda_1,\dots,\lambda_d) \in \mathbb{R}^{d\times d}$ where $\lambda_i \in \mathbb{R}$, $i = 1,\dots,d$. Then, $D^k = \operatorname{diag}(\lambda_1^k,\dots,\lambda_n^k)$, for any $k \in \mathbb{N}_0$, and we have

$$\begin{aligned}e^D &= \lim_{m\to\infty}\sum_{k=0}^{m}\frac{1}{k!}\operatorname{diag}(\lambda_1^k,\dots,\lambda_n^k) && \text{(A.12)}\\ &= \lim_{m\to\infty}\sum_{k=0}^{m}\operatorname{diag}\left(\frac{1}{k!}\lambda_1^k,\dots,\frac{1}{k!}\lambda_n^k\right) && \text{(A.13)}\\ &= \lim_{m\to\infty}\operatorname{diag}\left(\sum_{k=0}^{m}\frac{1}{k!}\lambda_1^k,\dots,\sum_{k=0}^{m}\frac{1}{k!}\lambda_n^k\right) && \text{(A.14)}\\ &= \operatorname{diag}\left(\lim_{m\to\infty}\sum_{k=0}^{m}\frac{1}{k!}\lambda_1^k,\dots,\lim_{m\to\infty}\sum_{k=0}^{m}\frac{1}{k!}\lambda_n^k\right) && \text{(A.15)}\\ &= \operatorname{diag}(e^{\lambda_1^k},\dots,e^{\lambda_n^k}) && \text{(A.16)}\end{aligned}$$

Here, we have used the absolute convergence of the series and that these matrices are elements of the ring $R^{d\times d}$.

The final property follows immediately from (A.9). □

Example A.2.4. We will perform an exemplary calculation of a matrix exponential. Consider

$$A = \begin{pmatrix}0 & 1\\ k^2 & 0\end{pmatrix}.$$

As the matrix exponential of a diagonal matrix is simply a diagonal matrix with the exponential of the entries, we diagonalize A.

To diagonalize A, note that the eigenvalues λ_1, λ_2 of A are $\lambda_1 = k$ and $\lambda_2 = -k$. Let $D = \operatorname{diag}(\lambda_1, \lambda_2) = \operatorname{diag}(k, -k)$. The corresponding eigenvectors are $\psi_1 = \big(1, k\big)^T$ and $\psi_2 = \big(1, -k\big)$. The matrix $\Psi = (\psi_1|\psi_2) \in \mathbb{R}^{2\times 2}$ satisfies

$$A = \Psi^{-1} D \Psi.$$

The inverse of Ψ is given as

$$\Psi^{-1} = \frac{1}{2}\begin{pmatrix} 1 & {}^1\!/\!_k \\ 1 & -{}^1\!/\!_k \end{pmatrix}$$

and with the above lemma we can now calculate

$$e^A = \Psi e^D \Psi^{-1} = \frac{1}{2}\begin{pmatrix} e^k + e^{-k} & {}^1\!/\!_k(e^k - e^{-k}) \\ k(e^k - e^{-k}) & e^k + e^{-k} \end{pmatrix} = \begin{pmatrix} \cosh(k) & {}^1\!/\!_k \sinh(k) \\ k\sinh(k) & \cosh(k) \end{pmatrix}.$$

A.3 The Banach fixed-point theorem

Theorem A.3.1 (Banach fixed-point theorem): Let $\Omega \subset \mathbb{R}$ be a closed set and $f\colon \Omega \to \Omega$ a contraction, i.e. there exists $\gamma \in (0,1)$ such that $|f(x) - f(y)| \leq \gamma |x - y|$. Then, there exists a unique $x^* \in \Omega$ such that $f(x^*) = x^*$.

Proof. Let $x_0 \in \Omega$ and define $x_{k+1} = f(x_k)$. First, we prove existence using the Cauchy-criterion. Let $k, n \in \mathbb{N}_0$ and consider

$$|x_k - x_{k+m}| = |f(x_{k-1}) - f(x_{k+m-1})| \leq \gamma |x_{k-1} - x_{k+m-1})|.$$

Iteratively, we get

$$|x_k - x_{k+m}| \leq \gamma^k |x_0 - x_m|.$$

We now write $x_0 - x_m = x_0 - x_1 + x_1 - x_2 + \cdots + x_{m-1} - x_m$. The triangle-inequality then yields the estimate

$$\begin{aligned} |x_k - x_{k+m}| &\leq \gamma^k \left(|x_0 - x_1| + |x_1 - x_2| + \cdots + |x_{m-1} - x_m|\right) \\ &\leq \gamma^k |x_0 - x_1| \left(1 + \gamma + \gamma^2 + \cdots + \gamma^{m-1}\right) \leq \frac{\gamma^k}{1-\gamma}|x_0 - x_1|. \end{aligned}$$

As k gets larger this estimate goes to zero.

Concerning uniqueness, let x^* and y^* be fixpoints. Then,

$$|x^* - y^*| = |f(x*) - f(y^*)| \leq \gamma |x^* - y^*|$$

Since $\gamma \in (0,1)$ we immediately obtain $|x^* - y^*| = 0$. Using that $|a| = 0$ if and only if $a = 0$ yields $y^* = x^*$. This concludes the proof. □

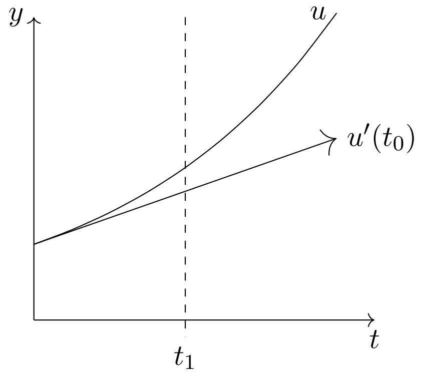
A.4 The implicit and explicit Euler-method

The explicit resp. implicit Euler is given by the one-step method

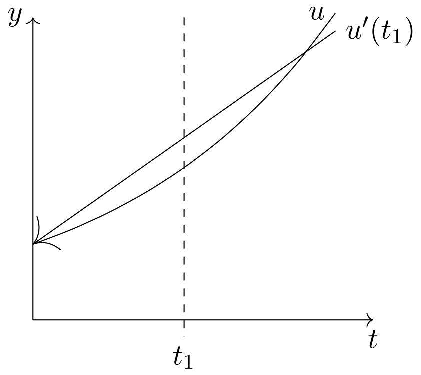
$$y_1 = y_0 + hf(y_0) \qquad \text{resp.} \qquad y_1 = y_0 + hf(y_1)$$

Clearly, the explicit Euler is a rather easy calculation since all one needs are f, h and y_0. The implicit Euler is more difficult to compute since for calculating y_1 we need the value of f at y_1. The goal of this section is to visualize and give an intuition for the two algorithms.

Consider the following visualizations.



For the explicit Euler we take u_0 and u_0'. y_1, our approximated solution for u_1, is chosen as the intersection point of t_1 and $g(t) = y_0 + t \cdot u'(t_0)$.

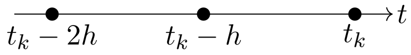


For implicit Euler we go backwards. On the t_1-axis we are looking for an the affine function g that fulfills $g(0) = u_0$ and $g'(t_1) = f(t_1)$. Then we set $y_1 = g(t_1)$.

A.5 Derivation of a BDF-scheme

The BDF formulae use the approximations of the solution at the previous time steps $t_k - sh, \dots, t_k - h$ and the unkown value y_k at t_k that we would like to determine. With the Lagrange polynomial given by $L_i(t) = \prod_{j=0, j\neq i}^{s} \frac{t-t_i}{t_j-t_i}$ we let $y(t) = \sum_{j=0}^{s} y_{k-j} L_{s-j}(t)$. Then, we will assume that y solves the IVP in the point t_k and obtain a linear system from which we derive the desired value y_k.

We now aim to derive the scheme for BDF(2): Let the points $t_k - 2h$, $t_k - h$ and t_k be given.



For the corresponding Lagrange polynomials we have, resp.,

$$L_0(t) = \tfrac{(t-t_k+h)(t-t_k)}{2h^2}, \; L_1(t) = \tfrac{(t-t_k)(t-t_kj-2h)}{h^2} \; \text{ and } \; L_2(t) = \tfrac{(t-t_k+2h)(t-t_k+h)}{2h^2}.$$

By assumption the interpolation polynomial fulfilles the IVP in the point t_k, i.e. there holds $f_k := f(t_k, y(t_k)) = y'(t_k) = \sum_{j=1}^{s} y_{k-j} L_{k-j}'(t)$. Since

$$L_0'(t) = \tfrac{2t-2t_k+h}{2h^2}\,,\; L_1'(t) = -\tfrac{2t-2t_k+2h}{h^2} \quad\text{and}\quad L_2'(t) = \tfrac{2t-2t_k+3h}{2h^2}\,,$$

evaluation at $t = t_k$ yields

$$f_k = \tfrac{1}{2h} y_{k-2} - \tfrac{2}{h} y_{k-1} + \tfrac{3}{2h} y_k\,.$$

The final BDF(2)-scheme is obtained by multiplication with $\frac{2h}{3}$:

$$y_k - \tfrac{4}{3} y_{k-1} + \tfrac{1}{3} y_{k-2} = \tfrac{2}{3h} f_k.$$

Bibliography

[But96] J. C. Butcher. A history of Runge-Kutta methods. *Appl. Numer. Math.*, 20(3):247–260, 1996.

[DB08] P. Deuflhard and F. Bornemann. *Numerische Mathematik 2. Gewöhnliche Differentialgleichungen.* de Gruyter, 3. auflage edition, 2008.

[Heu86] H. Heuser. *Lehrbuch der Analysis. Teil 2.* Teubner, 3. auflage edition, 1986.

[HNW09] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations I. Nonstiff problems*, volume 8 of *Springer Series in Computational Mathematics.* Springer, Berlin, second edition edition, 2009.

[HS52] M. R. Hestenes and E. Stiefel. Method of conjugate gradient for solving linear equations. *J. Res. Natl. Bur. Stand.*, 49:409–436, 1952.

[HW10] E. Hairer and G. Wanner. *Solving ordinary differential equations II. Stiff and differential-algebraic problems*, volume 14 of *Springer Series in Computational Mathematics.* Springer-Verlag, Berlin, second edition edition, 2010.

[LR05] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*, volume 14 of *Cambridge Monographs on Applied and Computational Mathematics.* Cambridge University Press, Cambridge, 2005.

[Lub] Ch. Lubich. Chapter VI. Symplectic Integration of Hamiltonian Systems. Lecture Notes, Universität Tübingen. `https://na.uni-tuebingen.de/~lubich/chap6.pdf`.

[NW06] J. Nocedal and S. J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[Ran17a] R. Rannacher. *Numerik 0: Einführung in die Numerische Mathematik.* Heidelberg University Publishing, 2017. DOI: 10.17885/heiup.206.281.

[Ran17b] R. Rannacher. *Numerik 1: Numerik gewöhnlicher Differentialgleichungen.* Heidelberg University Publishing, 2017. DOI: 10.17885/heiup.258.342.

[Run95] C. Runge. Über die numerische Auflösung von Differentialgleichungen. *Math. Ann.*, 46:167–178, 1895.