Numerics of differential equations

Michael Innerberger, Lothar Nannen, Dirk Praetorius

Note: These lecture notes provide the way in which I would have liked to present the course material. If you face any typos, please let me know:
`lothar.nannen@tuwien.ac.at`.

CHAPTER 1

Introduction

Differential equations can be used to model chemical, mechanical, biological, economical, or physical processes. We call an equation a differential equation, if in this equation (partial) derivatives of a sought function appear. In this lecture, we mainly focus on ordinary differential equations, i.e. only derivatives in one direction are involved. We will call this direction in most cases t for time. Note, that we will treat a single equation and a system of differential equations at the same time.

The topics of this lecture are numerical methods for solving ordinary differential equations. Nevertheless, we will give in the introduction some basic informations to the existence and uniqueness of solutions to boundary value problems. Moreover, we will state at least one stability theorem showing, that the problems under consideration are to some extent well posed.

THEOREM 1.1 (Picard-Lindelöf). *Let $n \in \mathbb{N}$, $a, b > 0$, $t_0 \in \mathbb{R}$, $y_0 \in \mathbb{R}^n$, $R := \{(t,y) \in \mathbb{R} \times \mathbb{R}^n : |t - t_0| < a, \|y - y_0\| \le b\}$ and $f \in C(R; \mathbb{R}^n)$ Lipschitz continuous in y, i.e. there exists a fixed constant $L > 0$ such that*

$$\forall\, (t,y), (t,\widetilde{y}) \in R : \quad \|f(t,y) - f(t,\widetilde{y})\| \le L\, \|y - \widetilde{y}\|. \tag{1.1}$$

We define $M := \sup_{(t,y)\in R} \|f(t,y)\|$, $\alpha < \min\{a, b/M, 1/L\}$, and $J := [t_0 - \alpha, t_0 + \alpha]$.

Then, there exists a solution $y \in C^1(J, \mathbb{R}^n)$ of the initial value problem

$$\forall t \in J : \, y'(t) = f(t, y(t)), \qquad y(t_0) = y_0. \tag{1.2}$$

This solution is unique in the following sense: For a different solution $\tilde{y} \in C^1(\tilde{J}, \mathbb{R}^n)$ of (1.2) there holds $y = \tilde{y}$ for all $t \in J \cap \tilde{J}$.

PROOF. (1.2) is equivalent to the fixed point problem $y = \Phi y$ with the integral operator Φ defined by

$$(\Phi y)\,(t) := y_0 + \int_{t_0}^{t} f(\tau, y(\tau)) d\tau, \qquad t \in (t_0 - a, t_0 + a). \tag{1.3}$$

We define

$$U := \{y \in C(J, \mathbb{R}^n) : y(t_0) = y_0 \wedge \|y(t) - y_0\| \le b\, \forall t \in J\}.$$

Equipped with the supremum norm $\|y\|_\infty = \max_{t\in J} \|y(t)\|$, the space U is a Banach space and $\Phi : U \to U$, since Φy is continuous for continuous y, $(\Phi y)(t_0) = y_0$, and

$$\forall t \in J : \quad \|(\Phi y)(t) - y_0\| \le \left\| \int_{t_0}^{t} f(\tau, y(\tau)) d\tau \right\| \le \alpha M \le b.$$

Moreover, Φ is a contraction, since for all $t \in J$ there holds

$$\|(\Phi y)\,(t) - (\Phi \widetilde{y})\,(t)\| \le \left| \int_{t_0}^{t} \|f(\tau, y(\tau)) - f(\tau, \widetilde{y}(\tau))\|\, d\tau \right| \le L\alpha \|y - \widetilde{y}\|.$$

Hence, $\|\Phi y - \Phi \tilde{y}\|_\infty < q\|y - \tilde{y}\|_\infty$ with $q < 1$ and the claims follows with the fixed point theorem of Banach. □

A problem is well posed in the sense of Hadamard, if there exists a unique solution and if the solution depends continuously on the initial data. The latter is very important, since we have to expect errors in the initial data. We want to bound the resulting errors in the solutions with respect to these errors. If the problem is well defined, small errors in the initial data will lead to small errors in the solution.

LEMMA 1.2 (Gronwall). *Let $J \in \mathbb{R}$ be an interval and $t_0 \in J$. Moreover, $A, B \geq 0$, $L > 0$ and the non-negative function $v \in C(J, \mathbb{R})$ satisfies the inequality*

$$v(t) \leq A + B|t - t_0| + L \left| \int_{t_0}^{t} v(\tau) d\tau \right|, \qquad t \in J.$$

Then

$$v(t) \leq A e^{L|t-t_0|} + \frac{B}{L} \left(e^{L|t-t_0|} - 1 \right), \qquad t \in J.$$

PROOF. We prove the assertion only for $t > t_0 = 0$. The general case can be deduced in the same way. Let

$$y(t) := Bt + L \int_0^t v(\tau) d\tau, \qquad t \in J.$$

Then

$$y'(t) = B + Lv(t) \leq B + L \left(A + Bt + L \int_0^t v(\tau) d\tau \right) = B + L \left(A + y(t) \right).$$

Hence, $y'(t) - Ly(t) \leq B + LA$ and a multiplication with e^{-Lt} leads to

$$\left(e^{-Lt} y(t) \right)' = e^{-Lt} \left(y'(t) - Ly(t) \right) \leq e^{-Lt}(LA + B).$$

Since $y(0) = 0$, integration over $(0, t)$ leads to

$$e^{-Lt} y(t) = \frac{LA + B}{-L} (e^{-Lt} - 1) = \frac{LA + B}{L} (1 - e^{-Lt}).$$

Finally, a multiplication with e^{Lt} gives the assertion

$$v(t) \leq A + Bt + L \int_0^t v(\tau) d\tau = A + y(t) \leq A e^{Lt} + \frac{B}{L} \left(e^{Lt} - 1 \right).$$

□

THEOREM 1.3. *Let J be an open interval, $\Omega \subset \mathbb{R}^n$ be an open domain and $f \in C(J \times \Omega, \mathbb{R}^n)$ Lipschitz continuous in y (see (1.1)). Moreover, let $y \in C^1(J, \mathbb{R}^n)$ be a solution to the differential equation $y'(t) = f(t, y(t))$ and $z \in C^1(J, \mathbb{R}^n)$ an approximate solution to the differential equation, i.e. there exists $\delta > 0$ such that*

$$\forall t \in J: \quad \|z'(t) - f(t, z(t))\| \leq \delta.$$

Then

$$\forall t \in J: \quad \|y(t) - z(t)\| \leq \|y(t_0) - z(t_0)\| e^{L|t-t_0|} + \frac{\delta}{L} \left(e^{L|t-t_0|} - 1 \right). \tag{1.4}$$

In other words, the error at each time t increases at most exponentially with respect to the error in the initial value and the error in the differential equation.

PROOF. The results is a consequence of the Gronwall Lemma for $v(t) := \|y(t)-z(t)\|$, $A := v(t_0)$, $B := \delta$ and

$$\begin{aligned} v(t) = \|y(t) - z(t)\| &\leq \|y(t_0) - z(t_0)\| + \left\| \int_{t_0}^{t} (y'(\tau) - z'(\tau))\, d\tau \right\| \\ &\leq A + \left\| \int_{t_0}^{t} (f(\tau, y(\tau)) - f(\tau, z(\tau)))\, d\tau \right\| + \left\| \int_{t_0}^{t} (z'(\tau) - f(\tau, z(\tau)))\, d\tau \right\| \\ &\leq A + L \left| \int_{t_0}^{t} \underbrace{\|y(\tau) - z(\tau)\|}_{=v(\tau)}\, d\tau \right| + \delta |t - t_0|. \end{aligned}$$

□

CHAPTER 2

Explicit one-step methods

Throughout this section, we consider the following model problem: Let $[t_0, T]$ be a given time-interval. For given $n \in \mathbb{N}$, let $f \in C([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ and suppose that f is Lipschitz continuous in y, i.e.,

$$\forall\, t \in [t_0, T]\, \forall\, y, \widetilde{y} \in \mathbb{R}^n : \quad \|f(t, y) - f(t, \widetilde{y})\| \leq L\, \|y - \widetilde{y}\|, \tag{2.1}$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$ and $L > 0$ is a fixed constant. For any initial value $y_0 \in \mathbb{R}^n$, we are looking for solutions $y \in C^1([t_0, T]; \mathbb{R}^n)$ of

$$y(t_0) = y_0 \quad \text{and} \quad y'(t) = f(t, y(t)) \quad \text{for all } t \in [t_0, T]. \tag{2.2}$$

Remark 2.1. If $f(t, y) = g(t)$, then the exact solution of (2.2) reads

$$y(t) = y_0 + \int_{t_0}^{t} g(\tau)\, \mathrm{d}\tau,$$

i.e., y is the antiderivative of g. Therefore, the solution of an initial value problem (2.2) ist also called **integration of the ODE** and the numerical solvers are also called **(numerical) integrators**.

2.1. Notation

In practice, the solution of the initial boundary value problem (2.2) cannot be computed in closed form. If quantitative results are required, one usually considers numerical methods, which provide approximations $y_\ell \approx y(t_\ell)$ at certain time-steps $t_\ell \in [t_0, T]$. From now on, we shall implicitly use the following notation:

Definition 2.2. Suppose that we have given time-steps $t_0 < t_1 < \cdots < t_N = T$. The set $\Delta := \{t_0, \ldots, t_N\}$ is called **mesh** (or **grid**) of the time interval $[t_0, T]$. The quantities $h_\ell := t_{\ell+1} - t_\ell$ are called **step-sizes**. Moreover, the **maximum step-size** of Δ is defined as $h_\Delta := \max\limits_{\ell=0,\ldots,N-1} h_\ell$.

Definition 2.3 (One-step method). Given an **incremental function** $\Phi : [t_0, T] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}^n$, the inductive procedure

$$y_{\ell+1} := y_\ell + h_\ell\, \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \quad \text{for all } \ell = 0, \ldots, N-1 \tag{2.3}$$

is called a **one-step method**. If Φ does not depend explicitly on $y_{\ell+1}$, we call the method an **explicit one-step method** and omit the third argument. Otherwise, the method is called **implicit one-step method**.

Given the value $y_\ell \approx y(t_\ell)$, we compute $y_{\ell+1} \approx y(t_{\ell+1})$. Note that this computation involves only the last time-step, but *not* the full history $y_0, \ldots, y_{\ell-1}$. This will be different for *multi-step methods* considered later.

Remark 2.4. Each time-step of the implicit one-step method will require the solution of one (possibly nonlinear) equation. It is not a priori clear, that such a solution exists. Moreover, usually the solution of a nonlinear equation is more costly than the direct evaluation needed for the explicit method. Nevertheless, we will see later on, that implicit methods are for some problems better suited than explicit ones.

EXAMPLE 2.5 (Explicit Euler method / forward Euler method). If $f \in C^1([t_0,T] \times \mathbb{R}^n;\mathbb{R}^n)$ and hence $y \in C^2([t_0,T];\mathbb{R}^n)$, the Taylor theorem proves that

$$y(t+h) = y(t) + h\,y'(t) + \mathcal{O}(h^2) = y(t) + h\,f\bigl(t, y(t)\bigr) + \mathcal{O}(h^2). \tag{2.4}$$

With $y(t_{\ell+1}) \approx y_{\ell+1}$ and $y(t_\ell) \approx y_\ell$, the explicit Euler method reads

$$y_{\ell+1} = y_\ell + h_\ell\,\Phi(t_\ell, y_\ell, h_\ell) \quad \text{with} \quad \Phi(t_\ell, y_\ell, h_\ell) := f\bigl(t_\ell, y_\ell\bigr). \tag{2.5}$$

EXAMPLE 2.6 (Implicit Euler method / backward Euler method). If $f \in C^1([t_0,T] \times \mathbb{R}^n;\mathbb{R}^n)$ and hence $y \in C^2([t_0,T];\mathbb{R}^n)$, the Taylor theorem proves that

$$y(t) = y(t+h) - h\,y'(t+h) + \mathcal{O}(h^2) = y(t+h) - h\,f\bigl(t+h, y(t+h)\bigr) + \mathcal{O}(h^2). \tag{2.6}$$

With $y(t_{\ell+1}) \approx y_{\ell+1}$ and $y(t_\ell) \approx y_\ell$, the implicit Euler method reads

$$y_{\ell+1} = y_\ell + h_\ell\,\Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \quad \text{with} \quad \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) := f\bigl(t_{\ell+1}, y_{\ell+1}\bigr), \tag{2.7}$$

where we note that $t_{\ell+1} = t_\ell + h_\ell$.

2.2. Consistency

DEFINITION 2.7. Let $\Phi(t, y, h)$ be the incremental function of an explicit one-step method and y be the exact solution of (2.2) with initial value $y_0 = y(t)$ at the time step t. We call the error

$$\tau(t, y, h) := \|y(t+h) - \bigl[y(t) + h\,\Phi(t, y(t), h)\bigr]\|, \tag{2.8}$$

i.e. the error if one step of the method is applied to the exact solution, the **consistency error**. We say that the one-step method is **consistent**, if

$$\forall\, t \in [t_0, T): \quad \lim_{h \to 0^+} \frac{\tau(t, y, h)}{h} = 0, \tag{2.9}$$

i.e. the discretization error of *one step* vanishes as $h \to 0$. For $p \geq 1$, we say that the one-step method has **consistency order** p, if for all $f \in C^p([t_0,T] \times \mathbb{R}^n;\mathbb{R}^n)$ and hence $y \in C^{p+1}([t_0,T];\mathbb{R}^n)$, it holds that

$$\exists\, C > 0: \ \forall\, t \in [t_0, T)\, \forall\, h \in (0, T-t]: \quad \tau(t, y, h) \leq C\,h^{p+1}. \tag{2.10}$$

REMARK 2.8. Suppose that the incremental function $\Phi(t, y(t), h)$ is continuous at $h = 0$. Then,

$$\begin{aligned} \text{consistency (2.9)} \quad &\iff \quad \forall\, t \in [t_0, T): \quad \lim_{h \to 0^+} \Bigl\| \frac{y(t+h) - y(t)}{h} - \Phi(t, y(t), h) \Bigr\| = 0 \\ &\iff \quad \forall\, t \in [t_0, T): \quad f(t, y(t)) = \Phi(t, y(t), 0), \end{aligned}$$

since $f(t, y(t)) = y'(t) = \lim\limits_{h \to 0^+} \dfrac{y(t+h) - y(t)}{h}$.

Leonhard Euler (1707–1783) *was a Swiss mathematician. From 1720–1723, he was studying at the University of Basel (attending also classes by Johann I Bernoulli). In 1727, he moved to the Imperial Russian Academy of Sciences in Saint Petersburg, where he became professor of mathematics in 1733. In 1741, he moved to the Prussian Academy of Sciences in Berlin, but returned to the Imperial Russian Academy of Sciences in Saint Petersburg in 1766.*

EXAMPLE 2.9 (Explicit Euler has consistency order $\boldsymbol{p} = \mathbf{1}$). The explicit Euler method (2.5) is always consistent, since $\Phi(t, y, h) = f(t, y)$. According to (2.4) it has consistency order $p = 1$. Moreover, the constant C in the consistency estimate (2.10) takes the form $C = \|y''\|_{\infty,[t_0,T]}/2$, since the Taylor expansion yields that

$$y(t+h) = y(t) + h\,y'(t) + \frac{h^2}{2}\,y''(\xi) \overset{(2.5)}{=} y(t) + h\,\Phi(t, y(t), h) + \frac{h^2}{2}\,y''(\xi)$$

for some appropriate $\xi \in [t, t+h]$.

REMARK 2.10. If the mesh Δ has step-size h, then it takes $N \sim \frac{T-t_0}{h} = \mathcal{O}(h^{-1})$ steps of the numerical scheme to obtain the approximation $y_N \approx y(T)$. If each step has a (cumulative consistency) error of order $\mathcal{O}(h^{p+1})$, one can hope for a total error estimate

$$\|y(T) - y_N\| \sim \sum_{j=0}^{N-1} \mathcal{O}(h^{p+1}) = \mathcal{O}(Nh^{p+1}) = \mathcal{O}(h^p).$$

In the following section, we aim to rigorously prove this expectation.

2.3. Convergence

LEMMA 2.11 (Discrete Gronwall lemma). *Let $A > 0$, $B \geq 0$, $h_\ell, a_\ell \geq 0$ such that*

$$0 \leq a_{\ell+1} \leq (1 + h_\ell A)a_\ell + h_\ell B \quad \textit{for all } \ell \in \mathbb{N}_0. \tag{2.11}$$

Then, it holds that

$$0 \leq a_\ell \leq \frac{B}{A}\left(\exp\Big(A\sum_{j=0}^{\ell-1} h_j\Big) - 1\right) + a_0 \exp\Big(A\sum_{j=0}^{\ell-1} h_j\Big) \quad \textit{for all } \ell \in \mathbb{N}_0. \tag{2.12}$$

PROOF. It holds that

$$1 + x \leq \sum_{j=0}^{\infty} \frac{x^j}{j!} = \exp(x) \quad \text{for all } x \geq 0. \tag{2.13}$$

Note that (2.12) holds for $\ell = 0$ (even with equality $a_\ell = a_0$). Arguing by induction on ℓ, we may suppose that (2.12) holds up to some $\ell \in \mathbb{N}_0$. For $\ell + 1$, note that

$$\begin{aligned}
a_{\ell+1} &\overset{(2.11)}{\leq} (1 + h_\ell A)a_\ell + h_\ell B \\
&\overset{(2.12)}{\leq} (1 + h_\ell A)\left[\frac{B}{A}\left(\exp\Big(A\sum_{j=0}^{\ell-1} h_j\Big) - 1\right) + a_0 \exp\Big(A\sum_{j=0}^{\ell-1} h_j\Big)\right] + h_\ell B \\
&\leq \frac{B}{A}\left(\underbrace{(1 + h_\ell A)}_{\overset{(2.13)}{\leq} \exp(h_\ell A)} \exp\Big(A\sum_{j=0}^{\ell-1} h_j\Big) - (1 + h_\ell A)\right) + a_0 \underbrace{(1 + h_\ell A)}_{\overset{(2.13)}{\leq} \exp(h_\ell A)} \exp\Big(A\sum_{j=0}^{\ell-1} h_j\Big) + h_\ell B \\
&\leq \frac{B}{A}\left(\exp\Big(A\sum_{j=0}^{\ell} h_j\Big) - (1 + h_\ell A)\right) + a_0 \exp\Big(A\sum_{j=0}^{\ell} h_j\Big) + h_\ell B \\
&= \frac{B}{A}\left(\exp\Big(A\sum_{j=0}^{\ell} h_j\Big) - 1\right) + a_0 \exp\Big(A\sum_{j=0}^{\ell} h_j\Big).
\end{aligned}$$

This concludes the proof. □

THEOREM 2.12 (Consistency plus stability implies convergence). *Let $\Phi(t, y, h)$ be the incremental function of an explicit one-step method. Suppose stability*

$$\forall\, t \in [t_0, T)\,\forall\, h \in (0, T-t]\,\forall\, y, \widetilde{y} \in \mathbb{R}^n: \quad \|\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)\| \le C_{\rm stab}\, \|y - \widetilde{y}\| \tag{2.14}$$

for some constant $C_{\rm stab} > 0$. Let $f \in C^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ with corresponding solution $y \in C^{p+1}([t_0, T]; \mathbb{R}^n)$. Then, consistency order $p \ge 1$ implies that

$$\max_{\ell=1,\dots,N} \|y_\ell - y(t_\ell)\| \le \frac{C_{\rm cons}}{C_{\rm stab}} \big[\exp\big(C_{\rm stab}(T - t_0)\big) - 1\big]\, h_\Delta^p, \tag{2.15}$$

where $C_{\rm cons} > 0$ is the consistency constant; see (2.10) *in Definition 2.7.*

PROOF. We define the consistency error

$$\eta(t_{\ell+1}) := y(t_{\ell+1}) - \big[y(t_\ell) + h_\ell\, \Phi(t_\ell, y(t_\ell), h_\ell)\big] \overset{(2.10)}{=} \mathcal{O}(h_\ell^{p+1}).$$

With this notation, the error satisfies

$$\begin{aligned} E_{\ell+1} &:= y_{\ell+1} - y(t_{\ell+1}) \\ &= \big[y_\ell + h_\ell\, \Phi(t_\ell, y_\ell, h_\ell)\big] - \big[y(t_\ell) + h_\ell \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}) \\ &= \big[y_\ell - y(t_\ell)\big] + h_\ell\,\big[\Phi(t_\ell, y_\ell, h_\ell) - \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}). \end{aligned}$$

With $a_\ell := \|y_\ell - y(t_\ell)\|$, stability (2.14) and consistency (2.10) lead to

$$0 \le a_{\ell+1} \le a_\ell\,(1 + h_\ell C_{\rm stab}) + C_{\rm cons}\, h_\ell^{p+1} \le a_\ell\,(1 + h_\ell C_{\rm stab}) + C_{\rm cons}\, h_\Delta^p\, h_\ell.$$

With $a_0 = 0$ and Lemma 2.11, we infer that

$$\|y_\ell - y(t_\ell)\| = a_\ell \overset{(2.12)}{\le} \frac{C_{\rm cons}}{C_{\rm stab}} \left(\exp\Big(C_{\rm stab} \sum_{j=0}^{\ell-1} h_j\Big) - 1\right) h_\Delta^p \quad \text{for all } \ell \in \mathbb{N}_0.$$

The claim follows from $\sum_{j=0}^{\ell-1} h_j \le T - t_0$. □

REMARK 2.13 (Stability of explicit Euler method). For the explicit Euler method, it holds that $\Phi(t, y, h) = f(t, y)$. Hence, stability (2.14) with $C_{\rm stab} = L$ follows from Lipschitz continuity of f in y.

COROLLARY 2.14 (One step methods are stable). *Under the assumptions of Theorem 2.12, we suppose that the* computed approximations *face rounding errors such that*

$$\widetilde{y}_0 = y_0 + \varepsilon_0 \quad \textit{and} \quad \widetilde{y}_{\ell+1} = \widetilde{y}_\ell + h_\ell\, \Phi(t_\ell, \widetilde{y}_\ell, h_\ell) + \delta_\ell \quad \textit{for all } \ell = 0, \dots, N-1, \tag{2.16}$$

where $\|\varepsilon_0\| \le \varepsilon$ and $\|\delta_\ell\| \le \delta\, h_\ell$. Then,

$$\max_{\ell=1,\dots,N} \|\widetilde{y}_\ell - y(t_\ell)\| \le C\,(h_\Delta^p + \delta) + \varepsilon\, \exp\big(C_{\rm stab}(T - t_0)\big), \tag{2.17}$$

where $C = \dfrac{\max\{1, C_{\rm cons}\}}{C_{\rm stab}} \big[\exp\big(C_{\rm stab}(T - t_0)\big) - 1\big]$.

PROOF. We argue as for the proof of Theorem 2.12. Define the consistency error

$$\eta(t_{\ell+1}) := y(t_{\ell+1}) - \big[y(t_\ell) + h_\ell\, \Phi(t_\ell, y(t_\ell), h_\ell)\big] = \mathcal{O}(h_\ell^{p+1}).$$

With this notation, the perturbed error satisfies

$$\begin{aligned} \widetilde{E}_{\ell+1} &:= \widetilde{y}_{\ell+1} - y(t_{\ell+1}) \\ &= \big[\widetilde{y}_\ell + h_\ell\, \Phi(t_\ell, \widetilde{y}_\ell, h_\ell) + \delta_\ell\big] - \big[y(t_\ell) + h_\ell \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}) \\ &= \big[\widetilde{y}_\ell - y(t_\ell)\big] + h_\ell\,\big[\Phi(t_\ell, \widetilde{y}_\ell, h_\ell) - \Phi(t_\ell, y(t_\ell), h_\ell)\big] - \eta(t_{\ell+1}) + \delta_\ell \end{aligned}$$

With $a_\ell := \|\widetilde{y}_\ell - y(t_\ell)\|$, stability (2.14) and consistency (2.10) lead to

$$\begin{aligned} 0 \le a_{\ell+1} &\le a_\ell\,(1 + h_\ell C_{\mathrm{stab}}) + C_{\mathrm{cons}}\, h_\ell^{p+1} + \delta_\ell \\ &\le a_\ell\,(1 + h_\ell C_{\mathrm{stab}}) + \max\{1, C_{\mathrm{cons}}\}\,(h_\Delta^p + \delta)\, h_\ell \end{aligned}$$

With $a_0 = \|\widetilde{y}_0 - y_0\| \le \varepsilon$ and Lemma 2.11, we infer that

$$\begin{aligned} \|\widetilde{y}_\ell - y(t_\ell)\| = a_\ell \overset{(2.12)}{\le} (h_\Delta^p + \delta)\,\frac{\max\{1, C_{\mathrm{cons}}\}}{C_{\mathrm{stab}}} &\Big(\exp\Big(C_{\mathrm{stab}} \sum_{j=0}^{\ell-1} h_j\Big) - 1\Big) \\ &+ \varepsilon \exp\Big(C_{\mathrm{stab}} \sum_{j=0}^{\ell-1} h_j\Big)\Big) \quad \text{for all } \ell \in \mathbb{N}_0. \end{aligned}$$

With $\sum_{j=0}^{\ell-1} h_j \le T - t_0$, this concludes the proof. □

2.4. Examples

We aim to construct explicit one-step methods with higher convergence order. One natural approach is to start from the Taylor expansion of the exact solution

$$y(t+h) = y(t) + \sum_{k=1}^{p} \frac{y^{(k)}(t)}{k!}\, h^k + \mathcal{O}(h^{p+1}), \tag{2.18}$$

where we implicitly assume smoothness of $f \in C^p([t_0,T] \times \mathbb{R}^n; \mathbb{R}^n)$ and hence of $y \in C^{p+1}([t_0,T]; \mathbb{R}^n)$. Since y is unknown (and so is $y^{(k)}$), we need to express the derivatives in terms of the given right-hand side f. First, recall that

$$y'(t) = f\big(t, y(t)\big). \tag{2.19}$$

To write the second derivative of f, define $g(t) := \big(t, y(t)\big)$. Then, the chain rule gives

$$\begin{aligned} y''(t) = \mathrm{d}_t f\big(t, y(t)\big) = \mathrm{d}_t\big[f\big(g(t)\big)\big] &= Df\big(t, y(t)\big) Dg(t) \\ &= (\partial_t f)\big(t, y(t)\big) + (\partial_y f)\big(t, y(t)\big)\, y'(t) \\ &= (\partial_t f)\big(t, y(t)\big) + (\partial_y f)\big(t, y(t)\big)\, f\big(t, y(t)\big). \end{aligned} \tag{2.20}$$

Proceeding with the chain rule, we can express all derivatives of y in terms of partial derivatives of f. We leave this to the interested reader, but state the following example for $p = 2$.

Example 2.15 (Second-order one-step method based on Taylor expansion). Given $f \in C^2([t_0,T] \times \mathbb{R}^n; \mathbb{R}^n)$, define the incremental function

$$\Phi(t, y, h) := f(t, y) + \frac{h}{2}\,\big[(\partial_t f)(t, y) + (\partial_y f)(t, y)\, f(t, y)\big]. \tag{2.21}$$

Then, the corresponding explicit one-step method has consistency order $p = 2$.

One drawback of the latter construction is that it requires high regularity of f to write down the incremental function. Moreover, one has to provide explicit formulas for the derivatives of f and, more importantly, stability (2.14) of the one-step method requires additional assumptions.

Alternatively, one can consider nested evaluations of f to avoid the evaluation of the derivatives of f.

Example 2.16 (Modified Euler method (Runge 1895)). Starting from the integral equation

$$y(t+h) = y + \int_t^{t+h} f(\tau, y(\tau))\,\mathrm{d}\tau$$

we can expect a consistency order $p = 2$, if the integral is replaced by the midpoint rule

$$\int_t^{t+h} f(\tau, y(\tau))\,\mathrm{d}\tau = hf\left(t + \frac{h}{2}, y\left(t + \frac{h}{2}\right)\right) + \mathcal{O}(h^3).$$

Now, the unknown value $y\left(t + \frac{h}{2}\right)$ is approximated by the explicit Euler method

$$y\left(t + \frac{h}{2}\right) = y(t) + \frac{h}{2} f(t, y(t)) + \mathcal{O}(h^2)$$

leading to the nested evaluations

$$\Phi(t, y, h) := f\left(t + \frac{h}{2}, y + \frac{h}{2} f(t, y)\right). \tag{2.22}$$

The following proposition generalizes this approach.

Proposition 2.17. *Given $a, b_1, b_2, c \in \mathbb{R}$, define the incremental function*

$$\Phi(t, y, h) := b_1 f(t, y) + b_2 f\big(t + ch, y + ah\, f(t, y)\big). \tag{2.23}$$

Then, the following statements are equivalent:

(i) *The corresponding one-step method has consistency order $p = 2$.*
(ii) $a = c$, $\quad b_1 + b_2 = 1$, $\quad$ *and* $\quad b_2 a = 1/2$.

Remark 2.18. The conditions in Proposition 2.17(ii) show that one does not have four degrees of freedom (i.e., $a, b_1, b_2, c \in \mathbb{R}$), but only one to ensure second-order consistency. Fixing any of these constants, all other constants are determined as well. Moreover, second-order consistency requires $a \neq 0 \neq c$. Finally, we note that a reasonable method will additionally impose the restriction $c \in (0, 1]$ to ensure that the evaluation point $t+ch$ will belong to the time interval $[t_0, T]$.

Proof of Proposition 2.17. Let $f \in C^2([t_0, T]\times\mathbb{R}^n; \mathbb{R}^n)$ and hence $y \in C^3([t_0, T]; \mathbb{R}^n)$ be the exact solution. The Taylor expansion of Φ around $h = 0$ proves that

$$\Phi(t, y, h) = \Phi(t, y, 0) + h\,\partial_h \Phi(t, y, 0) + \mathcal{O}(h^2).$$

Note that

$$\begin{aligned} \Phi(t, y, 0) &= b_1 f(t, y) + b_2 f(t, y) = (b_1 + b_2)\, f(t, y), \\ \partial_h \Phi(t, y, 0) &= b_2 c\, \partial_t f(t, y) + b_2 a\, \partial_y f(t, y)\, f(t, y). \end{aligned}$$

Hence,

$$\Phi(t, y, h) = (b_1 + b_2)\, f(t, y) + h\left[b_2 c\, \partial_t f(t, y) + b_2 a\, \partial_y f(t, y)\, f(t, y)\right] + \mathcal{O}(h^2). \tag{2.24}$$

Carl Runge (1856–1927) *was a German mathematician. He studied at LMU München and University of Bremen, before he took his PhD in 1880 at HU Berlin supervised by Weierstrass. He did his habilitation in 1883 at HU Berlin and became Professor at University of Hannover in 1886. In 1904, he accepted the Chair of Applied Mathematics at University of Göttingen (the first chair for applied mathematics in Germany). Runge is the father-in-law of Richard Courant.*

Taylor expansion of y around t proves that

$$\begin{aligned} y(t+h) &= y(t) + h\,y'(t) + \frac{h^2}{2}\,y''(t) + \mathcal{O}(h^3) \\ &\overset{(2.20)}{=} y(t) + h\,f\big(t,y(t)\big) + \frac{h^2}{2}\,\big[(\partial_t f)\big(t,y(t)\big) + (\partial_y f)\big(t,y(t)\big)\,f\big(t,y(t)\big)\big] + \mathcal{O}(h^3). \end{aligned} \tag{2.25}$$

Combining (2.24)–(2.25), we can identify the consistency error. To ease the presentation, we omit the arguments $(t, y(t))$ for the f-terms. Then,

$$\begin{aligned} &y(t+h) - \big[y(t) + h\,\Phi\big(t,y(t),h\big)\big] \\ &\quad = h\,f + \frac{h^2}{2}\,\big[\partial_t f + (\partial_y f)\,f\big] - h\,(b_1+b_2)\,f - h^2\,\big[b_2 c\,\partial_t f + b_2 a\,(\partial_y f)\,f\big] + \mathcal{O}(h^3) \\ &\quad = h\,\big[1-(b_1+b_2)\big]\,f + h^2\,\big[1/2 - b_2 c\big]\,\partial_t f + h^2\,\big[1/2 - b_2 a\big]\,(\partial_y f)\,f + \mathcal{O}(h^3). \end{aligned}$$

Consequently,

$$y(t+h) - \big[y(t) + h\,\Phi\big(t,y(t),h\big)\big] = \mathcal{O}(h^3)$$

is equivalent to the fact that all lower-order powers of h vanish, i.e.,

$$b_1 + b_2 = 1, \quad b_2 c = 1/2, \quad b_2 a = 1/2.$$

In particular, it follows that $a \neq 0$, $c \neq 0$, $b_2 \neq 0$, and $a = c$. This concludes the proof. □

Chosing $a = 1/2$ leads with Proposition 2.17(ii) to the modified Euler method 2.16.

Example 2.19 (Heun method (1900)). Choose $a = 1$. Then, Proposition 2.17(ii) suggests the choice $c = 1$, $b_1 = 1/2 = b_2$. Define

$$\Phi(t,y,h) := \frac{1}{2}\,f(t,y) + \frac{1}{2}\,f\big(t+h,\; y + h\,f(t,y)\big). \tag{2.26}$$

According to Proposition 2.17, the resulting one-step method has consistency order $p = 2$. Clearly, Lipschitz continuity of f in y yields stability (2.14). For the implementation, the Heun method computes

$$\begin{aligned} k_1 &:= f(t,y), \\ k_2 &:= f(t+h,\; y + hk_1), \\ \Phi(t,y,h) &:= \frac{k_1+k_2}{2}. \end{aligned}$$

Remark 2.20. As far as the effectivity of an integrator is concerned, one should plot the error vs. the number of f-evaluations. Then, higher-order methods pay, if the solution is smooth: For a uniform time-step size h, the number of time-steps satisfies $N = (T - t_0)/h$.

- The explicit Euler method has N evaluations of f, and the error decays like $\mathcal{O}(h) = \mathcal{O}(N^{-1})$, i.e., we get an algebraic decay with rate 1.
- The modified Euler method has $2N$ evaluations of f, and the error decays like $\mathcal{O}(h^2) = \mathcal{O}(N^{-2})$, i.e., we get an algebraic decay with rate 2.

Asymptotically, the modified Euler method will thus beat the explicit Euler method with respect to accuracy and computational time.

Karl Heun (1859–1929) *was a German mathematician. He took his PhD from the University of Göttingen in 1881 and completed his habilitation at LMU München. From 1890–1902, he worked as a teacher in Berlin, before he obtained the chair of theoretical mechanics at Technische Hochschule Darmstadt, where he retired in 1922.*

Remark 2.21. Second-order consistency / convergence is only visible if f (and hence y) are sufficiently smoooth. If f is not smooth, then numerical experiments with uniform meshes lead to order reductions.

2.5. Explicit Runge–Kutta methods

In this section, we generalize / formalize the idea of Proposition 2.17.

Definition 2.22. Let $A \in \mathbb{R}^{m\times m}$ be strictly lower triangular (i.e., $A_{ij} = 0$ for $i \le j$), $b, c \in \mathbb{R}^m$ with $0 \le c_1 \le c_2 \le \dots \le c_m \le 1$. Then, a one-step method with incremental function

$$\Phi(t, y, h) := \sum_{j=1}^{m} b_j k_j, \tag{2.27}$$

where the so-called **increments** satisfy that

$$k_i = f\Big(t + c_i h,\ y + h \sum_{j=1}^{i-1} A_{ij} k_j\Big) \quad \text{for all } i = 1, \dots, m, \tag{2.28}$$

is called **explicit m-stage Runge–Kutta method**. The y-arguments $y + h\sum_{j=1}^{i-1} A_{ij}k_j$ of the increments k_i are called **stages**. Usually, Runge–Kutta methods are denoted by their **Butcher tableau** $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$. If the data are explicitly given, usually the zero entries of A are omitted (see examples below).

Example 2.23.

- The **explicit Euler method** has the Butcher tableau

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

- The **modified Euler** has the Butcher tableau

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

- The **Heun method** has the Butcher tableau

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

Martin Wilhelm Kutta (1867–1944) *was a German mathematician. He studied mathematics at TU München. In 1901, we took his PhD with the thesis* "Beitrag zur näherungsweisen Integration totaler Differentialgleichungen"*, where he was generalizing the ideas of Runge. Later, he was associate professor at University of Jena (1909–1910), RWTH Aachen (1910–1912), before he became full professor at University of Stuttgart in 1912. He retired in 1935.*

John Charles Butcher (born 1933) *is a mathematician from New Zealand. He took his PhD at University of Sydney in 1961. In 1966, he became professor at the University of Auckland. He retired in 1999. He has made fundamental contributions to the mathematical understanding of Runge–Kutta methods. In 2010, Butcher was awarded the Jones Medal from the Royal Society of New Zealand for his "exceptional lifetime work on numerical methods for the solution of differential equations".*

PROPOSITION 2.24 (Consistent Runge–Kutta methods provide quadrature rules). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an (explicit) m-stage Runge–Kutta method with consistency order $p \geq 1$. Then, the vectors $c, b \in \mathbb{R}^m$ provide a quadrature, which is exact for polynomials $q \in \mathbb{P}_{p-1}$, i.e.,*

$$\sum_{j=1}^{m} b_j q(c_j) = \int_0^1 q(s)\,\mathrm{d}s \quad \textit{for all } q \in \mathbb{P}_{p-1}. \tag{2.29}$$

PROOF. Consider the integration problem $f(t, y) := t^p$ and $y(0) = 0$. The exact solution is the antiderivative $y(t) = \int_0^t t^p\,\mathrm{d}s = \frac{t^{p+1}}{p+1}$. On the one hand, it holds that

$$k_j \overset{(2.28)}{=} (t + c_j h)^p = \sum_{i=0}^{p} \binom{p}{i} t^{p-i}\, c_j^i\, h^i$$

and hence

$$\Phi(t, y, h) = \sum_{j=1}^{m} b_j k_j = \sum_{i=0}^{p} \binom{p}{i} t^{p-i}\, h^i \sum_{j=1}^{m} b_j c_j^i. \tag{2.30}$$

On the other hand, note that

$$y^{(i)}(t) = \frac{1}{p+1}\,\frac{(p+1)!}{(p+1-i)!}\, t^{p+1-i} = \frac{p!}{(p+1-i)!}\,\frac{(i-1)!}{(i-1)!}\, t^{p+1-i} = \binom{p}{i-1}\,(i-1)!\, t^{p+1-i}$$

and hence

$$y(t+h) - y(t) = \sum_{i=1}^{p+1} \frac{y^{(i)}(t)}{i!}\, h^i = \sum_{i=1}^{p+1} \binom{p}{i-1} \frac{1}{i}\, t^{p+1-i}\, h^i = \sum_{i=0}^{p} \binom{p}{i} \frac{1}{i+1}\, t^{p-i}\, h^{i+1}. \tag{2.31}$$

Combining (2.30)–(2.31), we obtain for the consistency error that

$$\mathcal{O}(h^{p+1}) = y(t+h) - \big[y(t) + h\,\Phi\big(t, y(t), h\big)\big] = \sum_{i=0}^{p} \binom{p}{i}\, t^{p-i}\, h^{i+1} \left[\frac{1}{i+1} - \sum_{j=1}^{m} b_j c_j^i\right].$$

Consequently, the lower-order powers of h must vanish, i.e.,

$$\int_0^1 s^i\,\mathrm{d}s = \frac{1}{i+1} = \sum_{j=1}^{m} b_j c_j^i \quad \text{for all } i = 0, \dots, p-1. \tag{2.32}$$

The identity (2.32) proves that the quadrature rule is exact for all monomials $1, s, s^2, \dots, s^{p-1}$. Due to linearity, this proves (2.29). □

EXAMPLE 2.25 (Classical Runge–Kutta method RK4 (Runge 1901)). Recall the Simpson rule

$$\int_0^1 g\,\mathrm{d}s \approx \frac{1}{6}\,\big[g(0) + 4\,g(1/2) + g(1)\big], \tag{2.33}$$

which is exact for polynomials of degree 3, i.e.,

$$\int_0^1 q\,\mathrm{d}s = \frac{1}{6}\,\big[q(0) + 4\,q(1/2) + q(1)\big]. \tag{2.34}$$

Due to Proposition 2.24, we aim to extend the Simpson rule to a Runge–Kutta method of order $p = 4$. Runge introduced the method

$$\begin{array}{c|cccc} 0 & & & & \\ 1/2 & 1/2 & & & \\ 1/2 & 0 & 1/2 & & \\ 1 & 0 & 0 & 1 & \\ \hline & 1/6 & 1/3 & 1/3 & 1/6 \end{array} \quad . \tag{2.35}$$

In explicit terms, the method reads

$$\begin{aligned} k_1 &= f(t, y), \\ k_2 &= f\Big(t + \frac{h}{2}\,,\, y + \frac{h}{2}\,k_1\Big), \\ k_3 &= f\Big(t + \frac{h}{2}\,,\, y + \frac{h}{2}\,k_2\Big), \\ k_4 &= f(t + h\,,\, y + h\,k_3), \\ \Phi(t, y, h) &= \frac{1}{6}\,\big[k_1 + 2\,k_2 + 2\,k_3 + k_4\big]. \end{aligned}$$

Note that (2.35) in fact induces the Simpson rule. By Taylor expansion, one can show that this method has consistency order $p = 4$. Moreover, we will see below that it needs (at least) $m = 4$ stages to get consistency order $p = 4$. In this sense, RK4 is optimal.

Example 2.26 (Another extension of the Simpson rule). Consider the Butcher tableau

$$\begin{array}{c|ccc} 0 & & & \\ 1/2 & 1/2 & & \\ 1 & -1 & 2 & \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \quad . \tag{2.36}$$

Note that (2.36) also induces the Simpson rule. Together with Runge's method from Example 2.25, we thus see that the extension of a quadrature rule to a Runge–Kutta method is not unique. By Taylor expansion, one can show that the method (2.36) has consistency order $p = 3$. Moreover, we will see below that this consistency order is maximal (due to $m = 3$ stages).

Theorem 2.27 (Stability of explicit Runge–Kutta methods). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an (explicit) m-stage Runge–Kutta method. Then, there exist $C_{\text{stab}} > 0$ such that*

$$\forall\, t \in [t_0, T)\, \forall\, h \in (0, T - t]\, \forall\, y, \widetilde{y} \in \mathbb{R}^n : \quad \|\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)\| \le C_{\text{stab}}\, \|y - \widetilde{y}\|. \tag{2.37}$$

Let $L > 0$ be the Lipschitz constant of f in y. Then, it holds that

$$C_{\text{stab}} \le L\, p(hL), \quad \textit{where} \quad p(s) = \sum_{j=0}^{m-1} \mu_j s^j \in \mathbb{P}_{m-1} \tag{2.38}$$

with $\mu_j \ge 0$ and $\mu_0 = \sum_{j=1}^{m} |b_j|$

Proof. Recall the increments $k_i = f\big(x + c_i h,\, y + h \sum_{j=1}^{i-1} A_{ij} k_j\big)$. Let $\widetilde{k}_i$ denote the increments with respect to $\widetilde{y}$. The proof is split into two steps.

Step 1. We prove by induction on $i = 1, \dots, m$ that

$$\|k_i - \widetilde{k}_i\| \le L\, q_i(hL)\, \|y - \widetilde{y}\| \quad \text{with} \quad q_i(s) = \sum_{j=0}^{i-1} \lambda_j s^j \in \mathbb{P}_{i-1}, \quad \lambda_j \ge 0, \quad \lambda_0 = 1. \tag{2.39}$$

For $i = 1$, Lipschitz continuity of f in y proves that

$$\|k_1 - \widetilde{k}_1\| = \|f(t,y) - f(t,\widetilde{y})\| \le L\, \|y - \widetilde{y}\|.$$

This proves (2.39) for $i = 1$. In the induction step $(i-1) \rightsquigarrow i$, note that

$$\begin{aligned}
\|k_i - \widetilde{k}_i\| &\le L \Big\| y - \widetilde{y} + h \sum_{j=1}^{i-1} A_{ij}(k_j - \widetilde{k}_j) \Big\| \\
&\le L \Big[\|y - \widetilde{y}\| + h \sum_{j=1}^{i-1} |A_{ij}|\, \|k_j - \widetilde{k}_j\| \Big] \\
&\le L\, \|y - \widetilde{y}\| \underbrace{\Big[1 + hL \sum_{j=1}^{i-1} |A_{ij}|\, q_j(hL) \Big]}_{=:q_i(hL)}.
\end{aligned}$$

Hence, (2.39) holds for all $i = 1, \dots, m$.

Step 2. Finally, it follows that

$$\|\Phi(t,y,h) - \Phi(t,\widetilde{y},h)\| = \Big\| \sum_{j=1}^{m} b_j (k_j - \widetilde{k}_j) \Big\| \overset{(2.39)}{\le} L \underbrace{\sum_{j=1}^{m} |b_j|\, q_j(hL)}_{=:p(hL)} \|y - \widetilde{y}\|.$$

This concludes the proof. □

Theorem 2.28 (Necessary consistency conditions for Runge–Kutta methods). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an (explicit) m-stage Runge–Kutta method with consistency order $p \ge 1$. Then, there hold the following statements:*

(i) $\displaystyle\sum_{j=1}^{m} b_j = 1.$

(ii) $\displaystyle\sum_{j=1}^{m} b_j c_j^\ell = \frac{1}{\ell+1}$ *for all $\ell = 0, \dots, p-1$.*

(iii) $\displaystyle\sum_{j=1}^{m} b_j (A^\ell \mathbb{1})_j = \frac{1}{(\ell+1)!}$ *for all $\ell = 0, \dots, p-1$, where $\mathbb{1} = (1, \dots, 1)^\top \in \mathbb{R}^m$.*

Proof. (i) follows already from (ii) and (iii) for $\ell = 0$. (ii) has already been proved in Proposition 2.24; see the identity (2.32). Hence, it only remains to prove (iii).

To prove (iii), we consider the scalar problem $y' = f(t,y) := y$ on $[0,1]$ with $y(0) = 1$. The unique solution is $y(t) = e^t$. Let $k := (k_1, \dots, k_m)^\top$ and note that

$$k_i = f\Big(t + c_i h,\, y + h \sum_{j=1}^{i-1} A_{ij} k_j\Big) = y + h \sum_{j=1}^{i-1} A_{ij} k_j = y + h\,(Ak)_i \quad \text{for all } i = 1, \dots, m.$$

In vector form, this identity reads

$$(I - hA)\,k = k - h\,Ak = y\,\mathbb{1}.$$

For small h, it holds that $h\,\|A\| < 1$. Hence, the so-called Neumann series proves that

$$I - hA \quad \text{is invertible with} \quad (I-hA)^{-1} = \sum_{\ell=0}^{\infty} (hA)^\ell.$$

Therefore,

$$\Phi(t,y,h) = \sum_{j=1}^{m} b_j k_j = b \cdot (I-hA)^{-1}(y\mathbb{1}) = b \cdot \sum_{\ell=0}^{\infty} (hA)^\ell (y\mathbb{1}) = y \sum_{\ell=0}^{\infty} h^\ell b \cdot (A^\ell \mathbb{1}).$$

On the other hand, Taylor expansion for $y(t) = e^t$ gives

$$y(t+h) - y(t) = \sum_{\ell=1}^{\infty} \frac{y^{(\ell)}(t)}{\ell!}\, h^\ell = y(t) \sum_{\ell=1}^{\infty} \frac{h^\ell}{\ell!} = y(t) \sum_{\ell=0}^{\infty} \frac{h^{\ell+1}}{(\ell+1)!}.$$

Combining these two identities, we obtain for the consistency error that

$$\mathcal{O}(h^{p+1}) = y(t+h) - \big[y(t) + h\,\Phi\big(t, y(t), h\big)\big] = y(t) \sum_{\ell=0}^{\infty} h^{\ell+1} \Big[\frac{1}{(\ell+1)!} - b \cdot (A^\ell \mathbb{1})\Big].$$

Hence, lower-order powers of h must vanish, i.e.,

$$\frac{1}{(\ell+1)!} = b \cdot (A^\ell \mathbb{1}) \quad \text{for all } \ell = 0, \dots, p-1.$$

This concludes the proof. □

COROLLARY 2.29 (Naive Butcher barriers for explicit Runge–Kutta methods). *For any explicit m-step Runge–Kutta method, the consistency order $p \geq 1$ satisfies that $p \leq m$.*

PROOF. Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be the Butcher tableau of an explicit m-stage Runge–Kutta method. Recall that $A = (A_{ij}) \in \mathbb{R}^{m\times m}$ is strictly lower triangular, i.e., $A_{ij} = 0$ for all $i \leq j$. Let $A^\ell = (A_{ij}^{(\ell)})$. We show by induction on $\ell = 1, \dots, m$ that

$$A_{ij}^{(\ell)} = 0 \quad \text{for all } i \leq j + \ell - 1. \tag{2.40}$$

Obviously, the claim is OK for $\ell = 1$. For the induction step $(\ell-1) \rightsquigarrow \ell$, note that

$$A_{ij}^{(\ell)} = \sum_{k=1}^{m} A_{ik} A_{kj}^{(\ell-1)} = \sum_{k=1}^{i-1} A_{ik} A_{kj}^{(\ell-1)} \overset{(2.40)}{=} \sum_{k=j+\ell-1}^{i-1} A_{ik} A_{kj}^{(\ell-1)}.$$

The latter sum is empty (and hence 0) if $i - 1 < j + \ell - 1$, i.e., $i < j + \ell$. This concludes the proof of (2.40).

For $\ell = m$, it follows from (2.40) that $A^m = 0$. Hence, Theorem 2.28(iii) cannot be satisfied for $\ell = m$. Therefore, we conclude that $p \leq m$. □

EXERCISE 2.30. Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be an (explicit) m-stage Runge–Kutta method. Show that the method has at least consistency order $p = 1$, if and only if $\sum_{j=1}^m b_j = 1$.

Exercise 2.31. Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be an (explicit) m-stage Runge–Kutta method. Suppose that

$$\sum_{j=1}^{m} b_j = 1, \quad \sum_{j=1}^{m} b_j c_j = \frac{1}{2}, \quad \text{and} \quad \sum_{j=1}^{i-1} A_{ij} = c_i \quad \text{for all } i = 1, \dots, m. \tag{2.41}$$

Argue as for Proposition 2.17 to show that the considered Runge–Kutta method has at least consistency order $p = 2$.

Remark 2.32. In Exercise 2.31, we have seen that the consistency conditions of Theorem 2.28 are (essentially) sharp for $m = 2 = p$. However, it is known that the Butcher barriers $p \leq m$ are *not sharp* in general. For $m \geq 5$, there are no explicit Runge–Kutta methods with consistency order $p = m$, i.e., it holds that $m - p \geq 1$; see [**But08**, Theorem 324B]. Moreover $m - p \geq 2$ for $p \geq 7$ and $m - p \geq 3$ for $p \geq 8$; see [**But08**, page 188].

The precise growth of the maximal order $p_{\max}(m)$ with respect to m as well as the the minimal stage number $m_{\min}(p)$ with respect to p are still unknown, however $p_{\max}(m) \to \infty$ as $m \to \infty$; see [**But08**, Theorem 324C].

The monograph [**SWP12**, Satz 2.4.6] refers to the original literature for the table

p	1	2	3	4	5	6	7	8	$p \geq 9$
$m_{\min}(p)$	1	2	3	4	6	7	9	11	$m_{\min}(p) \geq p + 3$

Conversely,

m	1	2	3	4	5	6	7	8	9	$m \geq 9$
$p_{\max}(m)$	1	2	3	4	4	5	6	6	7	$p_{\max}(m) \leq m - 2$

For $p = 10$, Hairer (1978) found an explicit Runge–Kutta method with $m = 17$. If $m_{\min}(10) < 17$ appears to be still open.

2.6. Adaptive time-step control

In this section, we aim to (heuristically) design an algorithm of the following type:

Algorithm 2.33 (Theoretical adaptive algorithm).

Input: Time interval $[t_0, T]$, initial value y_0, right-hand side $f(\cdot, \cdot)$, one-step method with incremental function $\Phi(t, y, h)$, tolerance $\tau > 0$, initial time-step size h_0, counter $\ell = 0$.

```
repeat
    Determine $h > 0$ such that $t + h \leq T$ and $\| z(t_\ell + h) - [y_\ell + h\,\Phi(t_\ell, y_\ell, h)] \| \approx \tau\, h$,
    where $z$ solves $z(t_\ell) = y_\ell$ and $z' = f(t, z)$ in $[t_\ell, T]$.
    Define $h_\ell := h$, $t_{\ell+1} := t_\ell + h_\ell$, $y_{\ell+1} := y_\ell + h_\ell\,\Phi(t_\ell, y_\ell, h_\ell)]$.
    Update counter $\ell \mapsto \ell + 1$.
until $t_\ell = T$
```

Output: Mesh $\Delta = \{t_0, \dots, t_N = T\}$, approximations $y_\ell \approx y(t_\ell)$ for all $\ell = 0, \dots, N$.

An adaptive choice of time-step is reasonable in terms of

Ernst Hairer *(born 1949) is an Austrian mathematician. He took is PhD in 1972 at University of Innsbruck (supervized by Gerhard Wanner). In 1985, he became Associate Professor at University of Geneva, Switzerland. Since 1999, he is full professor at University of Geneva. He has coauthored the two-volume monograph "Solving Ordinary Differential Equations", which is the main reference for research in the field of numerical integrators. Ernst Hairer is the father of the mathematician Martin Hairer, who won the fields medal in 2014.*

- **Efficiency:** We want to compute / approximate $y(T)$ as cheaply as possible, i.e., with as few evaluations of f as possible such that the error satisfies that $\|y(T) - y_N\| \approx \tau$.
- **Reliability:** Even if the solution is smooth, the error behavior $\|y(T) - y_N\| = \mathcal{O}(h_\Delta^p)$ hides a multiplicative constant. Possibly, a chosen uniform step-size $h = h_\Delta > 0$ is to coarse to ensure that $\|y(T) - y_N\| \approx \tau$.
- If the solution y is non-smooth, then this will spoil the experimental convergence order. For uniform meshes, one usually obtains a convergence $\mathcal{O}(h^q) = \mathcal{O}(N^{-q})$, where $q > 0$ is (much) smaller than the consistency order $p \geq 1$. However, if the mesh is locally adapted to the points of reduced regularity (i.e., the *singularities*), in many cases one can recover $\|y(T) - y_N\| = \mathcal{O}(N^{-p})$ for the error.

Since the exact solutions to ODEs are in general unknown, the algorithm cannot evaluate the consistency error to ensure that

$$\|z(t_\ell + h) - \big[y_\ell + h\,\Phi(t_\ell, y_\ell, h)\big]\| \approx \tau\, h,$$

where z solves $z(t_\ell) = y_\ell$ and $z' = f(t, z)$ in $[t_\ell, T]$. Therefore, we require appropriate heuristics to circumvent this lack of knowledge. In the most common approach, instead of the unknown value $z(t_\ell + h)$ a second numerical approximation is used. If this approximation is better than $\big[y_\ell + h\,\Phi(t_\ell, y_\ell, h)\big]$, we can use it to estimate the consistency error.

In principle, there exist two ways to construct this second approximation:

(1) $\boldsymbol{h - h/2}$ strategy: We use the same numerical method, but two steps of size $h/2$ instead of one step of size h.
(2) $\boldsymbol{p - (p+1)}$ strategy: We use the same step-size but a different numerical method of higher order.

Here, we start with the second strategy. Recall that consistency proofs are obtained by means of the Taylor expansion. For instance, let us consider the explicit Euler method. If the solution is smooth, then

$$\begin{aligned} z(t+h) &= z(t) + h\,z'(t) + \frac{h^2}{2}\,z''(t) + \mathcal{O}(h^3) \\ &= z(t) + h\,\Phi\big(t, z(t), h\big) + \Big[\frac{h^2}{2}\,z''(t) + \mathcal{O}(h^3)\Big]. \end{aligned}$$

For a general one-step method with consistency order $p \geq 1$, one can similarly show that

$$z(t+h) = z(t) + h\,\Phi\big(t, z(t), h\big) + h^{p+1}\,c(t) + \mathcal{O}(h^{p+2}) \tag{2.42}$$

with $c(t) \sim z^{(p+1)}(t)$. Define

$$\begin{aligned} z_1 &:= z(t) + h\,\Phi\big(t, z(t), h\big), \\ \widetilde{z}_1 &:= z(t) + h\,\widetilde{\Phi}\big(t, z(t), h\big), \end{aligned}$$

where $\widetilde{\Phi}$ belongs to a method with consistency order $p+1$. Then, it holds that

$$\begin{aligned} z(t+h) - z_1 &= h^{p+1}\,c(t) + \mathcal{O}(h^{p+2}), \\ z(t+h) - \widetilde{z}_1 &= \mathcal{O}(h^{p+2}). \end{aligned}$$

Hence,

$$z_1 - \widetilde{z}_1 = \big[z(t+h) - \widetilde{z}_1\big] - \big[z(t+h) - z_1\big] = -c(t)\,h^{p+1} + \mathcal{O}(h^{p+2}).$$

We obtain that

$$\|z_1 - \widetilde{z}_1\| = \|c(t)\|\, h^{p+1} + \mathcal{O}(h^{p+2})$$

and consequently

$$\|c(t)\| = \frac{\|z_1 - \widetilde{z}_1\|}{h^{p+1}} + \mathcal{O}(h). \tag{2.43}$$

We aim to find an H such that

$$\begin{aligned} \tau\, H \overset{!}{=} \big\|z(t+H) - \big[z(t) + H\,\Phi\big(t, z(t), H\big)\big]\big\| &\overset{(2.42)}{=} \|c(t)\|\, H^{p+1} + \mathcal{O}(H^{p+2}) \\ &\overset{(2.43)}{=} \frac{\|z_1 - \widetilde{z}_1\|}{h^{p+1}} H^{p+1} + \mathcal{O}(hH^{p+1}) + \mathcal{O}(H^{p+2}). \end{aligned}$$

If we neglect the higher-order terms $\mathcal{O}(hH^{p+1})$ and $\mathcal{O}(H^{p+2})$, we obtain that

$$\tau\, H \approx \frac{\|z_1 - \widetilde{z}_1\|}{h^{p+1}} H^{p+1}.$$

Rearranging this estimate, we are led to

$$H^p \approx \frac{\tau\, h^{p+1}}{\|z_1 - \widetilde{z}_1\|} = \frac{\tau\, h^p}{\big\|\Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big)\big\|},$$

where we recall that $z_1 - \widetilde{z}_1 = h\,\big[\Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big)\big]$. Note, that the right hand side is computable.

Heuristics 2.34 (Step-size control based on an $\boldsymbol{p - (p+1)}$ strategy). Let $\Phi(t, y, h)$ be the incremental function of an explicit one-step method with consistency order $p \geq 1$. Let $\widetilde{\Phi}(t, y, h)$ be the incremental function of an explicit one-step method with consistency order $p+1$. Let $h > 0$ be given and $\Phi(t, y, h)$ as well as $\widetilde{\Phi}(t, y, h)$ be computed. Then, (up to higher-order terms)

$$H := \left[\frac{\tau}{\big\|\Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big)\big\|}\right]^{1/p} h \tag{2.44a}$$

would be OK to ensure that

$$z(t+H) - \big[z(t) + H\,\Phi\big(t, z(t), H\big)\big] \approx H\,\tau. \tag{2.44b}$$

In explicit terms: If you compute with step-size h, you get a feedback on the appropriate step-size H.

Algorithm 2.35 (Practical adaptive algorithm).
Input: Time interval $[t_0, T]$, initial value y_0, right-hand side $f(\cdot, \cdot)$, one-step method $\Phi(t, y, h)$ of order $p \geq 1$, auxiliary one-step method $\widetilde{\Phi}(t, y, h)$ of order $p+1$, tolerance $\tau > 0$, initial time-step size $h > 0$, minimal time-step size $h_{\min} > 0$, conformity factor $\lambda \geq 1$, safety factor $0 < \varrho \leq 1$, counter $\ell = 0$.

```
repeat
    $h := \min\big\{T - t_\ell,\ \max\{h_{\min}, h\}\big\}$
    $F := \widetilde{\Phi}\big(t_\ell, y_\ell, h\big)$
    $H := \varrho \left[\frac{\tau}{\|\Phi(t_\ell, y_\ell, h) - F\|}\right]^{1/p} h$
    if $h \leq H$ OR $h \leq h_{\min}$ then
        $t_{\ell+1} := t_\ell + h$
        $y_{\ell+1} := y_\ell + hF$
```

```
 8:         if $t_{\ell+1} < T$ then
 9:             $h := \min\{H, \lambda h\}$
10:             Update counter $\ell \mapsto \ell + 1$
11:         end if
12:     else
13:         $h := \min\{H, h/\lambda\}$
14:     end if
15: until $t_{\ell+1} = T$
```

Output: Mesh $\Delta = \{t_0, \dots, t_N = T\}$ and corresponding approximations $y_\ell \approx y(t_\ell)$ for all $\ell = 0, \dots, N$).

Remark 2.36 (Comments on Algorithm 2.35).

Line 2 and 5: To ensure that the algorithm is finite, we have to guarantee that $h \not\to 0$. In fact, the algorithm ensures that $h \geq h_{\min}$ up to the final time-step (where possibly $T - t < h_{\min}$).

Line 4 (and 3): Up to the safety factor $0 < \varrho \leq 1$, we use the formula for H derived in (2.44). The safety factor tries to cover the fact that (2.44) was derived by neglecting higher-order terms (which could have an impact in our computation).

Line 5: In order to avoid f-evaluations, we accept the time-step, if $h \leq H$, i.e., if the current step-size is smaller than the step-size allowed by our heuristics (2.44) (i.e., a longer time-step would have been OK).

Line 7 (and 3): Usually $\widetilde{\Phi}(t_\ell, y_\ell, h)$ is more accurate. Therefore, we use $\widetilde{\Phi}(t_\ell, y_\ell, h)$ instead of $\Phi(t_\ell, y_\ell, h)$ for the next time step.

Line 8 and 9: If the time-step was accepted, but the final time T has not been reached, the algorithm requires a guess for the step-size in the next time-step. To this end, we aim to choose $h = H$, but we enforce that the growth of the step-size is not too big (i.e., the ratio is bounded by λ if the step-size is increased). Again, we intend to avoid f-evaluations. Therefore, the step-size guess should be accepted in the next time-step.

Line 12 and 13: If $H < h$, then we cannot accept the time-step $t_\ell + h$ and we have to recompute with a smaller h. To this end, we aim to choose $h = H$, but enforce at least a uniform reduction of the current step-size (by the factor λ^{-1}).

Attention: Since we have made certain simplifications to get the error estimate (2.44) for the consistency error, we cannot rigorously guarantee that the adaptive algorithm is mathematically reliable in the sense that $\|y(T) - y_N\| \leq \tau$ is guaranteed.

Practical choices of the parameters are the following:

- $h \sim \tau^{1/p}$, e.g., $h := \tau^{1/p}/10$, since $\tau \approx \|y(T) - y_N\| = \mathcal{O}(h^p)$ for smooth y.
- $h_{\min} = \tau$.
- $\lambda = 2$.
- $\varrho = 0.8$.

For the implementation of Algorithm 2.35, one aims to choose $\Phi(t, y, h)$ and $\widetilde{\Phi}(t, y, h)$ in a way that minimizes the number of f-evaluations. Practically relevant are so-called **embedded Runge–Kutta methods**, which use the same increments k_j (and c_j), but differ only for the vector $b \in \mathbb{R}^m$. They are usually denoted by

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

where $b \in \mathbb{R}^m$ belongs to the higher-order method and $\beta \in \mathbb{R}^m$ belongs to the lower-order method, i.e.,

$$\Phi(t,y,h) = \sum_{j=1}^{m} \beta_j k_j \quad \text{and} \quad \widetilde{\Phi}(t,y,h) = \sum_{j=1}^{m} b_j k_j.$$

Example 2.37 (Bogacki–Shampine pair RK3(2)). In 1989, Bogacki and Shampine proposed the embedded 4-stage Runge–Kutta method

$$\begin{array}{c|cccc}
0 & & & & \\
1/2 & 1/2 & & & \\
3/4 & 0 & 3/4 & & \\
1 & 2/9 & 1/3 & 4/9 & \\
\hline
 & 2/9 & 1/3 & 4/9 & 0 \\
 & 7/24 & 1/4 & 1/3 & 1/8
\end{array}$$

The vector $b = (2/9, 1/3, 4/9, 0)^\top$ belongs to a third-order method, while the vector $\beta = (7/24, 1/4, 1/3, 1/8)^\top$ belongs to a second-order method.

For one time-step, the method needs 4 f-evaluations instead of $3 + 2 = 5$ for non-embedded methods. Moreover, after the first time-step, RK3(2) has only 3 f-evaluations because of the **FSAL property** (i.e., first same as last): Since the last row of A coincides with b (together with $c_1 = 0$ and $c_4 = 1$), the last increment of the current time-step coincides with the first increment of the next time-step (i.e., $k_1(t_{\ell+1}) = k_4(t_\ell)$).

An adaptive Bogacki–Shampine method is provided by MATLAB function `ode23`.

Example 2.38 (Dormand–Prince pair RK5(4)). In 1980, Dormand and Prince proposed the embedded 7-stage Runge–Kutta method

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

The vector $b = ({}^{35}/_{384}, 0, {}^{500}/_{1113}, {}^{125}/_{192}, -{}^{2187}/_{6784}, {}^{11}/_{84}, 0)^\top \in \mathbb{R}^7$ belongs to the fifth-order method, while $\beta = ({}^{5179}/_{57600}, 0, {}^{7571}/_{16695}, {}^{393}/_{640}, -{}^{92097}/_{339200}, {}^{187}/_{210}, {}^{1}/_{40})^\top$ belongs to the fourth-order method.

Przemyslaw Bogacki, *Professor at Old Dominion University, Virginia, USA; see* `[webpage]`

Lawrence F. Shampine, *Professor Emeritus at Southern Methodist University, Texas, USA; see* `[webpage]`

John R. Dormand: *According to his book on "Numerical methods for differential equations", Dormand took his PhD in Physics at the Unviersity of York and then used to be a senior lecturer at the Department of Mathematics and Statistics, Teesside Polytechnique, Middlesbrough, Cleveland, UK. According to Scopus, his 1980 paper on RK5(4) has been cited more than 1400 times. His last paper appeared in 2003.*

Peter J. Prince *used to work at Department of Mathematics and Statistics, Teesside Polytechnique, Middlesbrough, Cleveland, UK. From 1978–1999, he was publishing papers on Runge–Kutta methods (mainly together with John R. Dormand).*

For one time-step, the method needs 7 f-evaluations instead of $5+4=9$ for non-embedded methods. As the Bogacki–Shampine pair RK3(2) from Example 2.37, RK5(4) has the FASL property and hence requires only 6 f-evaluatiosn after the first time-step.

An adaptive Dormand–Prince method is provided by MATLAB function `ode45`.

2.7. Extrapolation

If no auxiliary one-step method of order $p+1$ is at hand, one can use the Richardson extrapolation to obtain the higher-order method.

LEMMA 2.39 (Two-step Richardson extrapolation of one-step method). *Let $\Phi(t,y,h)$ be the incremental function of a stable, explicit one-step method with consistency order $p \geq 1$. Moreover, the solutions z to the initial value problem should be sufficiently smooth and the asymptotic expansion*

$$z(t+h) = z(t) + h\,\Phi\big(t, z(t), h\big) + h^{p+1}\,c(t) + \mathcal{O}(h^{p+2}), \tag{2.45}$$

with $c(t) \sim z^{(p+1)}(t)$ at least C^1 should hold. We define

$$z_1 := z(t) + h\,\Phi\big(t, z(h), h\big), \tag{2.46}$$

i.e. one step of this method with step-size h. Moreover, let

$$\widehat{z}_1 := \widehat{z}_{1/2} + \frac{h}{2}\,\Phi\big(t+h/2,\, \widehat{z}_{1/2},\, h/2\big) \quad \textit{with} \quad \widehat{z}_{1/2} := z(t) + \frac{h}{2}\,\Phi\big(t, z(t),\, h/2\big) \tag{2.47}$$

be the result of two successive steps with step-size $h/2$.

Then, it holds that

$$z(t+h) - \frac{2^p\,\widehat{z}_1 - z_1}{2^p - 1} = \mathcal{O}(h^{p+2}). \tag{2.48}$$

PROOF. Define

$$z_1^\star := z(t+h/2) + \frac{h}{2}\,\Phi\big(t+h/2,\, z(t+h/2),\, h/2\big).$$

Then,

$$\begin{aligned} z(t+h) - \widehat{z}_1 &= \big[z(t+h) - z_1^\star\big] + \big[z_1^\star - \widehat{z}_1\big] \\ &\overset{(2.45)}{=} \big[(h/2)^{p+1}\,c(t+h/2) + \mathcal{O}(h^{p+2})\big] \\ &\quad + \Big[z(t+h/2) - \widehat{z}_{1/2} + \frac{h}{2}\Big(\Phi\big(t+h/2,\, z(t+h/2),\, h/2\big) - \Phi\big(t+h/2,\, \widehat{z}_{1/2},\, h/2\big)\Big)\Big]. \end{aligned}$$

If Φ is stable in the sense of (2.14), consistency order p guarantees that

$$\begin{aligned} \big\|\Phi\big(t+h/2,\, z(t+h/2),\, h/2\big) - \Phi\big(t+h/2,\, \widehat{z}_{1/2},\, h/2\big)\big\| &\leq L\,\|z(t+h/2) - \widehat{z}_{1/2}\| \\ &= \mathcal{O}(h^{p+1}). \end{aligned}$$

Hence, we get that

$$\begin{aligned} z(t+h) - \widehat{z}_1 &= (h/2)^{p+1}\,c(t+h/2) + \big[z(t+h/2) - \widehat{z}_{1/2}\big] + \mathcal{O}(h^{p+2}) \\ &\overset{(2.45)}{=} (h/2)^{p+1}\,\big[c(t+h/2) + c(t)\big] + \mathcal{O}(h^{p+2}) \\ &= 2\,(h/2)^{p+1}\,c(t) + \mathcal{O}(h^{p+2}), \end{aligned} \tag{2.49}$$

where we have finally used the Taylor expansion $c(t+h/2) = c(t) + \mathcal{O}(h)$. Together with

$$z(t+h) - z_1 \overset{(2.45)}{=} c(t)\,h^{p+1} + \mathcal{O}(h^{p+2}),$$

we are led to

$$\widehat{z}_1 - z_1 = \big[z(t+h) - z_1\big] - \big[z(t+h) - \widehat{z}_1\big] = c(t)\, h^{p+1}\, \big[1 - 2^{-p}\big] + \mathcal{O}(h^{p+2}).$$

This yields that

$$c(t) = \frac{\widehat{z}_1 - z_1}{1 - 2^{-p}}\, h^{-(p+1)} + \mathcal{O}(h) \tag{2.50}$$

and hence

$$\begin{aligned} z(t+h) - \widehat{z}_1 &\overset{(2.49)}{=} 2\,(h/2)^{p+1}\, c(t) + \mathcal{O}(h^{p+2}) \\ &\overset{(2.50)}{=} 2\,(h/2)^{p+1}\, \frac{\widehat{z}_1 - z_1}{1 - 2^{-p}}\, h^{-(p+1)} + \mathcal{O}(h^{p+2}) \\ &= 2^{-p}\, \frac{\widehat{z}_1 - z_1}{1 - 2^{-p}} + \mathcal{O}(h^{p+2}) \\ &= \frac{\widehat{z}_1 - z_1}{2^p - 1} + \mathcal{O}(h^{p+2}). \end{aligned} \tag{2.51}$$

Since

$$\widehat{z}_1 + \frac{\widehat{z}_1 - z_1}{2^p - 1} = \frac{(2^p - 1)\,\widehat{z}_1 + (\widehat{z}_1 - z_1)}{2^p - 1} = \frac{2^p\,\widehat{z}_1 - z_1}{2^p - 1},$$

we conclude that

$$z(t+h) - \frac{2^p\,\widehat{z}_1 - z_1}{2^p - 1} = z(t+h) - \left[\widehat{z}_1 + \frac{\widehat{z}_1 - z_1}{2^p - 1}\right] \overset{(2.51)}{=} \mathcal{O}(h^{p+2}),$$

i.e., the extrapolated method has consistency order $p+1$. □

Heuristics 2.40 (Step-size control based on an $\boldsymbol{h - h/2}$ strategy). We use the notation of the last lemma. For given $h > 0$, let z_1 and $\widehat{z}_1$ be computed by (2.46)–(2.47). Then, (up to higher-order terms)

$$H := \left[\frac{2^p - 1}{2^p}\, \frac{\tau\, h^{p+1}}{\|z_1 - \widehat{z}_1\|}\right]^{1/p} \tag{2.52a}$$

would be OK to ensure that

$$z(t+H) - \big[z(t) + H\,\Phi\big(t, z(t), H\big)\big] \approx H\,\tau. \tag{2.52b}$$

In explicit terms: If you compute with step-size h, you get a feedback on the appropriate step-size H. In particular, we can also employ (2.52) to steer Algorithm 2.35.

We build on Heuristics 2.34 and use the auxiliary method provided by Lemma 2.39: Note that

$$\begin{aligned} h\left\{\Phi\big(t, z(t), h\big) - \widetilde{\Phi}\big(t, z(t), h\big)\right\} = z_1 - \frac{2^p\,\widehat{z}_1 - z_1}{2^p - 1} &= \frac{(2^p - 1)\, z_1 - (2^p\,\widehat{z}_1 - z_1)}{2^p - 1} \\ &= \frac{2^p}{2^p - 1}\, (z_1 - \widehat{z}_1). \end{aligned}$$

Therefore, (2.44) implies (2.52).

CHAPTER 3

Implicit one-step methods

Throughout this section, we consider the following model problem: Let $[t_0, T]$ be a given time-interval. For given $n \in \mathbb{N}$, let $f \in C([t_0,T] \times \mathbb{R}^n; \mathbb{R}^n)$ and suppose that f is Lipschitz continuous in y, i.e.,

$$\forall\, t \in [t_0,T]\, \forall\, y, \widetilde{y} \in \mathbb{R}^n : \quad \|f(t,y) - f(t,\widetilde{y})\| \le L\, \|y - \widetilde{y}\|, \tag{3.1}$$

where $\|\cdot\|$ is an arbitrary norm on $\mathbb{R}^n$ and $L > 0$ is a fixed constant. Then, for any initial value $y_0 \in \mathbb{R}^n$, the Picard–Lindelöf theorem guarantees existence and uniqueness of $y \in C^1([t_0,T]; \mathbb{R}^n)$ such that

$$y(t_0) = y_0 \quad \text{and} \quad y'(t) = f(t, y(t)) \quad \text{for all } t \in [t_0, T]. \tag{3.2}$$

Let $\Delta = \{t_0 < t_1 < \dots < t_N = T\}$ be a given mesh with local mesh-sizes $h_\ell := t_{\ell+1} - t_\ell > 0$ for all $\ell = 0, \dots, N-1$ and maximum mesh-size $h_\Delta := \max_{\ell=0,\dots,N-1} h_\ell$. As before, our task is to compute approximations

$$y_\ell \approx y(t_\ell) \quad \text{for all } \ell = 1, \dots, N. \tag{3.3}$$

3.1. Motivation

Let us assume for simplicity that $n = 1$ and $f(t,y) = \lambda y$ with $\Re(\lambda) < 0$. Of course, the solution to (3.2) is given by $y(t) = y_0 \exp(\lambda(t - t_0))$. So there is no need for a numerical method. Nevertheless, let us study the explicit and implicit Euler method applied to this problem.

The explicit Euler method (2.5) applied to this problem becomes

$$y_{\ell+1} = y_\ell + h_\ell \lambda y_\ell.$$

For an equidistant mesh, i.e. $h_\ell = h > 0$, the discrete solution of the explicit Euler method is given by

$$y_\ell = (1 + \lambda h)^\ell y_0, \qquad \ell = 0, \dots, N. \tag{3.4}$$

The implicit Euler method (2.6) leads to

$$y_\ell = \left(\frac{1}{1 - \lambda h}\right)^\ell y_0, \qquad \ell = 0, \dots, N. \tag{3.5}$$

Due to our assumptions $\Re(\lambda) < 0$ and $h > 0$, we have

$$|1 - \lambda h| \ge 1 - h\Re(\lambda) > 1.$$

Hence, the implicit Euler method is well-defined for all $h > 0$ and y_ℓ decays exponentially for $\ell \to \infty$. This correlates with the exponential decay of the correct solution $y(t)$ for $t \to \infty$. The explicit Euler method is well-defined for all $h > 0$, too. But only for sufficiently small h the discrete solution y_ℓ will emulate the exponential decay. So the implicit Euler method seems to be better suited for this specific problem than the explicit Euler method.

3.2. Fundamentals

This section aims to briefly transfer concepts and results from *explicit* one-step methods to *implicit* one-step methods.

We recall Definition 2.3: For a given **incremental function** $\Phi : [t_0, T] \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}_{>0} \to \mathbb{R}^n$, the inductive procedure

$$y_{\ell+1} := y_\ell + h_\ell \, \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \quad \text{for all } \ell = 0, \dots, N-1 \tag{3.6}$$

is called **implicit one-step method**.

Exercise 3.1. Suppose that the incremental function $\Phi(t, y, z, h)$ is Lipschitz continuous in z, i.e.,

$$\begin{aligned} &\exists\, L > 0 \;\forall\, t \in [t_0, T] \;\forall\, y \in \mathbb{R}^n \;\forall\, z, \widetilde{z} \in \mathbb{R}^n \;\forall\, h \in (0, T-t) : \\ &\qquad\qquad \|\Phi(t, y, z, h) - \Phi(t, y, \widetilde{z}, h)\| \le L \, \|z - \widetilde{z}\|. \end{aligned} \tag{3.7}$$

Then, it holds that

$$\begin{aligned} &\exists\, H > 0 \;\forall\, t \in [t_0, T) \;\forall\, y \in \mathbb{R}^n \;\forall\, 0 < h \le \min\{T-t, H\} \;\exists!\, z \in \mathbb{R}^n : \\ &\qquad\qquad z = y + h \, \Phi(t, y, z, h). \end{aligned} \tag{3.8}$$

In particular, for sufficiently small $h_\Delta > 0$, the implicit one-step method (3.6) is well-defined.

Remark 3.2. (i) While the proof of the last exercise is based on the Banach fixpoint theorem, one rather uses the Newton method than the Banach fixpoint iteration to compute $y_{\ell+1}$ from (3.6) in practice. Usually, the fixpoint iteration requires the restrictive condition $h_\Delta L < 1$, while the Newton method converges (in practice!) under much weaker conditions.

(ii) If $\Phi(t, y, z, h) = M(t, y, h) z + b(t, y, h)$ is affine in z, then one has to solve only one linear system

$$\big(I - h_\ell \, M(t_\ell, y_\ell, h_\ell)\big) y_{\ell+1} = y_\ell + h_\ell \, b(t_\ell, y_\ell, h_\ell)$$

to compute the solution $y_{\ell+1}$ of (3.6).

Definition 3.3. The implicit one-step method corresponding to the incremental function $\Phi(t, y, z, h)$ has consistency order $p \ge 1$, if the following is satisfied: Given $f \in C^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$, the exact solution $y \in C^{p+1}([t_0, T]; \mathbb{R}^n)$ of (3.2) satisfies that

$$\begin{aligned} &\exists\, C > 0 \,\forall\, t \in [t_0, T) \,\forall\, h \in (0, T-t] : \\ &\qquad\qquad \big\|y(t+h) - \big[y(t) + h \, \Phi(t, y(t), y(t+h), h)\big]\big\| \le C \, h^{p+1}. \end{aligned} \tag{3.9}$$

Example 3.4 (Implicit Euler method). Recall that

$$y_{\ell+1} := y_\ell + h_\ell \, f(t_{\ell+1}, y_{\ell+1}), \quad \text{i.e.,} \quad \Phi(t, y, z, h) := f(t+h, z).$$

According to the Taylor theorem for $t = (t+h) - h$, it holds that

$$y(t) = y(t+h) - h \, y'(t+h) + \mathcal{O}(h^2).$$

Hence, we see that

$$\begin{aligned} y(t+h) - \big[y(t) + h \, \Phi(t, y(t), y(t+h), h)\big] &= y(t+h) - \big[y(t) + h \, f(t+h, y(t+h))\big] \\ &= y(t+h) - \big[y(t) + h \, y'(t+h)\big] = \mathcal{O}(h^2). \end{aligned}$$

Therefore, the implicit Euler method has consistency order $p = 1$.

Example 3.5 (Implicit midpoint rule). The implicit midpoint rule is defined as

$$y_{\ell+1} := y_\ell + h_\ell\, f\Big(t_\ell + \frac{h_\ell}{2}, \frac{y_\ell + y_{\ell+1}}{2}\Big), \quad \text{i.e.,} \quad \Phi(t,y,z,h) := f\Big(t + \frac{h}{2}, \frac{y+z}{2}\Big).$$

According to the Taylor theorem for $(t + h/2) \pm h/2$, it holds that

$$\begin{aligned} y(t) &= y(t+h/2) - \frac{h}{2}\, y'(t+h/2) + \frac{h^2}{8}\, y''(t+h/2) + \mathcal{O}(h^3), \\ y(t+h) &= y(t+h/2) + \frac{h}{2}\, y'(t+h/2) + \frac{h^2}{8}\, y''(t+h/2) + \mathcal{O}(h^3). \end{aligned}$$

On the one hand, this shows that

$$y(t+h) - y(t) = h\, y'(t+h/2) + \mathcal{O}(h^3) = h\, f(t+h/2, y(t+h/2)) + \mathcal{O}(h^3).$$

On the other hand, this shows that

$$\frac{y(t+h) + y(t)}{2} = y(t+h/2) + \mathcal{O}(h^2)$$

and hence

$$f(t+h/2, y(t+h/2)) = f\Big(t + h/2, \frac{y(t+h)+y(t)}{2}\Big) + \mathcal{O}(h^2).$$

Overall, this results in

$$y(t+h) - \Big[y(t) + h\, f\Big(t + \frac{h}{2}, \frac{y(t+h)+y(t)}{2}\Big)\Big] = \mathcal{O}(h^3).$$

Therefore, the implicit midpoint rule has consistency order $p = 2$.

Exercise 3.6 (Trapezoidal rule). The trapezoidal rule is defined as

$$y_{\ell+1} := y_\ell + h_\ell\, \frac{f(t_\ell, y_\ell) + f(t_{\ell+1}, y_{\ell+1})}{2}, \quad \text{i.e.,} \quad \Phi(t,y,z,h) := \frac{f(t,y) + f(t+h,z)}{2}.$$

Show that the trapezoidal rule has consistency order $p = 2$.

Exercise 3.7 (Stability plus consistency implies convergence). Let $\Phi(t,y,z,h)$ be the incremental function of an implicit one-step method with consistency order $p \geq 1$. Suppose that $\Phi(t,y,z,h)$ is stable, i.e.,

$$\begin{aligned} &\exists\, L > 0\ \forall\, t \in [t_0, T)\ \forall\, y, \widetilde{y} \in \mathbb{R}^n\ \forall\, z, \widetilde{z} \in \mathbb{R}^n\ \forall\, h \in (0, T-t] : \\ &\qquad \|\Phi(t,y,z,h) - \Phi(t,\widetilde{y},\widetilde{z},h)\| \leq L\,\big(\|y - \widetilde{y}\| + \|z - \widetilde{z}\|\big). \end{aligned} \tag{3.10}$$

Then, the solution $y \in C^{p+1}([t_0,T];\mathbb{R}^n)$ satisfies that

$$\max_{\ell=1,\dots,N} \|y(t_\ell) - y_\ell\| = \mathcal{O}(h_\Delta^p), \tag{3.11}$$

whenever the discrete solutions exist (e.g., $h_\Delta \leq H$).

The following proposition shows that all implicit one-step methods are locally explicit. As far as, e.g., the adaptive step-size control is concerned, we can thus simply employ the ideas developed for explicit one-step methods (since our ideas above have built only on local Taylor expansions).

Proposition 3.8 (Implicit methods are locally explicit). *Let $\Phi(t,y,z,h)$ be the incremental function of an implicit one-step method with consistency order $p \geq 1$. Suppose that $\Phi(t,y,z,h)$ is Lipschitz continuous in z (see (3.7)) and continuously differentiable in (t,y,z). Let $h > 0$ and (t,y,z) with*

$$z = y + h\, \Phi(t,y,z,h) \quad \textit{and} \quad h\, \|D_z \Phi(t,y,z,h)\| < 1. \tag{3.12}$$

Then, there exist open sets $U \subset [0,T] \times \mathbb{R}^n$ and $V \subset \mathbb{R}^n$ with $(t,y) \in U$ and $z \in V$ as well as a function $g \in C^1(U;V)$ such that

$$\forall\,(\widetilde{t},\widetilde{y}) \in U \;\forall\,\widetilde{z} \in V: \quad \big(\widetilde{z} = \widetilde{y} + h\,\Phi(\widetilde{t},\widetilde{y},\widetilde{z},h) \quad \Longleftrightarrow \quad \widetilde{z} = \widetilde{g}(\widetilde{t},\widetilde{y})\big). \tag{3.13}$$

In particular, one step of the one-step method (3.6) *with step-size h is well-defined and even explicit, since $\Phi\big(t,y,z,h\big) = \Phi\big(t,y,g(t,y),h\big)$.*

Proof. Consider $F\big((\widetilde{t},\widetilde{y}),\widetilde{z}\big) := \widetilde{z} - \big[\widetilde{q} + h\,\Phi(\widetilde{t},\widetilde{y},\widetilde{z},h)\big]$. Then, $F \in C^1([t_0,T] \times \mathbb{R}^n \times \mathbb{R}^n;\mathbb{R}^n)$ and $F\big((t,y),z\big) = 0$ by assumption. Moreover, $\kappa := h\,\|D_z\Phi(t,y,z,h)\| < 1$ allows to employ the Neumann series to see that

$$D_zF\big((t,y),z\big) = I - h\,D_z\Phi(t,y,z,h)$$

is regular with

$$\big[D_zF\big((t,y),z\big)\big]^{-1} = \sum_{k=0}^{\infty} \big(h\,D_z\Phi(t,y,z,h)\big)^k.$$

Therefore, the claim follows from the implicit function theorem. □

3.3. Implicit Runge–Kutta methods

Definition 3.9. Let $A \in \mathbb{R}^{m\times m}$, $b,c \in \mathbb{R}^m$ with $0 \le c_1 \le c_2 \le \cdots \le c_m \le 1$. Then, a one-step method with incremental function

$$\Phi(t,y,h) := \sum_{j=1}^{m} b_j k_j, \tag{3.14}$$

where the so-called **stages** satisfy the implicit conditions

$$k_j = f\Big(t + c_j h,\; y + h\sum_{\ell=1}^{m} A_{j\ell}k_\ell\Big) \quad \text{for all } j = 1,\dots,m, \tag{3.15}$$

is called ***m*-stage Runge–Kutta method**. A method is called **implicit *m*-stage Runge–Kutta method**, if the matrix A is *not* strictly lower triangular. Usually, Runge–Kutta methods are denoted by their **Butcher tableau** $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$.

We stress that well-posedness of an implicit Runge–Kutta method is not obvious, since the equations for the stages (3.15) are (nonlinearly) coupled and implicit. However, well-posedness is shown in the following proposition, if the step-size h is sufficiently small.

Proposition 3.10. *Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be an m-stage Runge–Kutta method. Then, there exists $H > 0$ such that the stages* (3.15) *are well-defined, i.e.,*

$$\forall\, t \in [t_0,T]\; \forall\, y \in \mathbb{R}^n\; \forall\, 0 < h < \min\{T-t,H\}\; \exists!\, k_1,\dots,k_m \in \mathbb{R}^n\; \forall\, j = 1,\dots,m: \\ k_j = f\Big(t + c_j h,\; y + h\sum_{\ell=1}^{m} A_{j\ell}k_\ell\Big) \tag{3.16}$$

Proof. We write $K := (k_1,\dots,k_m) \in \mathbb{R}^{n\times m}$. Consider

$$\Psi : \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}, \quad \big(\Psi(K)\big)_j := f\Big(t + c_j h,\; y + h\sum_{\ell=1}^{m} A_{j\ell}k_\ell\Big) \in \mathbb{R}^n \quad \text{for all } j = 1,\dots,m.$$

Note that the stage conditions (3.15) are equivalent to $\Psi(K) = K$. Hence, we only need to show that Ψ has a unique fixpoint. To this end, we aim to apply the Banach fixpoint theorem. Note that

$$\begin{aligned}
&\|\Psi(K) - \Psi(\widetilde{K})\|_\infty := \max_{j=1,\dots,m} \Big\| f\Big(t + c_j h,\, y + h \sum_{\ell=1}^{m} A_{j\ell} k_\ell\Big) - f\Big(t + c_j h,\, y + h \sum_{\ell=1}^{m} A_{j\ell} \widetilde{k}_\ell\Big) \Big\| \\
&\quad \le hL \max_{j=1,\dots,m} \Big\| \sum_{\ell=1}^{m} A_{j\ell}(k_\ell - \widetilde{k}_\ell) \Big\| \le hL \max_{j=1,\dots,m} \sum_{\ell=1}^{m} |A_{j\ell}| \max_{i=1,\dots,m} \|k_i - \widetilde{k}_i\| \\
&\quad = \Big(hL \max_{j=1,\dots,m} \sum_{\ell=1}^{m} |A_{j\ell}| \Big) \|K - \tilde{K}\|_\infty.
\end{aligned}$$

For sufficiently small $h > 0$, the mapping Ψ is hence a contraction. Therefore, the Banach fixpoint theorem concludes the proof. □

Example 3.11. We consider the Runge–Kutta method $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$. By definition, one step of this method leads to $y_{\ell+1} = y_\ell + h\, k_1$, where

$$k_1 = f(t_\ell + h_\ell,\, y_\ell + h_\ell k_1) = f(t_{\ell+1},\, y_\ell + h_\ell k_1) = f(t_{\ell+1}, y_{\ell+1}).$$

We hence obtain that

$$y_{\ell+1} = y_\ell + h\, f(t_{\ell+1}, y_{\ell+1}),$$

which is the implicit Euler method.

Example 3.12. We consider the Runge–Kutta method $\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$. Then,

$$\begin{aligned}
k_1 &= f(t_\ell, y_\ell), \\
k_2 &= f\Big(t_\ell + h_\ell,\, y_\ell + h_\ell \,\frac{k_1 + k_2}{2}\Big), \\
y_{\ell+1} &= y_\ell + h_\ell \,\frac{k_1 + k_2}{2}.
\end{aligned}$$

From the last equation, we obtain that $k_2 = f(t_{\ell+1}, y_{\ell+1})$. Overall, we see that

$$y_{\ell+1} = y_\ell + h_\ell \frac{f(t_\ell, y_\ell) + f(t_{\ell+1}, y_{\ell+1})}{2},$$

which is the implicit trapezoidal rule.

Example 3.13. We consider the Runge–Kutta method $\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$. Then,

$$\begin{aligned}
k_1 &= f\Big(t_\ell + \frac{h_\ell}{2},\, y_\ell + \frac{h_\ell}{2}\, k_1\Big), \\
y_{\ell+1} &= y_\ell + h_\ell\, k_1.
\end{aligned}$$

From the last equation, we obtain that $h_\ell\, k_1 = y_{\ell+1} - y_\ell$ and hence $k_1 = f(t_\ell + \frac{h_\ell}{2},\, \frac{y_\ell + y_{\ell+1}}{2})$. Overall, we see that

$$y_{\ell+1} = y_\ell + h_\ell\, f\Big(t_\ell + \frac{h_\ell}{2},\, \frac{y_\ell + y_{\ell+1}}{2}\Big),$$

which is the implicit midpoint rule.

Exercise 3.14. Consider a general m-stage Runge–Kutta method $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$. Note that the corresponding stages $k_j = k_j(t, y, h)$ depend on (t, y, h) and that $\Phi(t, y, h) = \sum_{j=1}^m b_j k_j$. By exploiting the Lipschitz continuity of f in y, show that the Runge–Kutta method is stable in the sense of

$$\begin{aligned} &\exists\, H > 0\, \exists\, C > 0\, \forall t \in [t_0, T]\, \forall y, \widetilde{y} \in \mathbb{R}^n\, \forall\, 0 < h < \min\{T - t, H\} : \\ &\qquad\qquad\qquad \|\Phi(t, y, h) - \Phi(t, \widetilde{y}, h)\| \le C\, \|y - \widetilde{y}\|. \end{aligned} \tag{3.17}$$

Using stability and consistency, formulate and prove a convergence theorem for general Runge–Kutta methods.

Remark 3.15. We have already proved that an m-stage Runge–Kutta method with consistency order $p \ge 1$ satisfies the following statements (i)–(iii):

(i) $\displaystyle\sum_{j=1}^m b_j = 1.$

(ii) $\displaystyle\sum_{j=1}^m b_j c_j^\ell = \frac{1}{\ell + 1}$ for all $\ell = 0, \dots, p - 1$.

(iii) $\displaystyle\sum_{j=1}^m b_j (A^\ell \mathbb{1})_j = \frac{1}{(\ell + 1)!}$ for all $\ell = 0, \dots, p - 1$, where $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^m$.

Proposition 3.16 (Consistency $\boldsymbol{\le 2m}$). *Consider a general m-stage Runge–Kutta method $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ and suppose consistency order $p \ge 1$. Then, it holds that $p \le 2m$. Moreover, if $p = 2m$, then the vectors $b, c \in \mathbb{R}^m$ are unique: The coefficients of c are the nodes of the Gaussian quadrature rule on $[0, 1]$ and that of b are the corresponding weights.*

Proof. Consistency order p yields that

$$\sum_{j=1}^m b_j c_j^\ell \overset{!}{=} \frac{1}{\ell + 1} = \int_0^1 t^\ell \,\mathrm{d}t \quad \text{for all } \ell = 0, \dots, p - 1.$$

Hence, this quadrature is exact for polynomials of degree $p - 1$, i.e.,

$$\sum_{j=1}^m b_j q(c_j) = \int_0^1 q(t)\,\mathrm{d}t \quad \text{for all } q \in \mathbb{P}_{p-1}.$$

The claim follows from the basic lecture on numerical analysis: First, the maximum exactness for a quadrature rule with m nodes is $2m - 1$ and hence $p \le 2m$. Second, the Gaussian quadrature is the unique quadrature rule with exactness $2m - 1$. □

Remark 3.17. Runge–Kutta methods can be interpreted as the application of appropriate quadrature schemes employed to the integral representation of the exact solution. A simple substitution shows that

$$\begin{aligned} y(t + h) - y(t) &= \int_t^{t+h} y'(s)\,\mathrm{d}s = h \int_0^1 y'(t + sh)\,\mathrm{d}s = h \int_0^1 f(t + sh, y(t + sh))\,\mathrm{d}s \\ &\approx h \sum_{j=1}^m b_j f\big(t + c_j h, y(t + c_j h)\big) \approx h \sum_{j=1}^m b_j k_j. \end{aligned}$$

To derive a formula for $k_j \approx f(t+c_jh, y(t+c_jh))$, we proceed analogously:

$$\begin{aligned} y(t+c_jh)-y(t) &= \int_t^{t+c_jh} y'(s)\,\mathrm{d}s = h\int_0^{c_j} y'(t+sh)\,\mathrm{d}s = h\int_0^{c_j} f(t+sh, y(t+sh))\,\mathrm{d}s \\ &\approx h\sum_{\ell=1}^m A_{j\ell} f\big(t+c_\ell h, y(t+c_\ell h)\big) \approx h\sum_{\ell=1}^m A_{j\ell}k_\ell. \end{aligned}$$

With the foregoing interpretation, we will formulate and prove a sufficient criterion for the consistency of Runge–Kutta methods in terms of quadrature formulas. For this proof, we need two Lemmata on interpolation and quadrature, which will be shown first.

Lemma 3.18 (Lagrange interpolation). *Let $g \in C^s[a,b]$ and $a \le t_1 < \dots < t_s \le b$. Then, there exists a unique polynomial*

$$q \in \mathbb{P}_{s-1} \quad \textit{such that} \quad q(t_j) = g(t_j) \quad \textit{for all } j = 1,\dots,s, \tag{3.18}$$

and it holds that

$$q(t) = \sum_{j=1}^s g(t_j)L_j(t) \quad \textit{with the } \textbf{\textit{Lagrange polynomials}} \quad L_j(t) := \prod_{\substack{k=1\\k\neq j}}^s \frac{t-t_k}{t_j-t_k}. \tag{3.19}$$

Moreover, for all $k = 0,\dots,s-1$ and all $t \in [a,b]$, there holds the error identity

$$g^{(k)}(t) - q^{(k)}(t) = \frac{g^{(s)}(\xi)}{(s-k)!}\prod_{\ell=1}^{s-k}(t-\zeta_\ell) \tag{3.20}$$

with appropriate scalars $\xi = \xi(k,t), \zeta_\ell = \zeta_\ell(k) \in [a,b]$. In particular, it follows that

$$\|g^{(k)} - q^{(k)}\|_{\infty,[a,b]} \le \frac{\|g^{(s)}\|_{\infty,[a,b]}}{(s-k)!}\,|b-a|^{s-k}. \tag{3.21}$$

Proof. The proof is split into four steps.

Step 1. Consider the operator

$$\mathrm{T} : \mathbb{P}_{s-1} \to \mathbb{R}^s, \quad \mathrm{T}p := \big(p(t_1),\dots,p(t_s)\big).$$

Clearly, T is linear and $\dim \mathbb{P}_{s-1} = s$. Hence, T is bijective if and only if it is surjective (or injective). Given $z \in \mathbb{R}^s$, define $p := \sum_{j=1}^s z_jL_j$. Note that $L_j(t_j) = 1$, while $L_j(t_k) = 0$ for all $j,k = 1,\dots,n$ with $j \neq k$. Therefore, we obtain that $\mathrm{T}p = z$, i.e., T is surjective and hence bijective. In explicit terms, we have thus shown that $p = \sum_{j=1}^s z_jL_j$ is the unique polynomial in $\mathbb{P}_{s-1}$ such that

$$\forall j = 1,\dots,s: \quad p(t_j) = z_j.$$

Step 2. The error $e := g - q \in C^s[a,b]$ has at least s pairwise different zeros in $[a,b]$ (at the t_j). According to the mean value theorem, between two zeros of e, one has one zero of e'. Hence, $e' \in C^{s-1}[a,b]$ has at least $s-1$ zeros $\zeta_1^{(k)},\dots\zeta_{s-1}^{(k)}$. Inductively, we obtain that $e^{(k)} \in C^{s-k}[a,b]$ has at least $s-k$ zeros $\zeta_1^{(k)},\dots,\zeta_{s-k}^{(k)} \in [a,b]$.

Step 3. For $t = \zeta_\ell$, the error identity (3.20) is trivial. Without loss of generality, we can thus assume that $t \notin \{\zeta_1^{(k)},\dots,\zeta_{s-k}^{(k)}\}$. We consider the function

$$G(x) := e^{(k)}(t)\,\omega(x) - \omega(t)\,e^{(k)}(x), \quad \text{where} \quad \omega(x) := \prod_{\ell=1}^{s-k}(x-\zeta_\ell^{(k)}).$$

Clearly, $G \in C^{s-k}[a,b]$ has at least $s-k+1$ zeros in $[a,b]$. Inductively, the mean value theorem shows that $G^{(s-k)}$ has at least one zero $\xi \in [a,b]$. Hence,

$$0 = G^{(s-k)}(\xi) = e^{(k)}(t)\,\omega^{(s-k)}(\xi) - \omega(t)\,e^{(s)}(\xi) = e^{(k)}(t)\,(s-k)! - \omega(t)\,g^{(s)}(\xi).$$

Rearranging this estimate, we prove the error identity (3.20).

Step 4. The final estimate (3.21) follows from $|t - \zeta_\ell^{(k)}| \le |b-a|$ and hence

$$|g^{(k)}(t) - q^{(k)}(t)| \overset{(3.20)}{\le} \frac{\|g^{(s)}\|_{\infty,[a,b]}}{(s-k)!}\,|b-a|^{s-k}.$$

Taking the supremum over all $t \in [a,b]$, we conclude the proof. □

Lemma 3.19 (Interpolatory quadrature). *Let $g \in C^s[a,b]$ and $a \le t_1 < \cdots < t_s \le b$. Then,*

$$\Big|\int_a^b g\,\mathrm{d}t - \sum_{j=1}^s g(t_j)\int_a^b L_j\,\mathrm{d}t\Big| \le \frac{\|g^{(s)}\|_{\infty,[a,b]}}{s!}\,|b-a|^{s+1}. \tag{3.22}$$

Proof. Let $q := \sum_{j=1}^s g(t_j)L_j \in \mathbb{P}_{s-1}$ and recall the Lagrange interpolation (3.18)–(3.19). Then,

$$\sum_{j=1}^s g(t_j)\int_a^b L_j\,\mathrm{d}t = \int_a^b q\,\mathrm{d}t$$

and hence

$$\Big|\int_a^b g\,\mathrm{d}t - \sum_{j=1}^s g(t_j)\int_a^b L_j\,\mathrm{d}t\Big| = \Big|\int_a^b (g-q)\,\mathrm{d}t\Big| \overset{(3.21)}{\le} \frac{\|g^{(s)}\|_{\infty,[a,b]}}{s!}\,|b-a|^{s+1}.$$

This concludes the proof. □

Proposition 3.20 (Consistency in terms of quadrature rules). *Let $p \ge 1$. Consider a general Runge–Kutta method $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ and suppose that*

$$\sum_{j=1}^m b_j q(c_j) = \int_0^1 q\,\mathrm{d}t \quad \textit{for all } q \in \mathbb{P}_{p-1} \tag{3.23a}$$

as well as

$$\sum_{j=1}^m A_{\ell j} q(c_j) = \int_0^{c_\ell} q\,\mathrm{d}t \quad \textit{for all } q \in \mathbb{P}_{p-2} \textit{ and all } \ell = 1,\dots,m, \tag{3.23b}$$

where $\mathbb{P}_s$ denotes the space of all polynomials of degree $\le s$. Then, the Runge–Kutta method has at least consistency order p.

Proof. The proof is split into two steps.

Step 1. First, note that

$$\begin{aligned} y(t+c_\ell h) - y(t) &= \int_t^{t+c_\ell h} y'(s)\,\mathrm{d}s = h\int_0^{c_\ell} y'(t+sh)\,\mathrm{d}s \\ &= h\int_0^{c_\ell} f\big(t+sh, y(t+sh)\big)\,\mathrm{d}s \overset{(3.23b)}{=} h\Big(\sum_{j=1}^m A_{\ell j} f\big(t+c_j h, y(t+c_j h)\big) + \mathcal{O}(h^{p-1})\Big). \end{aligned}$$

Recall that $k_j = f\big(t + c_j h, y(t) + h \sum_{i=1}^m A_{ji} k_i\big)$. Using the last identity, we get that

$$\begin{aligned}
r_\ell &:= \Big\| y(t + c_\ell h) - \Big[y(t) + h \sum_{j=1}^m A_{\ell j} k_j \Big] \Big\| \\
&= h \, \Big\| \sum_{j=1}^m A_{\ell j} \Big[f\big(t + c_j h, y(t + c_j h)\big) - f\big(t + c_j h, y(t) + h \sum_{i=1}^m A_{ji} k_i\big) \Big] \Big\| + \mathcal{O}(h^p) \\
&\le hL \sum_{j=1}^m |A_{\ell j}| \, \Big\| y(t + c_j h) - \Big[y(t) + h \sum_{i=1}^m A_{ji} k_i \Big] \Big\| + \mathcal{O}(h^p) \\
&\le hL \max_{i=1,\dots,m} \sum_{j=1}^m |A_{ij}| \, \|r_j\| + \mathcal{O}(h^p).
\end{aligned}$$

Note that the right-hand side is independent of ℓ. Therefore, we get that

$$R := \max_{\ell=1,\dots,m} r_\ell \le \Big(hL \max_{i=1,\dots,m} \sum_{j=1}^m |A_{ij}| \Big) R + \mathcal{O}(h^p). \tag{3.24}$$

For sufficiently small $h > 0$, we thus obtain that $R = \mathcal{O}(h^p)$.

Step 2. Similarly, we see that

$$\begin{aligned}
y(t+h) - y(t) &= \int_t^{t+h} y'(s) \, \mathrm{d}s = h \int_0^1 y'(t + sh) \, \mathrm{d}s \\
&= h \int_0^1 f\big(t + sh, y(t + sh)\big) \, \mathrm{d}s \overset{(3.23a)}{=} h \Big(\sum_{j=1}^m b_j f\big(t + c_j h, y(t + c_j h)\big) + \mathcal{O}(h^p) \Big).
\end{aligned}$$

Finally, we can estimate the consistency error by

$$\begin{aligned}
&\big\| y(t+h) - \big[y(t) + h \, \Phi(t, y(t), h) \big] \big\| = \Big\| y(t+h) - \Big[y(t) + h \sum_{j=1}^m b_j k_j \Big] \Big\| \\
&= h \, \Big\| \sum_{j=1}^m b_j \, \Big[f\big(t + c_j h, y(t + c_j h)\big) - f\big(t + c_j h, y(t) + h \sum_{i=1}^m A_{ji} k_i\big) \Big] \Big\| + \mathcal{O}(h^{p+1}) \\
&\le hL \sum_{j=1}^m |b_j| \, \Big\| y(t + c_j h) - \Big[y(t) + h \sum_{i=1}^m A_{ji} k_i \Big] \Big\| + \mathcal{O}(h^{p+1}) \\
&\le hL \sum_{j=1}^m |b_j| \, R + \mathcal{O}(h^{p+1}) \overset{(3.24)}{=} \mathcal{O}(h^{p+1}).
\end{aligned}$$

This concludes the proof. □

3.4. Collocation methods

Definition 3.21. Let $0 \le c_1 < \cdots < c_m \le 1$. Then, the following inductive procedure is called **collocation method**: Given a time-step t_ℓ and the corresponding approximation $y_\ell \in \mathbb{R}^n$, let $q_\ell \in \mathbb{P}_m$ satisfy

$$q_\ell(t_\ell) = y_\ell \quad \text{and} \quad q_\ell'(t_\ell + c_j h_\ell) = f\big(t_\ell + c_j h_\ell, \, q_\ell(t_\ell + c_j h_\ell)\big) \quad \text{for all } j = 1, \dots, m. \tag{3.25}$$

Then define $y_{\ell+1} := q_\ell(t_{\ell+1})$.

Note that a collocation method is a very natural strategy to solve (3.2). A collocation method provides a continuous piecewise polynomial, which satisfies the given ODE pointwise at finitely many collocation nodes $t_\ell + c_j h_\ell$. We stress that well-posedness of a collocation method is not obvious, since the interpolation conditions for q_ℓ' in (3.25) are implicit and (thus possibly) nonlinear. However, we will show in Theorem 3.23 that collocation methods are (implicit) Runge–Kutta methods. Hence, well-posedness can be derived from the well-posedness of implicit Runge–Kutta methods, see Prop. 3.10.

EXAMPLE 3.22 (Collocation methods with $\boldsymbol{m = 1}$). Let $m = 1$. Then, the collocation polynomial q_ℓ from (3.25) is linear and hence its derivative is constant $q_\ell'(t) = \frac{y_{\ell+1}-y_\ell}{h_\ell}$.

- If $c_1 = 0$, then $\frac{y_{\ell+1}-y_\ell}{h_\ell} = q_\ell'(t_\ell) = f(t_\ell, q_\ell(t_\ell)) = f(t_\ell, y_\ell)$. Hence, we obtain the explicit Euler method.
- If $c_1 = 1$, then $\frac{y_{\ell+1}-y_\ell}{h_\ell} = q_\ell'(t_{\ell+1}) = f(t_{\ell+1}, q_\ell(t_{\ell+1})) = f(t_{\ell+1}, y_{\ell+1})$. Hence, we obtain the implicit Euler method.
- If $c = 1/2$, then $t_\ell + \frac{h_\ell}{2} = \frac{t_\ell + t_{\ell+1}}{2}$ and $q_\ell(t_\ell + \frac{h_\ell}{2}) = \frac{q_\ell(t_\ell)+q_\ell(t_{\ell+1})}{2} = \frac{y_\ell + y_{\ell+1}}{2}$. Hence,
$$\frac{y_{\ell+1}-y_\ell}{h_\ell} = q_\ell'\Big(\frac{t_\ell + t_{\ell+1}}{2}\Big) = f\Big(t_\ell + \frac{h_\ell}{2},\, q_\ell\Big(t_\ell + \frac{h_\ell}{2}\Big)\Big) = f\Big(\frac{t_\ell + t_{\ell+1}}{2},\, \frac{y_\ell + y_{\ell+1}}{2}\Big),$$
and we obtain the implicit midpoint scheme.

THEOREM 3.23 (Collocation methods are Runge–Kutta methods). *For any nodes $0 \le c_1 < \dots < c_m \le 1$, the corresponding collocation scheme is an (possibly implicit) m-stage Runge–Kutta method with consistency order $p \ge m$, where*
$$A_{ij} := \int_0^{c_i} L_j \,\mathrm{d}t \quad \textit{and} \quad b_j := \int_0^1 L_j \,\mathrm{d}t \quad \textit{for all } i,j = 1,\dots,m \tag{3.26}$$
with the Lagrange polynomials $L_j(t) := \prod_{\substack{k=1\\k\neq j}}^m \frac{t-c_k}{c_j-c_k}$.

PROOF. The proof is split into two steps.

Step 1. Let $q_\ell \in \mathbb{P}_m$ be the collocation polynomial (3.25). Then, $q_\ell' \in \mathbb{P}_{m-1}$ and hence
$$q_\ell'(t + s h_\ell) \overset{(3.19)}{=} \sum_{j=1}^m f\big(t + c_j h,\, q_\ell(t + c_j h)\big)\, L_j(s).$$
Therefore, we see that
$$\begin{aligned} q_\ell(t_\ell + c_i h_\ell) &= q_\ell(t_\ell) + \int_{t_\ell}^{t_\ell + c_i h_\ell} q_\ell'(t)\,\mathrm{d}t = y_\ell + h_\ell \int_0^{c_i} q_\ell'(t_\ell + s h_\ell)\,\mathrm{d}s \\ &= y_\ell + h_\ell \sum_{j=1}^m f\big(t + c_j h,\, q_\ell(t + c_j h)\big) \int_0^{c_i} L_j(s)\,\mathrm{d}s \\ &\overset{(3.26)}{=} y_\ell + h_\ell \sum_{j=1}^m A_{ij} f\big(t + c_j h,\, q_\ell(t + c_j h)\big). \end{aligned}$$
With $k_i := f\big(t + c_i h,\, q_\ell(t + c_i h)\big)$, the last identity proves that
$$k_i = f\big(t_\ell + c_i h_\ell,\, q_\ell(t_\ell + c_i h_\ell)\big) = f\Big(t_\ell + c_i h_\ell,\, y_\ell + h_\ell \sum_{j=1}^m A_{ij} k_j\Big).$$

The same argument proves that

$$\begin{aligned} y_{\ell+1} = q_\ell(t_{\ell+1}) = q_\ell(t_\ell) + \int_{t_\ell}^{t_\ell+h_\ell} q_\ell'(t)\,\mathrm{d}t = y_\ell + h_\ell \int_0^1 q_\ell'(t_\ell + sh_\ell)\,\mathrm{d}s \\ = y_\ell + h_\ell \sum_{j=1}^m f\big(t + c_j h,\, q_\ell(t + c_j h)\big) \int_0^1 L_j(s)\,\mathrm{d}s \overset{(3.26)}{=} y_\ell + h_\ell \sum_{j=1}^m b_j k_j. \end{aligned}$$

This shows that the collocation method is indeed a Runge–Kutta method.

Step 2. For each polynomial $q \in \mathbb{P}_{m-1}$, it holds that

$$q(s) \overset{(3.19)}{=} \sum_{j=1}^m q(c_j) L_j(s).$$

Hence, it follows that

$$\sum_{j=1}^m b_j q(c_j) \overset{(3.26)}{=} \sum_{j=1}^m q(c_j) \int_0^1 L_j\,\mathrm{d}t = \int_0^1 q\,\mathrm{d}t$$

as well as

$$\sum_{j=1}^m A_{ij} q(c_j) \overset{(3.26)}{=} \sum_{j=1}^m q(c_j) \int_0^{c_i} L_j\,\mathrm{d}t = \int_0^{c_i} q\,\mathrm{d}t.$$

Therefore, Proposition 3.20 proves that the consistency order is at least m. □

REMARK 3.24. Since Runge-Kutta methods are stable, collocation methods with m pairwise different nodes converge with order $p \geq m$, i.e.

$$\max_{\ell=1,\dots,N} \|y(t_\ell) - y_\ell\| = \mathcal{O}(h_\Delta^m),$$

whenever the discrete solutions exist (e.g., $h_\Delta \leq H$).

We will extend this result in two directions. We will show in Theorem 3.28 that collocation methods not only provide approximations to $y(t_\ell)$, but to the function y and its derivatives in the whole intervall. Moreover, in Theorem 3.30 we will show, that a collocation method has convergence order p if and only if the underlying quadrature formula $\sum_{j=1}^m b_j q(c_j) \approx \int_0^1 q(\tau)\,\mathrm{d}\tau$ is exact for polynomials $q \in \mathbb{P}_{p-1}$. Hence, the maximal order of convergence will be $p = 2m$ (Gauss methods).

COROLLARY 3.25. *The matrix A of a Butcher Scheme of a collocation method is regular if and only if for all $j = 1, \dots, m$ the nodes c_j belong to $(0, 1]$ and are piecewise different.*

PROOF. Chosing the polynomials $q \in \mathbb{P}_{m-1}$ in the second step of the last proof as monomials $t \mapsto t^k$ with $k = 0, \dots, m-1$, the last equation in this proof leads for all $i = 1, \dots, m$ and $k = 0, \dots, m-1$ to the matrix equation $AV = \tilde{V}$ with the matrices

$$V := \begin{pmatrix} c_1^0 & c_1^1 & \cdots & c_1^{m-1} \\ c_2^0 & c_2^1 & \cdots & c_2^{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_m^0 & c_m^1 & \cdots & c_m^{m-1} \end{pmatrix} \quad \text{and } \tilde{V} := \begin{pmatrix} c_1^1 & c_1^2/2 & \cdots & c_1^m/m \\ c_2^1 & c_2^2/2 & \cdots & c_2^m/m \\ \vdots & \vdots & \ddots & \vdots \\ c_m^1 & c_m^2/2 & \cdots & c_m^m/m \end{pmatrix}.$$

V is the Vandermonde matrix, which is regular for piecewise different nodes. The claim follows by

$$\det \tilde{V} = \left(\frac{1}{m!} \prod_{j=1}^{m} c_j \right) \det V.$$

□

One major advantage of collocation methods is that they provide an approximation of the sought solution y in the full time interval $[t_0, T]$ (and not only at the time-steps t_j). Moreover, we even get natural approximations of the derivatives of y. Before we state and prove the main convergence theorem of collocation methods, we need the following lemma on polynomial approximations.

LEMMA 3.26 (Polynomial approximation). *Let $z \in C^{r+1}[t, t+h]$, $q \in \mathbb{P}_r$, and $k \in \{1, \dots, r\}$. Then, it holds that*

$$\|z^{(k)} - q^{(k)}\|_{\infty,[t,t+h]} \le C \left(h^{r+1-k} \, \|z^{(r+1)}\|_{\infty,[t,t+h]} + h^{-k} \, \|z - q\|_{\infty,[t,t+h]} \right), \tag{3.27}$$

where $C > 0$ depends only on r, but is independent of z, q, and h.

PROOF. The proof follows from a so-called scaling argument and is split in three steps.

Step 1. Define $\widehat{z}(s) := z(t+sh)$ and $\widehat{q}(s) := q(t+sh)$. Note that, e.g., $\widehat{z}^{(k)}(s) = h^k \, z^{(k)}(t+sh)$. In particular, one sees that

$$\begin{aligned} \|\widehat{z}^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} &= h^k \, \|z^{(k)} - q^{(k)}\|_{\infty,[t,t+h]}, \\ \|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} &= h^{r+1} \, \|z^{(r+1)}\|_{\infty,[t,t+h]}. \end{aligned}$$

Step 2. Let $0 \le s_0 < \dots < s_r \le 1$ be interpolation nodes. Define

$$(I\widehat{z})(s) := \sum_{j=0}^{r} \widehat{z}(s_j) L_j(s), \quad \text{where} \quad L_j(s) := \prod_{\substack{k=0 \\ k \ne j}}^{r} \frac{s - s_k}{s_j - s_k}.$$

The error estimate (3.21) for the Lagrange interpolation yields that

$$\begin{aligned} \|\widehat{z} - I\widehat{z}\|_{\infty,[0,1]} &\le \frac{1}{(r+1)!} \, \|\widehat{z}^{(r+1)}\|_{\infty,[0,1]}, \\ \|\widehat{z}^{(k)} - (I\widehat{z})^{(k)}\|_{\infty,[0,1]} &\le \frac{1}{(r+1-k)!} \, \|\widehat{z}^{(r+1)}\|_{\infty,[0,1]}. \end{aligned}$$

Moreover, norm equivalence on the finite-dimensional space $\mathbb{P}_r$ provides a constant $C_k > 0$ such that

$$\begin{aligned} \|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} &\le \|I\widehat{z} - \widehat{q}\|_{\infty,[0,1]} + \|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \\ &\le C_k \max_{j=0,\dots,r} \big| (I\widehat{z})(s_j) - \widehat{q}(s_j) \big|, \end{aligned}$$

where we note that $\|p\| := \max_{j=0,\dots,r} \big|p(z_j)\big|$ is a norm on $\mathbb{P}_r$ due to Lemma 3.18. Since $(I\widehat{z})(s_j) = z(s_j)$ for all $j = 0, \dots, r$, we obtain that

$$\|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \le C_k \, \|\widehat{z} - \widehat{q}\|_{\infty,[0,1]}.$$

With the triangle inequality and $C := \max\limits_{k=1,\dots,r} \max\{C_k,\, 1/(r+1-k)!\}$, we conclude that

$$\begin{aligned}\|\widehat{z}^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} &\le \|z^{(k)} - (I\widehat{z})^{(k)}\|_{\infty,[0,1]} + \|(I\widehat{z})^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \\ &\le \frac{1}{(r+1-k)!}\,\|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} + C_k\,\|\widehat{z} - \widehat{q}\|_{\infty,[0,1]} \\ &\le C\,\big(\|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} + \|\widehat{z} - \widehat{q}\|_{\infty,[0,1]}\big).\end{aligned}$$

We stress that this already proves the claim (3.27) on the reference interval $[0,1]$.

Step 3. It holds that

$$\begin{aligned}\|z^{(k)} - q^{(k)}\|_{\infty,[t,t+h]} &= h^{-k}\,\|\widehat{z}^{(k)} - \widehat{q}^{(k)}\|_{\infty,[0,1]} \\ &\le C\,h^{-k}\,\big(\|\widehat{z}^{(r+1)}\|_{\infty,[0,1]} + \|\widehat{z} - \widehat{q}\|_{\infty,[0,1]}\big) \\ &\le C\,h^{-k}\,\big(h^{r+1}\,\|z^{(r+1)}\|_{\infty,[t,t+h]} + \|z - q\|_{\infty,[t,t+h]}\big).\end{aligned}$$

This concludes the proof. We stress that the norm equivalence argument has to be used on $[0,1]$ instead of $[t,t+h]$ to ensure that C is independent of h. □

THEOREM 3.27 (Global convergence of collocation methods). *Let $0 \le c_1 < \dots < c_m \le 1$ be given nodes of a collocation method. Define the spline $q : [t_0,T] \to \mathbb{R}^n$ by $q|_{[t_\ell,t_{\ell+1}]} := q_\ell$, where $q_\ell \in \mathbb{P}_m$ are the collocation polynomials* (3.25) *for all $\ell = 0,\dots,N-1$. Suppose that $f \in C^m([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$. Then, the solution $y \in C^{m+1}([t_0,T];\mathbb{R}^n)$ of* (3.2) *satisfies that*

$$\|y^{(k)} - q^{(k)}\|_{\infty,[t_0,T]} = \mathcal{O}(h_\Delta^{m-k}) \quad \textit{for all } 0 \le k \le m, \tag{3.28}$$

where $q^{(k)}$ is understood elementwise on $[t_\ell,t_{\ell+1}]$ for all $\ell = 0,\dots,N-1$.

PROOF. Note that

$$\|g\|_{\infty,[t_0,T]} = \max_{\ell=0,\dots,N-1} \|g\|_{\infty,[t_\ell,t_{\ell+1}]}.$$

Therefore, it suffices to prove (3.28) for one time interval $[t_\ell,t_{\ell+1}]$. Moreover, recall that $q|_{[t_\ell,t_{\ell+1}]} = q_\ell \in \mathbb{P}_m$. Therefore, Lemma 3.26 proves that

$$\|y^{(k)} - q^{(k)}\|_{\infty,[t_\ell,t_{\ell+1}]} \le C\,\big(h_\ell^{m+1-k}\,\|y^{(m+1)}\|_{\infty,[t_\ell,t_{\ell+1}]} + h_\ell^{-k}\,\|y - q\|_{\infty,[t_\ell,t_{\ell+1}]}\big).$$

Hence, it suffices to consider the case $k = 0$.

Step 1. Let $t \in [t_0,T)$ and $h > 0$. Let $Iy \in \mathbb{P}_m$ be the (unique) polynomial such that

$$Iy(t) = y(t) \quad\text{and}\quad \forall j = 1,\dots,m: \quad (Iy)'(t + c_j h) = y'(t + c_j h).$$

Let $0 \le \lambda \le 1$. With $L_j(s) = \prod\limits_{\substack{k=1\\k\ne j}}^{m} \dfrac{s - c_k}{c_j - c_k}$, Lemma 3.19 yields that

$$\begin{aligned}y(t+\lambda h) - Iy(t+\lambda h) &= \int_t^{t+\lambda h} (y - Iy)'(s)\,\mathrm{d}s = h\int_0^\lambda (y - Iy)'(t + sh)\,\mathrm{d}s \\ &\overset{(3.22)}{=} h\,\Big[\sum_{j=1}^m (y - Iy)'(t + c_j h)\int_0^\lambda L_j(s)\,\mathrm{d}s + \mathcal{O}(h^m)\Big] = \mathcal{O}(h^{m+1}).\end{aligned}$$

Since the right-hand side is independent of λ, this proves that

$$\|y - Iy\|_{\infty,[t,t+h]} = \mathcal{O}(h^{m+1}).$$

Step 2. Employ the notation $Iy \in \mathbb{P}_m$ from Step 1 for the interval $[t_\ell, t_\ell + h_\ell] = [t_\ell, t_{\ell+1}]$. The triangle inequality and Step 1 prove that

$$\begin{aligned} \|y - q\|_{\infty,[t_\ell,t_{\ell+1}]} &\leq \|y - Iy\|_{\infty,[t_\ell,t_{\ell+1}]} + \|Iy - q\|_{\infty,[t_\ell,t_{\ell+1}]} \\ &= \mathcal{O}(h_\ell^{m+1}) + \|Iy - q\|_{\infty,[t_\ell,t_{\ell+1}]}. \end{aligned} \tag{3.29}$$

Let $0 \leq \lambda \leq 1$. We argue as before:

$$\begin{aligned} Iy(t_\ell + \lambda h_\ell) - q(t_\ell + \lambda h_\ell) &= y(t_\ell) - q(t_\ell) + \int_{t_\ell}^{t_\ell + \lambda h_\ell} (Iy - q)'(s) \,\mathrm{d}s \\ &= y(t_\ell) - q(t_\ell) + h_\ell \int_0^\lambda (Iy - q)'(t + sh) \,\mathrm{d}s \\ &= y(t_\ell) - q(t_\ell) + h_\ell \sum_{j=1}^m (Iy - q)'(t_\ell + c_j h_\ell) \int_0^\lambda L_j(s) \,\mathrm{d}s. \end{aligned}$$

Recall that by Remark 3.24 we get convergence

$$y(t_\ell) - q(t_\ell) = y(t_\ell) - y_\ell = \mathcal{O}(h_\Delta^m).$$

Note that

$$(Iy - q)'(t_\ell + c_j h_\ell) = f\big(t_\ell + c_j h_\ell,\, y(t_\ell + c_j h_\ell)\big) - f\big(t_\ell + c_j h_\ell,\, q(t_\ell + c_j h_\ell)\big).$$

Combining the last three identities with the Lipschitz continuity of f, we obtain that

$$\begin{aligned} \|(Iy - q)(t_\ell + \lambda h_\ell)\| &\leq h_\ell L \sum_{j=1}^m \Big| \int_0^\lambda L_j(s) \,\mathrm{d}s \Big| \, \|(y - q)(t_\ell + c_j h_\ell)\| + \mathcal{O}(h_\Delta^m) \\ &\leq \left(h_\ell L \sum_{j=1}^m \int_0^1 |L_j(s)| \,\mathrm{d}s \right) \|y - q\|_{\infty,[t_\ell,t_{\ell+1}]} + \mathcal{O}(h_\Delta^m). \end{aligned}$$

Note that the right-hand side is independent of $0 \leq \lambda \leq 1$. Hence, we are led to

$$\|Iy - q\|_{\infty,[t_\ell,t_{\ell+1}]} \leq \left(h_\ell L \sum_{j=1}^m \int_0^1 |L_j(s)| \,\mathrm{d}s \right) \|y - q\|_{\infty,[t_\ell,t_{\ell+1}]} + \mathcal{O}(h_\Delta^m).$$

In combination with (3.29), we see that

$$\|y - q\|_{\infty,[t_\ell,t_{\ell+1}]} \leq \left(h_\ell L \sum_{j=1}^m \int_0^1 |L_j(s)| \,\mathrm{d}s \right) \|y - q\|_{\infty,[t_\ell,t_{\ell+1}]} + \mathcal{O}(h_\Delta^m) + \mathcal{O}(h_\ell^{m+1}).$$

For sufficiently small $h_\ell \leq h_\Delta$, we thus infer that

$$\|y - q\|_{\infty,[t_\ell,t_{\ell+1}]} = \mathcal{O}(h_\Delta^m) + \mathcal{O}(h_\Delta^{m+1}) = \mathcal{O}(h_\Delta^m).$$

□

Note, that we can improve the error bound to

$$\|y^{(k)} - q^{(k)}\|_{\infty,[t_0,T]} = \mathcal{O}(h_\Delta^{m+1-k}) \quad \text{for all } 0 \leq k \leq m,$$

if the collocation method has order $m+1$ and $f \in C^m([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$.

The next goal is the theorem that a collocation method has consistency order p if and only if the induced quadrature on $[0, 1]$ (see Proposition 3.16) is exact for all $q \in \mathbb{P}_{p-1}$. In particular, the m-stage Gaussian collocation method is the only collocation scheme with consistency order $p = 2m$.

Remark 3.28. Let $f \in C^p([t_0,T]\times\mathbb{R}^n;\mathbb{R}^n)$ and suppose that f is Lipschitz in y. Let $(t_0,y_0)\in\mathbb{R}\times\mathbb{R}^n$. Then,

$$Y(t,t_0,y_0) := y_0 + \int_{t_0}^{t} f\big(s, Y(s,t_0,y_0)\big)\,\mathrm{d}s \tag{3.30}$$

is the unique solution of

$$y'(t) = f\big(t,y(t)\big) \text{ in } [t_0,T] \quad \text{subject to} \quad y(t_0)=y_0.$$

Clearly, $y(t) = Y(t,t_0,y_0)$ satisfies $y \in C^{p+1}([t_0,T],\mathbb{R}^n)$. Moreover, one can show that $Y(t,t_0,y_0)$ is also C^p with respect to t_0 and y_0, and all derivatives depend only on the derivatives of f. We refer to [**Wal00**, Part III, Section §13, Subsection XI, Corollar].

Lemma 3.29 (Gröbner & Alekseev). *Let $f,\varepsilon \in C^1([t,t+h]\times\mathbb{R}^n;\mathbb{R}^n)$ be Lipschitz in y. Let $y,z\in C^2([t,t+h],\mathbb{R}^n)$ such that*

$$\begin{aligned} y(t) &= y_0, \quad y'(t) = f\big(t,y(t)\big),\\ z(t) &= y_0, \quad z'(t) = f\big(t,z(t)\big)+\varepsilon\big(t,z(t)\big). \end{aligned} \tag{3.31}$$

Then, it follows that

$$z(t+h)-y(t+h) = \int_t^{t+h} \partial_3 Y\big(t+h,s,z(s)\big)\varepsilon\big(s,z(s)\big)\,\mathrm{d}s \tag{3.32}$$

where $Y\big(t,s,z(s)\big)$ is given by (3.30)*.*

Proof. Let $N\in\mathbb{N}$. For $\ell=0,\dots,N$, let $t_\ell := t+\ell H$ with $H := h/N$. Note that $t_0 = t$ and $t_N = t+h$. The Taylor expansion shows that

$$\begin{aligned} d_i &:= z(t_{i+1}) - Y\big(t_{i+1},t_i,z(t_i)\big)\\ &= \big[\, z(t_i) + H\,\big\{f\big(t_i,z(t_i)\big)+\varepsilon\big(t_i,z(t_i)\big)\big\} + \mathcal{O}(H^2)\,\big] - \big[\, z(t_i) + H\,f\big(t_i,z(t_i)\big) + \mathcal{O}(H^2)\,\big]\\ &= H\,\varepsilon\big(t_i,z(t_i)\big) + \mathcal{O}(H^2)\\ &= H\,\varepsilon\big(t_{i+1},z(t_{i+1})\big) + \mathcal{O}(H^2). \end{aligned}$$

By definition of d_i, the uniqueness of ODE solutions proves that

$$Y\big(t+h,t_i,z(t_i)\big) = Y\big(t+h,t_{i+1},z(t_{i+1})-d_i\big).$$

Therefore, the Taylor expansion shows that

$$\begin{aligned} D_i &:= Y\big(t+h,t_{i+1},z(t_{i+1})\big) - Y\big(t+h,t_i,z(t_i)\big)\\ &= Y\big(t+h,t_{i+1},z(t_{i+1})\big) - Y\big(t+h,t_{i+1},z(t_{i+1})-d_i\big)\\ &= \partial_3 Y\big(t+h,t_{i+1},z(t_{i+1})\big)d_i + \mathcal{O}(|d_i|^2)\\ &= H\,\underbrace{\partial_3 Y\big(t+h,t_{i+1},z(t_{i+1})\big)\,\varepsilon\big(t_{i+1},z(t_{i+1})\big)}_{=:g(t_{i+1})} + \mathcal{O}(H^2). \end{aligned}$$

Note that $\mathcal{O}(H^2)=\mathcal{O}(N^{-2})$. Together with Riemann sum theory, it follows that

$$\sum_{i=0}^{N-1} D_i = \mathcal{O}(N^{-1}) + \sum_{i=0}^{N-1}\Big(\big(t_{i+1}-t_i\big)\,g(t_{i+1})\Big) \xrightarrow{N\to\infty} \int_t^{t+h} g(s)\,\mathrm{d}s,$$

since g is (at least) continuous. Moreover, the telescopic series proves that

$$\begin{aligned}\sum_{i=0}^{N-1} D_i &= Y\big(t+h, t_N, z(t_N)\big) - Y\big(t+h, t_0, z(t_0)\big)\\ &= Y\big(t+h, t+h, z(t+h)\big) - Y\big(t+h, t, y(t)\big) = z(t+h) - y(t+h).\end{aligned}$$

Combining the latter two identities, we prove (3.32). □

Theorem 3.30 (Quadrature vs. collocation). *Let $0 \le c_1 < \cdots < c_m \le 1$ and $b_j := \int_0^1 L_j \,\mathrm{d}t$, where $L_j \in \mathbb{P}_{m-1}$ are the Lagrange polynomials. Let $p \in \mathbb{N}$. Then, the following statements* (i)–(ii) *are equivalent:*

(i) $\displaystyle\sum_{j=1}^m b_j q(c_j) = \int_0^1 q \,\mathrm{d}t$ *for all $q \in \mathbb{P}_{p-1}$, i.e., the quadrature has exactness $p-1$.*

(ii) *The corresponding collocation method has consistency order p.*

In any case, it holds that $m \le p \le 2m$.

Proof. The implication (ii) $\Longrightarrow$ (i) as well as the bound $m \le p \le 2m$ are already known; see Proposition 3.16. It thus only remains to prove that (i) $\Longrightarrow$ (ii).

Step 1. Let $g \in C^p[t, t+h]$. By Lemma B.3 and assumption (i) it holds that

$$\Big| \int_t^{t+h} g \,\mathrm{d}s - h \sum_{j=1}^m b_j g(t + c_j h) \Big| \le \frac{\|g^{(p)}\|_{\infty,[t,t+h]}}{p!}\, h^{p+1}. \tag{3.33}$$

Step 2. Let $q \in \mathbb{P}_m$ be the collocation polynomial on $[t, t+h]$, i.e.,

$$q(t) = y(t) \quad \text{and} \quad q(t + c_j h) = f\big(t + c_j h,\, q(t + c_j h)\big) \quad \text{for all } j = 1, \dots, m.$$

We apply Lemma 3.29 for the collocation polynomial and

$$q'(t) = f\big(t, q(t)\big) + \big[q'(t) - f\big(t, q(t)\big)\big], \quad \text{i.e.,} \quad \varepsilon\big(t, q(t)\big) = q'(t) - f\big(t, q(t)\big).$$

This leads to

$$q(t+h) - y(t+h) = \int_t^{t+h} \underbrace{\partial_3 Y\big(t+h, s, q(s)\big)\big[q'(s) - f\big(s, q(s)\big)\big]}_{=:g(s)} \,\mathrm{d}s,$$

where we also define g for which we aim to apply Step 1. Note that $g(t + c_j h) = 0$ for all $j = 1, \dots, m$ by the collocation conditions. From Step 1, we thus infer that the consistency error satisfies

$$|y(t+h) - q(t+h)| = \Big| \int_t^{t+h} g \,\mathrm{d}s - \underbrace{h \sum_{j=1}^m b_j g(t + c_j h)}_{=0} \Big| \overset{(3.33)}{\le} \frac{\|g^{(p)}\|_{\infty,[t,t+h]}}{p!}\, h^{p+1}.$$

Step 3. It remains to argue that $\|g^{(p)}\|_{\infty,[t,t+h]}$ is uniformly bounded in terms of y and f. In particular, we have to find a bound which is independent of h. Note that

$$g(s) = \partial_3 Y\big(t+h, s, q(s)\big)\big[q'(s) - f\big(s, q(s)\big)\big],$$

where $Y\big(t+h, s, q(s)\big)$ solves

$$\partial_1 Y\big(t, s, q(s)\big) = f\big(t, Y(t, s, q(s))\big) \text{ in } [t, t+h] \quad \text{subject to} \quad Y\big(s, s, q(s)\big) = q(s).$$

According to the chain rule (and Remark 3.28), $g^{(p)}$ depends only on partial derivatives of f and derivatives of q. For $0 \leq k \leq m$, it holds that

$$\|q^{(k)}\|_{\infty,[t,t+h]} \leq \|y^{(k)} - q^{(k)}\|_{\infty,[t,t+h]} + \|y^{(k)}\|_{\infty,[t,t+h]} \overset{(3.28)}{=} \mathcal{O}(h^{m-k}) + \|y^{(k)}\|_{\infty,[t,t+h]} = \mathcal{O}(1).$$

For $k > m$, it holds that $q^{(k)} = 0$. Overall, we thus obtain that $\|g^{(p)}\|_{\infty,[t,t+h]} = \mathcal{O}(1)$. This concludes the proof. □

Joseph-Louis de Lagrange (1736–1813) *was an Italian mathematician (born in Turino as* Giuseppe Lodovico Lagrangia*). In 1755, he got the chair for mathematics at the military academy in Turino. In 1766, he became the successor of Euler at the Prussian Academy if Sciences in Berlin. In 1786, he moved to Paris to the French Academy of Sciences. From 1797, he was professor at the Ecole Polytechnique in Paris, where we was promoting the young Augustin-Louis Cauchy.*

CHAPTER 4

Stiff ODEs

4.1. Introduction

In 1952, it was observed by Curtiss and Hirschfelder that explicit methods fail for the numerical integration of certain ODEs which model chemical reactions. They said that the ODE is stiff, if certain components of the solution arrive in a very short time in their equilibrium (i.e., the fast reacting components), while other slowly changing components are more or less fixed (i.e., stiff).

Mathematically, this is, e.g., the case for

$$y'(t) = My(t) + f(t) \tag{4.1}$$

if the eigenvalues $\lambda_1, \dots, \lambda_n \in \mathbb{C}$ of $M \in \mathbb{R}^{n\times n}$, sorted by $\operatorname{Re}\lambda_1 \geq \cdots \geq \operatorname{Re}\lambda_n$, satisfy that

$$|\operatorname{Re}\lambda_1| \sim 1, \operatorname{Re}\lambda_1 \leq 0 \quad \text{but} \quad \operatorname{Re}\lambda_n \ll 0. \tag{4.2}$$

Example 4.1. Consider the symmetric matrix $M = \begin{pmatrix} a & b \\ b & a \end{pmatrix} \in \mathbb{R}^{2\times 2}$ with $a, b \in \mathbb{R}$. The characteristic polynomial is $p(\lambda) = \det(A - \lambda I) = (a-\lambda)^2 - b^2 = \lambda^2 - 2a\lambda + (a^2 - b^2)$. The eigenvalues of M are the zeros of p and hence M has the eigenvalues $\lambda = a \pm b$. Even for simple $a = -51$, $b = -50$, we get eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -101$, so that the corresponding ODE would be stiff.

Example 4.2 (Heat equation). The prototype of a parabolic partial differential equation is the heat equation. In one space dimension, we are looking for solutions u to

$$\partial_t u(x,t) = \partial_x^2 u(x,t) + f(x,t), \qquad t \in [0,T], \quad x \in [a,b]. \tag{4.3a}$$

In order to ensure uniqueness, we pose homogeneous Dirichlet boundary conditions

$$\forall t \in [0,T]: \quad u(a,t) = u(b,t) = 0 \tag{4.3b}$$

and an initial value function g with compact support in (a,b) such that

$$\forall x \in [a,b]: \quad u(x,0) = g(x). \tag{4.3c}$$

Charles Francis Curtiss *(1921–2007) was a American chemist. He lived his academic live at the University of Wisconsin, where he did his bachelor in 1942 and his PhD (supervised by Hirschfelder) in 1948. He became assistant professor in 1949, associate professor in 1954, and full professor in 1960. He retired in 1989.*

Joseph Oakland Hirschfelder *(1911–1990) was an American chemist in physicist. He studied natural sciences at University of Minnesota (1927–1929) and Yale (1929–1931) and finished his PhD in chemistry and physics at Princeton in 1936. From 1936–1937, he was postdoc with John von Neumann at the Princeton Institute for Advanced Studies. In 1937, he went to the University of Wisconsin, where he became assistant professor in chemistry in 1941. 1944/45, he was group leader at the Manhattan project at Los Alamos and later leading researcher in the American nuclear program. In 1946, he became full professor at the University of Wisconsin. He retired in 1981.*

Charles Francis Curtiss, Joseph Oakland Hirschfelder: Integration of stiff equations, *Proceedings of the National Academy of Sciences of the USA; 38 (1952), 235–243. [This work introduces the BDF multistep methods for the solution of stiff ODEs]*

The most simplest way of a numerical solver for this problem is to introduce a mesh $\Delta_x := \{x_0, \ldots, x_N\}$ of $[a, b]$ with $h := (b-a)/N$ and $x_j := a + jh$ for $j = 0, \ldots, N$, and to define $u_j(t) := u(x_j, t)$. Since $u_0(t) = u_N(t) = 0$ we collect the unknowns in the vector function $\mathbf{U}_h := (u_1, \ldots, u_{N-1})^\top$ and replace the partial derivative ∂_x^2 by the finite difference

$$\partial_x^2 u(x,t)|_{x=x_j} \approx \frac{1}{h^2}\left[u(x_{j-1},t) - 2u(x_j,t) + u(x_{j+1},t)\right], \qquad j = 1, \ldots, N-1,. \tag{4.4}$$

This leads to the initial value problem in time

$$\begin{aligned} \forall t \in [0,T]: \quad \mathbf{U}_h'(t) &= \mathbf{M}_h \mathbf{U}_h(t) + \mathbf{F}(t), && (4.5a) \\ \mathbf{U}_h(0) &= \mathbf{G}, && (4.5b) \end{aligned}$$

with $\mathbf{F}(t) := (f(x_1,t), \ldots, f(x_{N-1},t))^\top$, $\mathbf{G} := (g(x_1), \ldots, g(x_{N-1}))^\top$, and

$$\mathbf{M}_h := \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{pmatrix}. \tag{4.6}$$

Since the eigenvalues of $\mathbf{M}_h$ are given by $\lambda_j = \frac{2}{h^2}\left(-1 + \cos(\frac{j\pi}{N})\right)$ for $j = 1, \ldots, N-1$, this system is a prototype of a stiff ODE of the form (4.1) as well.

In the following we show, that solutions to (4.1) can be derived easily, if the eigenvalues and eigenvectors of $M \in \mathbb{R}^{n\times n}$ are known. So there would be no need for a numerical method for small values of n. But for large values of n computing all these eigenpairs will not be possible in general. Hence, the following results are for large n only of theoretical interest. Practically, we still need a numerical method to solve such systems.

Proposition 4.3. *Given $y_0 \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n\times n}$, consider the initial value problem*

$$y'(t) = My(t) \text{ in } [t_0, T] \quad \text{subject to} \quad y(t_0) = y_0. \tag{4.7}$$

Then, the unique solution $y \in C^1([t_0,T]; \mathbb{R}^n)$ reads

$$y(t) = e^{M(t-t_0)} y_0 \tag{4.8}$$

with the matrix exponential function

$$e^X := \sum_{k=0}^{\infty} \frac{1}{k!} X^k \quad \text{for } X \in \mathbb{R}^{n\times n}. \tag{4.9}$$

Suppose that M is diagonalizable, i.e., there exist $\{v_1, \ldots v_n\} \subset \mathbb{C}^n$ linearly independent eigenvectors for eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ such that $Mv_j = \lambda_j v_j$ for all $j = 1, \ldots, n$. Then,

$$y_0 = \sum_{j=1}^n \alpha_j v_j \quad \Longrightarrow \quad y(t) = \sum_{j=1}^n \left(\alpha_j e^{\lambda_j (t-t_0)}\right) v_j, \tag{4.10}$$

i.e., $y = \sum_{j=1}^n y_j v_j$, where the scalar functions $y_j \in C^1[t_0,T]$ solve the ODEs

$$y_j'(t) = \lambda_j y_j(t) \text{ in } [t_0, T] \quad \text{subject to} \quad y_j(t_0) = \alpha_j. \tag{4.11}$$

PROOF. Existence and uniqueness of the solution of (4.7) (resp. (4.11)) follows from the Picard–Lindelöf theorem. Consider y given by (4.8). Then, $y(t_0) = y_0$ and

$$\begin{aligned} y'(t) &= d_t\Big(\sum_{k=0}^{\infty} \frac{1}{k!}\, M^k (t-t_0)^k y_0\Big) \\ &= \sum_{k=1}^{\infty} \frac{1}{k!}\, k\, M^k (t-t_0)^{k-1} y_0 \\ &= M\Big(\sum_{k=1}^{\infty} \frac{1}{(k-1)!}\, M^{k-1} (t-t_0)^{k-1} y_0\Big) \\ &= M\Big(\sum_{k=0}^{\infty} \frac{1}{k!}\, M^k (t-t_0)^k y_0\Big) = My(t). \end{aligned}$$

If M is diagonalizable and $y_0 = \sum_{j=1}^n \alpha_j v_j$, then

$$\begin{aligned} y(t) &= \sum_{k=0}^{\infty} \frac{1}{k!}\, M^k (t-t_0)^k y_0 \\ &= \sum_{j=1}^{n} \Big(\sum_{k=0}^{\infty} \frac{1}{k!}\, M^k (t-t_0)^k \alpha_j v_j\Big) \\ &= \sum_{j=1}^{n} \Big(\alpha_j \sum_{k=0}^{\infty} \frac{1}{k!}\, \lambda_j^k (t-t_0)^k v_j\Big) \\ &= \sum_{j=1}^{n} \Big(\alpha_j e^{\lambda_j (t-t_0)}\Big)\, v_j. \end{aligned}$$

Since $y_j = e^{\lambda_j(t-t_0)}\alpha_j$, this concludes the proof. □

More generally, one can even prove the following result. We note that the solution $z \in C^1([t_0,T];\mathbb{R}^n)$ of (4.13) can be obtained by solving n scalar ODEs (if $\widetilde{g}(t) := V^{-1}g(t)$ is explicitly known).

PROPOSITION 4.4. *Given $g \in C([t_0,T];\mathbb{R}^n)$, $M \in \mathbb{R}^{n\times n}$, and $y_0 \in \mathbb{R}^n$, consider the in homogeneous initial value problem*

$$y'(t) = My(t) + g(t) \text{ in } [t_0,T], \quad y(t_0) = y_0. \tag{4.12}$$

Suppose that the matrix $M \in \mathbb{R}^n$ is diagonalizable, i.e., there exist $\{v_1,\dots v_n\} \subset \mathbb{C}^n$ linearly independent eigenvectors for eigenvalues $\lambda_1,\dots,\lambda_n \in \mathbb{C}$ such that $Mv_j = \lambda_j v_j$ for all $j = 1,\dots,n$. Define $\Lambda := \operatorname{diag}(\lambda_1,\dots,\lambda_n) \in \mathbb{R}^{n\times n}$ and $V := (v_1,\dots,v_n) \in \mathbb{R}^{n\times n}$ and consider the problem

$$z'(t) = \Lambda z(t) + V^{-1}g(t) \text{ in } [t_0,T], \quad z(t_0) = V^{-1}y_0. \tag{4.13}$$

Then, (4.12)*–*(4.13) *have unique solutions and it holds that $y(t) = Vz(t)$.*

PROOF. Existence and uniqueness of the solutions of (4.12)–(4.13) follows from the Picard–Lindelöf theorem. Clearly, V is invertible and independent of t. Therefore, (4.13) implies that

$$(Vz)'(t) = V\Lambda z(t) + g(t).$$

Note that $V\Lambda = MV$, since the columns of V are the eigenvectors of M. Hence, $\widetilde{y} := Vz$ solves that

$$\widetilde{y}'(t) = M\widetilde{y}(t) + g(t) \quad \text{together with} \quad \widetilde{y}(t_0) = Vz(t_0) = y_0.$$

From the uniqueness of solutions, we thus conclude that $\widetilde{y} = y$. □

EXERCISE 4.5. Consider the setting of Proposition 4.4. Let $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ be an explicit m-stage Runge–Kutta method. Let $y_\ell, z_\ell \in \mathbb{R}^n$ be the resulting Runge–Kutta iterates of (4.12)–(4.13). Prove that $y_\ell = Vz_\ell$ for all $\ell = 0, \dots, N$, i.e., explicit Runge–Kutta methods commute with diagonalization of the ODE system.

If we consider the simple homogeneous model problem from Proposition 4.3, it follows that, for a stiff ODE, we should employ Runge–Kutta methods, which appropriately integrate the scalar ODEs

$$y_j'(t) = \lambda_j y(t) \quad \text{in } [t_0, T]$$

simultaneously for all eigenvalues $\lambda_j \in \mathbb{C}$ of $M \in \mathbb{R}^{n \times n}$.

REMARK 4.6. In Section 3.1 of Chapter 3 we have already considered the explicit and the implicit Euler method for such problems. The explicit Euler method applied to such problems leads to (see (3.4))

$$y_\ell = (1 + h\lambda)^\ell \, y_0 \quad \text{for all } \ell \geq 0,$$

while the implicit Euler method leads for $h\lambda \neq 1$ to (see (3.5))

$$y_\ell = (1 - h\lambda)^{-\ell} \, y_0 \quad \text{for all } \ell \geq 0.$$

We make the following observation: If $\operatorname{Re}\lambda \gg 0$, then both methods will need small h, since the exact solution grows exponentially. If $\operatorname{Re}\lambda < 0$, then the explicit Euler method needs small h so that (3.4) leads to $y_\ell \to 0$ as $\ell \to \infty$. If $|1 + h\lambda| > 1$, then (3.4) leads to blow-up and oscillations. On the other hand, the implicit Euler method (3.5) guarantees always $y_\ell \to 0$ as $\ell \to \infty$, if $\operatorname{Re}\lambda < 0$.

EXAMPLE 4.7. We consider the ODE

$$y' = My \text{ in } \mathbb{R}_{\geq 0} \quad \text{with} \quad M = \begin{pmatrix} a & b \\ b & a \end{pmatrix}.$$

Obviously, the eigenvalues of M are $\lambda_{1/2} = a \pm b$ with eigenvectors $v_{1/2} = (1, \pm 1)^\top$. According to Proposition 4.3, the unique solution satisfies

$$y(t) = \alpha_1 e^{(a+b)t} v_1 + \alpha_2 e^{(a-b)t} v_2 \quad \text{provided that} \quad y(0) = \alpha_1 v_1 + \alpha_2 v_2.$$

From the initial condition, we get that

$$\left.\begin{array}{rcl} y_1(0) & = & \alpha_1 + \alpha_2 \\ y_2(0) & = & \alpha_1 - \alpha_2 \end{array}\right\} \quad \text{and hence} \quad \left\{\begin{array}{rcl} \alpha_1 & = & \frac{y_1(0)+y_2(0)}{2} \\ \alpha_2 & = & \frac{y_1(0)-y_2(0)}{2}. \end{array}\right.$$

Altogether, the solution reads

$$\begin{aligned} y_1(t) &= \frac{y_1(0) + y_2(0)}{2} \, e^{(a+b)t} + \frac{y_1(0) - y_2(0)}{2} \, e^{(a-b)t}, \\ y_2(t) &= \frac{y_1(0) + y_2(0)}{2} \, e^{(a+b)t} - \frac{y_1(0) - y_2(0)}{2} \, e^{(a-b)t}. \end{aligned}$$

Let $a = -51$, $b = -50$, so that $a + b = -101$ and $a - b = -1$. For large $t > 0$, a good approximation of $y(t)$ is obtained, if we neglect the first summands with $e^{(a+b)t}$. But the

ODE is stiff. The explicit Euler method requires $|1 + h(a \pm b)| < 1$ to avoid oscillations (in both components) of y. However, note that the "component" $y(t) \cdot (1, -1)^\top$ (which corresponds to the "good" eigenvalue $\lambda_2 = a - b = -1$ will nevertheless be approximated well for any $0 < h < 1$.

4.2. Stability domains

The idea of all notions of stability is that the discrete solution should reflect certain properties of the continuous solution, at least for model problems, where the solution (or its qualitative behavior) is known. For stability domains (as well as A-stability and L-stability), one considers the scalar model problem

$$y'(t) = \lambda y(t) \text{ in } \mathbb{R}_{\geq 0}, \quad y(0) = y_0, \tag{4.14}$$

where $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda < 0$.

Definition 4.8. Let $\Phi(t, y, z, h)$ be the incremental function of a one-step method. A function $R : \mathbb{C} \to \mathbb{C}$ is called **stability function** of the method, if

$$y_{\ell+1} := y_\ell + h_\ell \Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \overset{!}{=} R(\lambda h_\ell)\, y_\ell \tag{4.15}$$

for the model problems (4.14).

Example 4.9. We rephrase Remark 4.6: The stability function of the explicit Euler method is $R(z) = 1 + z$, and it is well-defined for all $z \in \mathbb{C}$. The stability function of the implicit Euler method is $R(z) = 1/(1 - z)$, and it is well-defined for all $z \in \mathbb{C} \backslash \{1\}$.

Theorem 4.10 (Stability function of Runge–Kutta methods). *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an m-step Runge–Kutta method. Then, the stability function reads*

$$R(z) = 1 + z b^\top (I - zA)^{-1} \mathbb{1}, \quad \textit{where} \quad \mathbb{1} = (1, \dots, 1)^\top \in \mathbb{R}^m. \tag{4.16}$$

It is well-defined for all $z \in \mathbb{C}$ such that $1/z \notin \sigma(A) := \big\{\lambda \in \mathbb{C} : \lambda$ is eigenvalue of $A\big\}$. Moreover, there hold the following statements:

(i) *If the method is explicit, then $R \in \mathbb{P}_m$.*
(ii) *If the method is implicit, then $R = P/Q$ with $P, Q \in \mathbb{P}_m$.*

Proof. Recall the Runge–Kutta increments

$$k_i = f\Big(t + c_i h\,,\, y_\ell + h \sum_{j=1}^m A_{ij} k_j\Big) \overset{!}{=} \lambda \Big(y_\ell + h \sum_{j=1}^m A_{ij} k_j\Big) \quad \text{for all } i = 1, \dots, m.$$

With the vector $k = (k_1, \dots, k_m)^\top$, it follows that $k = \lambda y_\ell \mathbb{1} + \lambda h A\, k$ and hence

$$k = \lambda y_\ell (I - \lambda h A)^{-1} \mathbb{1}, \quad \text{provided that} \quad 1/(\lambda h) \notin \sigma(A).$$

By definition of a Runge–Kutta method, it follows that

$$y_{\ell+1} = y_\ell + h \sum_{i=1}^m b_i k_i = y_\ell + h b^\top k = y_\ell \left[1 + \lambda h\, b^\top (I - \lambda h A)^{-1} \mathbb{1}\right] = y_\ell\, R(\lambda h).$$

This proves (4.16).

For the remainder of the proof, recall **Cramer's rule**: Let $V = (v_1, \ldots, v_n) \in \mathbb{R}^n$ be invertible with columns $v_j \in \mathbb{R}^n$. Let $w \in \mathbb{R}^n$. Define $V_i := (v_1, \ldots, v_{i-1}, w, v_{i+1}, \ldots, v_n) \in \mathbb{R}^{n \times n}$. Then, the vector $x := V^{-1}w \in \mathbb{R}^n$ satisfies that

$$x_i = \frac{\det(V_i)}{\det(V)} \quad \text{for all } i = 1, \ldots, m. \tag{4.17}$$

To apply Cramer's rule, we write

$$R(z) = 1 + zb^\top (I - zA)^{-1}\mathbb{1} = 1 + zb^\top \gamma \quad \text{where} \quad (I - zA)\gamma = \mathbb{1}.$$

This is rewritten as

$$\underbrace{\begin{pmatrix} I - zA & 0 \\ -zb^\top & 1 \end{pmatrix}}_{=:V} \begin{pmatrix} \gamma \\ R(z) \end{pmatrix} = \begin{pmatrix} \mathbb{1} \\ 1 \end{pmatrix}.$$

According to the Laplace cofactor expansion (for the $(m+1)$-th column), it holds that

$$\det(V) = (-1)^{(m+1)+(m+1)} \det(I - zA) = \det(I - zA) \neq 0 \quad \text{provided that} \quad 1/z \notin \sigma(A).$$

Hence, V is invertible and Cramer's rule yields (for $R(m)$ being the $(m+1)$-th coefficient of the solution) that

$$R(z) = \frac{\det \begin{pmatrix} I - zA & \mathbb{1} \\ -zb^\top & 1 \end{pmatrix}}{\det \begin{pmatrix} I - zA & 0 \\ -zb^\top & 1 \end{pmatrix}} = \frac{\det \begin{pmatrix} I - zA & \mathbb{1} \\ -zb^\top & 1 \end{pmatrix}}{\det(I - zA)} =: \frac{P(z)}{Q(z)}.$$

Clearly, the characteristic polynomial $Q(z) = \det(I - zA)$ satisfies $Q \in \mathbb{P}_m$. Moreover, if the method is explicit, then A is strictly lower triangular and hence $\det(I - zA) = 1$. Hence, it only remains to show that the numerator satisfies that $P \in \mathbb{P}_m$. With the Laplace cofactor expansion (for the $(m+1)$-th column), we write

$$P(z) = \det \begin{pmatrix} I - zA & \mathbb{1} \\ -zb^\top & 1 \end{pmatrix} =: \det \begin{pmatrix} S - zT & \mathbb{1} \end{pmatrix} = \sum_{i=1}^{m+1} (-1)^{i+(m+1)} \det \big(S_i - zT_i\big),$$

where $S_i, T_i \in \mathbb{R}^{m \times m}$ are obtained by canceling the i-th row of $S, T \in \mathbb{R}^{(m+1) \times m}$. A simple induction on the dimension m (together with the Laplace cofactor expansion) proves that $\det\big(S_i - zT_i\big) \in \mathbb{P}_m$. Therefore, we conclude that $P \in \mathbb{P}_m$. □

COROLLARY 4.11 (Stability function is an approximation of exp). *Let $R(z)$ be the stability function of a Runge–Kutta method with consistency order $p \geq 1$. Then,*

$$R(z) = \exp(z) + \mathcal{O}(z^{p+1}). \tag{4.18}$$

Moreover, if the method is explicit, then it is a polynomial with

$$R(z) = \sum_{j=1}^{p} \frac{z^j}{j!} + \mathcal{O}(z^{p+1}), \tag{4.19}$$

Gabriel Cramer *(1704–1752) was a Swiss Mathematician. He studied Mathematics at the University of Geneva and took his PhD in 1722. In 1724, he became professor at the University of Geneva. He had personal contact with his colleagues Johann Bernoulli, Leonhard Euler, and James Stirling.*

Pierre-Simon (Marquis de) Laplace *(1749–1827) was a French Mathematician and, from 1768–1771, a student of d'Alembert in Paris. In 1771, he became teacher for geometry, trigonometry, analysis, and statistics at the military academy in Paris. In 1773, be became associate member of the French Academy of Sciences.*

and $\mathcal{O}(z^{p+1})$ vanishes for an explicit p-stage method of order p.

Proof. Consider the first step of the Runge–Kutta method for $\lambda = -1$ and $y_0 = 1$. Then,

$$\exp(-h) = y(h) \approx y_1 = R(-h)y_0 = R(-h).$$

Moreover, consistency order p thus implies that

$$\exp(-h) - R(-h) = \mathcal{O}(h^{p+1}). \tag{4.20}$$

Note that $R(z) = 1 + zb^\top(I - zA)^{-1}\mathbb{1}$ is smooth locally around $z = 0$ (i.e., analytic). Hence, it holds that

$$f(z) := R(z) - \exp(z) = \sum_{n=0}^{p} \frac{f^{(n)}(0)}{n!}\, z^n + \mathcal{O}(z^{p+1}).$$

Due to (4.20), we see that $f^{(n)}(0) = 0$ for all $n = 0, \dots, p$. Therefore, it follows that

$$R(z) = \exp(z) + \big(R(z) - \exp(z)\big) = \exp(z) + \mathcal{O}(z^{p+1}).$$

If the method is explicit, then $R \in \mathbb{P}_m$, i.e., $R(z) = \sum_{j=0}^{m} a_j z^j$. Then,

$$\sum_{j=0}^{p} \Big(a_j - \frac{1}{j!}\Big)\, z^j = R(z) - \exp(z) + \mathcal{O}(z^{p+1}) \overset{(4.18)}{=} \mathcal{O}(z^{p+1}).$$

Consequently, the lower-order powers of z vanish, i.e., $a_j = 1/j!$ for $j = 0, \dots, p$. This concludes the proof. □

Example 4.12. The stability function of the Heun method (Example 2.19) and the modified Euler method (Example 2.16) is $R(z) = 1 + z + \frac{z^2}{2}$, since both methods are explicit 2-stage methods of order $p = 2$. The stability function of the classical RK4 method (Example 2.25) is $R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}$, since RK4 is a 4-step method of order $p = 4$.

Definition 4.13. Let $R(z)$ be the stability function of a one-step method. Then,

$$S := \big\{z \in \mathbb{C} \,:\, |R(z)| \le 1\big\} \tag{4.21}$$

denotes the corresponding **stability domain**.

Remark 4.14. If $\lambda h = z \in S$, then the approximations $y_\ell \approx y(t_\ell)$ of the model problem (4.14) remain bounded. Note that $y(t) = e^{\lambda t} y_0 \to 0$ as $t \to \infty$, since $\operatorname{Re}\lambda < 0$.

Example 4.15. The stability domain of the explicit Euler method is

$$S = \big\{z \in \mathbb{C} \,:\, |1 + z| \le 1\big\} = \overline{U_1(-1)}.$$

The stability domain of the implicit Euler method is

$$S = \Big\{z \in \mathbb{C} \,:\, \Big|\frac{1}{1-z}\Big| \le 1\Big\} = \big\{z \in \mathbb{C} \,:\, 1 \le |1 - z|\big\} = \mathbb{C}\backslash U_1(1).$$

Corollary 4.16. *For all Runge–Kutta methods with consistency order $p \ge 1$, it holds that $0 \in \partial S$.*

Proof. According to Corollary 4.11, it holds that

$$R(z) = \exp(z) + \mathcal{O}(z^{p+1}) = 1 + z + \mathcal{O}(z^2).$$

For all sufficiently small $z = \pm h$, it hence follows that

- $R(+h) = 1 + h + \mathcal{O}(h^2) \ge 1 + h/2 > 1$, i.e., $+h \notin S$;

- $R(-h) = 1 - h + \mathcal{O}(h^2) \leq 1 - h/2 < 1$, i.e., $-h \in S$.

Hence, each neighborhood of $z = 0$ contains one point $-h \in S$ and $+h \notin S$. This proves that $0 \in \partial S$. □

4.3. A-stability and L-stability

Definition 4.17. Let $R(z)$ be the stability function of a one-step method. The one-step method is

- **A-stable**, if $\sup_{\operatorname{Re} z \leq 0} |R(z)| \leq 1$;
- **L-stable**, if it is A-stable and $\lim_{\operatorname{Re} z \to -\infty} |R(z)| = 0$.

Remark 4.18. Due to the definition of the stability domain S in (4.21), A-stability is equivalent to $\mathbb{C}^- := \{z \in \mathbb{C} : \operatorname{Re} z \leq 0\} \subseteq S$. Hence, the explicit Euler method is *not* A-stable (see Example 4.15).

Remark 4.19. If a method is A-stable, then

$$|y_{\ell+1}| = |R(\lambda h)||y_\ell| \leq |y_\ell| \quad \text{for all } h > 0 \text{ and } \operatorname{Re} \lambda \leq 0,$$

i.e., the discrete solutions of the model problem (4.14) are non-expansive and, in particular, remain at least bounded.

If a method is L-stable, then

$$|y_{\ell+1}| = |R(\lambda h)||y_\ell| \to 0 \quad \text{as } h \to \infty \text{ for all } \operatorname{Re} \lambda \leq 0,$$

i.e., the discrete solutions of the model problem (4.14) decay with larger time-steps.

Example 4.20. The implicit Euler method satisfies that $R(z) = \frac{1}{1-z}$ and $S = \mathbb{C} \backslash U_1(1) \supset \mathbb{C}^-$. Hence, the implicit Euler method is L-stable and A-stable.

Exercise 4.21. Show that the stability function of the implicit midpoint rule is $R(z) = \frac{1+z/2}{1-z/2}$. Determine the stability domain of the implicit midpoint rule! Is the implicit midpoint rule A-stable and/or L-stable?

Theorem 4.22 (Explicit Runge–Kutta methods fail). *No explicit Runge–Kutta method is L-stable. No consistent explicit Runge–Kutta method is A-stable. In particular, any A-stable and/or L-stable Runge–Kutta method is implicit.*

Proof (L-stability fails). Recall that the stability function of an m-stage Runge–Kutta method satisfies that $R = P/Q$ with $P, Q \in \mathbb{P}_m$. Obviously, L-stability thus is equivalent to the fact that $P \in \mathbb{P}_{\nu-1}$ and $Q \in \mathbb{P}_\nu \backslash \mathbb{P}_{\nu-1}$ for some $1 \leq \nu \leq m$ to ensure that $P/Q = \mathcal{O}(1/z)$ as $|z| \to \infty$. □

Proof (A-stability fails). The stability function of an explicit m-stage Runge–Kutta method $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ satisfies that $R \in \mathbb{P}_m$. Consistency implies that $\sum_{j=1}^m b_j = 1$ and hence the method has consistency order $p \geq 1$. Together with Corollary 4.11, this implies that

$$R(z) = \sum_{j=0}^{\nu} a_j z^j \quad \text{for some } 1 \leq \nu \leq m \text{ with } a_\nu \neq 0.$$

The triangle inequality thus proves that

$$|R(z)| \geq |a_\nu||z|^\nu - \sum_{j=0}^{\nu-1} |a_j||z|^j = |z|^\nu \Big(|a_\nu| - \sum_{j=0}^{\nu-1} |a_j| \frac{1}{|z|^{\nu-j}}\Big) \xrightarrow{\operatorname{Re} z \to -\infty} \infty.$$

Hence, $\sup\limits_{\operatorname{Re} z \leq 0} |R(z)| = \infty$ and the method is *not* A-stable. $\square$

THEOREM 4.23. *Let* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *be an implicit Runge–Kutta method such that A is invertible and A-stable. Then, the following statements* (i)–(ii) *are equivalent:*

(i) $b^\top A^{-1} \mathbb{1} = 1$ *with* $\mathbb{1} = (1, \dots, 1)^\top \in \mathbb{R}^m$.

(ii) *The method is L-stable.*

PROOF. If $\|\frac{1}{z} A^{-1}\| < 1$ (e.g., $|z| > 2\|A^{-1}\|$), then the Neumann series proves that

$$\Big(I + \frac{1}{z} A^{-1}\Big)^{-1} = \sum_{j=0}^{\infty} \big(-\frac{1}{z} A^{-1}\big)^j = I + \mathcal{O}\left(\frac{1}{z}\right)$$

Moreover,

$$I - zA = (A^{-1} - zI)A = -z\Big(I - \frac{1}{z} A^{-1}\Big) A.$$

This leads to

$$(I - zA)^{-1} = -\frac{1}{z} A^{-1} \Big(I + \frac{1}{z} A^{-1}\Big)^{-1} = -\frac{1}{z} A^{-1} + \mathcal{O}\left(\frac{1}{z^2}\right).$$

With Theorem 4.10, it follows that

$$R(z) \overset{(4.16)}{=} 1 + z b^\top (I - zA)^{-1} \mathbb{1} = 1 - b^\top A^{-1} \mathbb{1} + \mathcal{O}\left(\frac{1}{z}\right).$$

Hence, L-stability is equivalent to $1 - b^\top A^{-1} \mathbb{1} = 0$. This concludes the proof. $\square$

REMARK 4.24 (Radau-IIA methods). The Radau-IIA methods are collocation methods with $c_m = 1$. The remaining nodes c_j of $c \in \mathbb{R}^m$ are chosen such that the induced quadrature rule has maximal exactness, see Section B.6 in Chapter B. Arguing as for the Gaussian quadrature rule, one obtains that the nodes $0 < c_1 < c_2 < \cdots < c_m = 1$ are unique. The quadrature rule has exactness $2m - 2$ (i.e., one order less than the Gaussian quadrature). Hence, the Radau-IIA methods have consistency order $p = 2m - 1$. Recall from Theorem 3.23 that the Runge–Kutta data of a collocation method read

$$A_{ij} = \int_0^{c_i} L_j(\tau) \,\mathrm{d}\tau \quad \text{and} \quad b_j = \int_0^1 L_j(\tau) \,\mathrm{d}\tau \quad \text{for all } i, j = 1, \dots, m.$$

For Radau-IIA methods, the last row of $A \in \mathbb{R}^{m \times m}$ thus coincides with $b \in \mathbb{R}^m$ (since $c_m = 1$). Moreover, A is invertible by Cor.3.25, since $c_j > 0$. With the m-th unit vector $e_m \in \mathbb{R}^m$, this means that $b^\top = e_m^\top A$ and hence $b^\top A^{-1} = e_m^\top$. Consequently,

$$b^\top A^{-1} \mathbb{1} = e_m^\top \mathbb{1} = 1.$$

Jean Charles Rodolphe Radau *(1835–1911) was a German-French mathematician. Born in Germany, he studied mathematics and astronomy at the University of Königsberg. In 1858, he moved to Paris and later became French citizen in 1873. In 1897, he became member of the French Academy of Sciences.*

According to Theorem 4.23, the Radau-IIA methods are L-stable, if they are A-stable. The latter will be shown in the following. The first Butcher tableau with consistency orders $p = 1$, $p = 3$, and $p = 5$ are

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

Note, that the one stage Radau method is the implicit Euler method.

4.4. Dissipative systems and B stability

Definition 4.25. A function $f \in C(\Omega;\mathbb{R}^n)$ with $\Omega \subset \mathbb{R}^n$ is called dissipative with respect to the scalar product $\langle \cdot\,,\,\cdot\rangle$, if

$$\forall y, \tilde{y} \in \Omega : \quad \langle f(y) - f(\tilde{y})\,,\, y - \tilde{y}\rangle \leq 0. \tag{4.22}$$

We call the autonomous system $y' = f(y)$ with $y(t_0) = y_0$ dissipative, if the right hand side f is dissipative.

For $n = 1$ and the usual scalar product in $\mathbb{R}$ a function is dissipative if and only if the function is monotonically decreasing.

Proposition 4.26. *Suppose $f \in C(\mathbb{R}^n;\mathbb{R}^n)$ is dissipative with respect to the scalar product $\langle \cdot\,,\,\cdot\rangle$. Then, for the corresponding autonomous initial value problem $y' = f(y)$ with $y(t_0) = y_0$ the following statements hold.*

(i) *The solution y exists for all $t \geq t_0$.*

(ii) *The problem is non-expansive, i.e. for all initial values $y_0, \tilde{y}_0 \in \mathbb{R}^n$ with corresponding solutions y, $\tilde{y}$ there holds*

$$\forall t \geq t_0 : \quad \|y(t) - \tilde{y}(t)\| \leq \|y_0 - \tilde{y}_0\|. \tag{4.23}$$

Thereby $\|\cdot\|$ denotes the norm induces by the scalar product, i.e. $\|\cdot\| := \sqrt{\langle \cdot\,,\,\cdot\rangle}$.

Proof. We first show (ii) for all $t > t_0$ such that the solutions y and $\tilde{y}$ exist. Let us define the scalar, non-negative function u by $u(t) := \|y(t) - \tilde{y}(t)\|^2$. Using

$$u'(t) = 2\,\langle y'(t) - \tilde{y}'(t)\,,\, y(t) - \tilde{y}(t)\rangle = 2\,\langle f(y(t)) - f(\tilde{y}(t))\,,\, y(t) - \tilde{y}(t)\rangle \leq 0$$

the claim follows by

$$\|y(t) - \tilde{y}(t)\|^2 = u(t) = u(0) + \int_{t_0}^{t} u'(\tau)\,\mathrm{d}\tau \leq u(0) = \|y_0 - \tilde{y}_0\|^2.$$

Second, we prove by contradiction that the solutions exists for all $t \geq t_0$. Let us assume, that there exists a $t_+ > t_0$ and $t_+ < \infty$ such that the maximal interval of existence for the solution y with initial value y_0 is $[t_0, t_+)$. Since the domain of definition for f is the whole $\mathbb{R}^n$, there needs to be a blow-up at t_+, i.e. $\|y(t)\| \to \infty$ for $t \to t_+$.

Let $\hat{t} \in (t_0, t_+)$ and $\tilde{y}_0 := y(\hat{t})$. Since the differential equation is autonomous, we have for the solution $\tilde{y}$ with initial value $\tilde{y}(t_0) = \tilde{y}_0$

$$\forall t \in (t_0, t_+ - \hat{t} + t_0) : \qquad \tilde{y}(t) = y(t + \hat{t} - t_0).$$

Hence, $\|\tilde{y}(t)\| \to \infty$ for $t \to t_+ - \hat{t} + t_0 < t_+$. Using the inequality (4.23), which we have already shown for $t \in (t_0, t_+ - \hat{t} + t_0)$, leads to the contradiction

$$\|\tilde{y}(t)\| \leq \|y(t) - \tilde{y}(t)\| + \|y(t)\| \leq \|y_0 - \tilde{y}_0\| + \|y(t)\| < \infty.$$

□

The concept of A stability as defined in Def. 4.17 uses the stability function and is therefore related to linear differential equations. The concept of B stability we are going to introduce is related to dissipative, autonomous, but non-linear systems.

Definition 4.27. Let $\Phi(t, y, z, h)$ be the incremental function of an (implicit) one-step method. Then the one-step method is called **B-stable**, if for all dissipative $f \in C(\mathbb{R}^n; \mathbb{R}^n)$ the method guarantees that for all h such that the one-step method is well-defined there holds

$$\forall t_0 \in \mathbb{R} \quad \forall y_0, \tilde{y}_0 \in \mathbb{R}^n : \quad \|y_1 - \tilde{y}_1\| \le \|y_0 - \tilde{y}_0\|, \tag{4.24}$$

where

$$y_1 := y_0 + h\Phi(t_0, y_0, y_1, h) \quad \text{and } \tilde{y}_1 := \tilde{y}_0 + h\Phi(t_0, \tilde{y}_0, \tilde{y}_1, h).$$

Again, $\|\cdot\|$ is the norm induced by the scalar product for which f is dissipative.

Lemma 4.28. *Each B-stable Runge-Kutta method is A-stable.*

Proof. By Def. 4.17 we have to show that the stability function R of the one-step method satisfies $|R(z)| \le 1$ for $z \in \mathbb{C}_{\mathrm{Re}\le 0} := \{z \in \mathbb{C} : \mathrm{Re}(z) \le 0\}$. Let $\lambda \in \mathbb{C}_{\mathrm{Re}\le 0}$ and consider the (complex) initial value problem

$$y' = \lambda y, \qquad y(0) = 1. \tag{4.25}$$

Note, that this was the model problem used for the definition of stability functions in Def. 4.8.

We first show, that this model problem is equivalent to a real, dissipative system of ODEs. To this end, let $y = u + \mathrm{i}v$ and $\lambda = \alpha + \mathrm{i}\beta$ with $u(t), v(t), \alpha, \beta \in \mathbb{R}$ for all $t > 0$. Since $\lambda \in \mathbb{C}_{\mathrm{Re}\le 0}$, we have $\alpha \le 0$. (4.25) is equivalent to the real system of Ode's

$$\begin{pmatrix} u \\ v \end{pmatrix}' = M \begin{pmatrix} u \\ v \end{pmatrix}, \quad \begin{pmatrix} u \\ v \end{pmatrix}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \text{with } M := \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}. \tag{4.26}$$

The right hand side $f(\left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)) := M \left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)$ is linear and dissipative with respect to the Euclidean scalar product in $\mathbb{R}^2$, since

$$\left\langle M \begin{pmatrix} u \\ v \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \right\rangle = \begin{pmatrix} \alpha u - \beta v \\ \beta u + \alpha v \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} = \alpha(u^2 + v^2) \le 0.$$

Hence, the concept of B stability applies and

$$\|y_{\ell+1} - \tilde{y}_{\ell+1}\|_{\mathbb{C}} = \left\| \begin{pmatrix} u_{\ell+1} \\ v_{\ell+1} \end{pmatrix} - \begin{pmatrix} \tilde{u}_{\ell+1} \\ \tilde{v}_{\ell+1} \end{pmatrix} \right\|_{\mathbb{R}^2} \le \left\| \begin{pmatrix} u_\ell \\ v_\ell \end{pmatrix} - \begin{pmatrix} \tilde{u}_\ell \\ \tilde{v}_\ell \end{pmatrix} \right\|_{\mathbb{R}^2} = \|y_\ell - \tilde{y}_\ell\|_{\mathbb{C}} \,.$$

Since the right hand side of (4.25) is linear, the Runge-Kutta method is linear as well and the claim follows by $y_{\ell+1} - \tilde{y}_{\ell+1} = R(\lambda h_\ell)\,(y_\ell - \tilde{y}_\ell)$. □

Example 4.29. We will show next, that the implicit midpoint rule of Example 3.5 is B stable. Hence, it is A stable as well. The stability function of the midpoint rule is $R(z) = \frac{2+z}{2-z}$. The trapezoidal rule Ex. 3.6 has the same stability function, i.e. applied to linear systems both rules give identical approximations. So the trapezoidal rule is A stable as well. Nevertheless, it can be shown that it is not B stable.

Definition 4.30 (Gauss method). Gauss methods are collocation methods induced by the Gauss quadratures defined in Def. B.11, which are given in Exa. B.16. The first Butcher tableau are

$$\begin{array}{c|c} \frac12 & \frac12 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} \frac12-\frac{\sqrt3}{6} & \frac14 & \frac14-\frac{\sqrt3}{6} \\ \frac12+\frac{\sqrt3}{6} & \frac14+\frac{\sqrt3}{6} & \frac14 \\ \hline & {}^1\!/_2 & {}^1\!/_2 \end{array} \qquad \begin{array}{c|ccc} \frac12-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac29-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\ \frac12 & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac29 & \frac{5}{36}-\frac{\sqrt{15}}{24} \\ \frac12+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac29+\frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & {}^5\!/_{18} & {}^4\!/_9 & {}^5\!/_{18} \end{array}.$$

By Theorem 3.30 these methods have the maximal consistency order $p = 2$, $p = 4$, and $p = 6$. The one stage Gauss method is the implicit midpoint rule of Example 3.5.

Theorem 4.31. *Gauss methods are B-stable and A-stable.*

Proof. Lemma 4.28 ensures A stability, if we can show B stability. Hence, we assume that f is dissipative with respect to $\langle\cdot\,,\,\cdot\rangle$ and sufficiently smooth. Moreover, we have a m-stage Gauss method with mesh-size h and collocation polynomials $q, \tilde q \in \mathbb{P}_m$ as defined in (3.25) to the initial values $q(0) = y_0$ and $\tilde q(0) = \tilde y_0$ respectively. Since $\|\cdot\| = \sqrt{\langle\cdot\,,\,\cdot\rangle}$ the function $u : [0, 1] \to \mathbb{R}$ defined by

$$u(\tau) := \|q(\tau h) - \tilde q(\tau h)\|^2$$

belongs to $\mathbb{P}_{2m}$. Moreover, by definition of the collocation polynomials in (3.25) there holds $q'(c_j h) = f(q(c_j h))$ and $\tilde q'(c_j h) = f(\tilde q(c_j h))$. Hence,

$$\forall j = 1, \dots, m : \quad u'(c_j) = 2h\,\langle f(q(c_j h)) - f(\tilde q(c_j h))\,,\, q(c_j h) - \tilde q(c_j h)\rangle \le 0,$$

since f is dissipative. Finally, we have

$$\|y_1 - \tilde y_1\|^2 = u(1) = u(0) + \int_0^1 u'(\tau)\,\mathrm{d}\tau = \|y_0 - \tilde y_0\|^2 + \int_0^1 u'(\tau)\,\mathrm{d}\tau.$$

So the method is B stable if and only if $\int_0^1 u'(\tau)\,\mathrm{d}\tau \le 0$. This is guaranteed, since for $u' \in \mathbb{P}_{2m-1}$ the Gauss quadrature is exact, i.e.

$$\int_0^1 u'(\tau)\,\mathrm{d}\tau = \sum_{j=1}^{m} \alpha_j u'(c_j).$$

Note, that the quadrature weighs α_j are positive for all $j = 1, \dots, m$ by Theorem B.15. □

Remark 4.32. The Radau methods of Remark 4.24 are B-stable as well. Since Radau quadratures are exact only for polynomials of degree $2m-2$, the last step of the preceding proof has to be modified by introducing the error induced by the Radau quadrature. It can be shown, that this additional error term has the correct sign such that $\int_0^1 u'(\tau)\,\mathrm{d}\tau \le 0$ still holds.

There are many more stability concepts for numerical integrators of ODEs. Most of them try to guarantee that a specific behavior of the true solution is carried over to the discrete level. Note, that there is a severe difference between these concepts and the stability introduced first in (2.14). The latter is necessary to ensure convergence of the method. A-, B-, and L-stability are not relevant for the convergence. Explicit as well as implicit one-step methods will converge to the true solution, if the mesh-size is sufficiently small. Using A-, B-, or L-stable one-step methods we try to avoid unnecessary small time-steps.

CHAPTER 5

Multi-Step methods

5.1. Idea and definition

Up to now, we considered one-step methods

$$y_{\ell+1} := y_\ell + h_\ell\,\Phi(t_\ell, y_\ell, y_{\ell+1}, h_\ell) \quad \text{for all } \ell = 0, \dots, N-1.$$

To compute the next approximation $y_{\ell+1} \approx y(t_\ell + h_\ell)$ only the value $y_\ell \approx y(t_\ell)$ was used. The computational costs per step depend for explicit Runge-Kutta methods on the number of function evaluations, i.e. on the number of stages. Since the number of stages increases with the convergence order of the method, the computational costs per step increase as well. Note, that the overall costs depend on the number of steps $N \approx {}^{(T-t_0)}/_{h^p_\Delta}$ for a method of order p and a sufficiently smooth right hand side f of the ODE.

In this section we try to construct a high order method such that only one function evaluation per step is required. This can be archived, if we use not only y_ℓ but the approximations $y_{\ell-j}$ for $j = 1, \dots, k-1$ as well. Since multi-step methods with non-uniform step-sizes are awkward, we confine ourselves to uniform time discretizations.

Definition 5.1 (Multi-Step-Method). Let $\Delta := \{t_0, \dots, t_N\}$ be a uniform mesh on $[t_0, T]$, i.e. $t_\ell := t_0 + \ell h$ for $\ell = 0, \dots, N$ and $h := {}^{(T-t_0)}/_N$. Moreover, for $k \in \mathbb{N}$ the data $y_0, \dots, y_{k-1} \in \mathbb{R}^n$ should be given. Then

$$\sum_{j=0}^{k} \alpha_{k-j} y_{\ell+1-j} = h\Phi(t_\ell, y_{\ell+1}, \dots, y_{\ell+1-k}, h), \qquad \ell = k-1, \dots, N-1 \tag{5.1}$$

with $\alpha_k := 1$, constants $\alpha_0, \dots, \alpha_{k-1} \in \mathbb{R}$ and $\Phi : [t_0, T) \times (\mathbb{R}^n)^{k+1} \times \mathbb{R}_+ \to \mathbb{R}^n$ is called a *k-step method*. The method is called *linear k-step method*, if

$$\sum_{j=0}^{k} \alpha_{k-j} y_{\ell+1-j} = h \sum_{j=0}^{k} \beta_{k-j} f\left(t_{\ell+1-j}, y_{\ell+1-j}\right), \qquad \ell = k-1, \dots, N-1 \tag{5.2}$$

with constants $\beta_0, \dots, \beta_k \in \mathbb{R}$ and $|\alpha_0| + |\beta_0| \neq 0$.

For $k = 1$ (5.1) reads

$$y_{\ell+1} + \alpha_0 y_\ell = h\Phi(t_\ell, y_{\ell+1}, y_\ell, h), \qquad \ell = 0, \dots, N-1.$$

Hence, if $\alpha_0 = -1$ we have the standard form of a one-step-method. We will focus in the following on linear multi-step methods (5.2). For $k = 1$, $\alpha_0 = -1$, $\beta_0 = 0$, and $\beta_1 = 1$ we get the implicit Euler method

$$y_{\ell+1} = y_\ell + hf\left(t_{\ell+1}, y_{\ell+1}\right), \qquad \ell = 0, \dots, N-1,$$

and for $k = 1$, $\alpha_0 = -1$, $\beta_0 = 1$, and $\beta_1 = 0$ the explicit Euler method

$$y_{\ell+1} = y_\ell + hf\left(t_\ell, y_\ell\right), \qquad \ell = 0, \dots, N-1.$$

REMARK 5.2. Linear multi-step methods are explicit, if $\beta_k = 0$, and implicit otherwise. If they are implicit, existence of a unique solution is guaranteed for sufficiently small h by Banach fixed point theorem, if f is Lipschitz with respect to the second argument. The same holds true for multi-step methods of the form (5.1), if Φ is Lipschitz with respect to the argument $y_{\ell+1}$.

REMARK 5.3. Only the value y_0 will be given for an initial value problem. Hence, in order to start a multi-step method, the data $y_1, \ldots, y_{k-1}$ have to be computed first with a different method.

5.2. Adams methods

Recall for arbitrary $r \in \mathbb{N}_0$ the integral form of the initial value problem

$$y(t_{\ell+1}) = y(t_{\ell-r}) + \int_{t_{\ell-r}}^{t_{\ell+1}} f(\tau, y(\tau))\,\mathrm{d}\tau, \qquad \ell = r, \ldots, N-1. \tag{5.3}$$

Runge-Kutta methods can be motivated by (interpolation) quadrature formulas for the integral with $r = 0$. For those formulas typically quadrature nodes in the interval $[t_{\ell-r}, t_{\ell+1}]$ are used. Instead, we use for arbitrary $r, s, k \in \mathbb{N}_0$ the $k+1$ interpolation points

$$t_{\ell+1-s-j}, \qquad j = 0, \ldots, k.$$

Using the short hand notation

$$f_\ell := f(t_\ell, y_\ell), \qquad \ell = 0, \ldots, N, \tag{5.4}$$

we replace f in the integrand by its interpolation polynomial $p \in \mathbb{P}_k$ given by

$$p = \sum_{j=0}^{k} f_{\ell+1-s-j} L_j, \qquad \text{with } L_j(t) := \prod_{\substack{m=0\\ m\neq j}}^{k} \frac{t - t_{\ell+1-s-m}}{t_{\ell+1-s-j} - t_{\ell+1-s-m}}.$$

Hence, the approximation $y_{\ell+1}$ of $y(t_{\ell+1})$ is given by

$$y_{\ell+1} = y_{\ell-r} + h \sum_{j=0}^{k} b_j f\left(t_{\ell+1-s-j}, y_{\ell+1-s-j}\right) \tag{5.5a}$$

with

$$b_j := \frac{1}{h} \int_{t_{\ell-r}}^{t_{\ell+1}} L_j(\tau)\,\mathrm{d}\tau = \int_{-r}^{1} L_j(t_\ell + \tilde{\tau} h)\,\mathrm{d}\tilde{\tau} = \int_{-r}^{1} \prod_{\substack{m=0\\ m\neq j}}^{k} \frac{\tilde{t} - (1-s-m)}{m-j}\,\mathrm{d}\tilde{\tau}, \tag{5.5b}$$

where we used the uniform mesh $t_\ell := t_0 + \ell h$. Note, that b_j are independent of h and f. They can be computed once for all types of problems and all mesh-sizes h. Using a non-uniform mesh would complicate this a lot, since then b_j would depend on all the different mesh-sizes.

Note, that up to now convergence of such multi-step methods is not guaranteed. We have to discuss the consistency of a multi-step method. Moreover, in contrast to one-step methods, stability of multi-step methods is non-trivial. We will show later, that an additional stability condition is needed for convergence.

REMARK 5.4. (5.5) is for all $r, s, k \in \mathbb{N}_0$ a linear multi-step method of the form (5.2) with $m := \min\{r+1, s+k\}$ steps,

$$\alpha_m = 1, \qquad \alpha_{m-1-r} = -1, \qquad \forall j \in \{0, \ldots, m-1\} \setminus \{r+1\} : \alpha_{m-j} = 0$$

and

$$\forall j \in \{0, \dots, s-1\} : \beta_{m-j} = 0, \qquad \forall j \in \{s, \dots, m\} : \beta_{m-j} = b_{j-s}.$$

Hence, (5.5) is explicit for $s \geq 1$.

Example 5.5 (Adams-Bashforth methods). For $s = 1$ (explicit) and $r = 0$ the so-called *Adams-Bashforth methods* are

$$\begin{aligned}
k = 0: &\qquad y_{\ell+1} = y_\ell + h f_\ell && \text{(explicit Euler method)}\\
k = 1: &\qquad y_{\ell+1} = y_\ell + \frac{h}{2}(3f_\ell - f_{\ell-1})\\
k = 2: &\qquad y_{\ell+1} = y_\ell + \frac{h}{12}(23f_\ell - 16f_{\ell-1} + 5f_{\ell-2})\\
k = 3: &\qquad y_{\ell+1} = y_\ell + \frac{h}{24}(55f_\ell - 59f_{\ell-1} + 37f_{\ell-2} - 9f_{\ell-3})
\end{aligned}$$

Example 5.6 (Adams-Moulton methods). For $s = 0$ (implicit) and $r = 0$ the so-called *Adams-Moulton methods* are

$$\begin{aligned}
k = 0: &\qquad y_{\ell+1} = y_\ell + h f_{\ell+1} && \text{(implicit Euler method)}\\
k = 1: &\qquad y_{\ell+1} = y_\ell + \frac{h}{2}(f_{\ell+1} + f_\ell) && \text{(trapezoidal rule)}\\
k = 2: &\qquad y_{\ell+1} = y_\ell + \frac{h}{12}(5f_{\ell+1} + 8f_\ell - f_{\ell-1})\\
k = 3: &\qquad y_{\ell+1} = y_\ell + \frac{h}{24}(9f_{\ell+1} + 19f_\ell - 5f_{\ell-1} + 1f_{\ell-2})
\end{aligned}$$

Example 5.7 (Nyström method). For $s = 1$ (explicit)and $r = 1$ the methods are called *Nyström methods*. For $k = 0$ we have the explicit midpoint rule

$$y_{\ell+1} = y_{\ell-1} + 2hf_\ell.$$

Remark 5.8. Adams methods rely on the integral form of the initial value problem and Lagrange interpolation of the right hand side f. This is not the only possible construction of multi-step methods. One could first use Lagrange interpolation at the points $(t_{\ell+1-j}, y_{\ell+1-j})$ for $j = 0, \dots, k$. The unknown value $y_{\ell+1}$ can be fixed with the assumption, that the interpolation polynomial should satisfy the differential equation at $t_{\ell+1}$. Combinations of these approaches are possible as well.

5.3. Consistency

Similar to one-step method we first introduce the consistency error, i.e. the local error if one step of the method is performed using exact initial data.

Definition 5.9 (Consistency of multi-step methods). Let (α, Φ) as in Def. 5.1 be a k-step method and y be the exact solution of $y'(t) = f(t, y(t))$ with $y(t_0) = y_0$. Let $y_{\ell+1}$ with $\ell \in \{k-1, \dots, N-1\}$ be for sufficiently small $h > 0$ the unique solution to (5.1) with $y_{\ell-j} := y(t_{\ell-j})$ for $j = 0, \dots, k-1$.

Then we call

$$\tau_\ell(h) := y(t_{\ell+1}) - y_{\ell+1}, \qquad \ell = k-1, \dots, N-1, \tag{5.6}$$

the **consistency error** of the multi-step method. For $p \geq 1$, we say that the k-step method has **consistency order** p, if for all $f \in C^p([t_0, T] \times \mathbb{R}^n; \mathbb{R}^n)$ and hence $y \in$

$C^{p+1}([t_0,T];\mathbb{R}^n)$ there exist $C > 0$ and $h_0 > 0$ such that

$$\forall h \in (0, h_0): \qquad \sup_{\ell=k-1,\dots,N-1} \|\tau_\ell(h)\| \le C\, h^{p+1}. \tag{5.7}$$

In the literature, often the consistency error is defined using the following so called truncation error, since the latter is easier to handle. We will show, that both definitions are somehow equivalent.

Definition 5.10 (Truncation error). Let (α, Φ) as in Def. 5.1 be a k-step method. The for all $y \in C([t_0,T];\mathbb{R}^n)$, all $h > 0$, and all $\ell = k-1, \dots, N-1$ we call

$$\eta_\ell(y,h) := h\Phi(t_\ell, y\left(t_{\ell+1}\right), \dots, y\left(t_{\ell+1-k}\right), h) - \sum_{j=0}^{k} \alpha_{k-j} y\left(t_{\ell+1-j}\right), \quad \ell = k-1, \dots, N-1, \tag{5.8}$$

the **truncation error** of the multi-step method.

Remark 5.11. Let us assume that y is the exact solution of the initial value problem. Then

$$\tilde{\eta}_\ell(y,h) := \sum_{j=0}^{k} \{h\beta_j y'(t_{\ell+1-k} + jh) - \alpha_j y(t_{\ell+1-k} + jh)\} \tag{5.9}$$

is the truncation error of a linear k-step method, i.e. $\tilde{\eta}_\ell(y,h) = \eta_\ell(y,h)$. Note, that in Def. 5.9 $y(t_{\ell+1}) \ne y_{\ell+1}$.

Lemma 5.12. *If y is the solution to $y'(t) = f(t, y(t))$ with $y(t_0) = y_0$ and if f is Lipschitz continuous with respect to the second argument, then the truncation error $\eta_\ell(y,h)$ and the consistency error $\tau_\ell(h)$ of linear k-step methods are equivalent, i.e. there exist constants $C_1, C_2, h_0 > 0$ such that*

$$\forall h \in (0, h_0) \quad \forall \ell \in \{k-1, \dots, N-1\}: \quad C_1\|\tau_\ell(h)\| \le \|\eta_\ell(y,h)\| \le C_2\|\tau_\ell(h)\|. \tag{5.10}$$

Proof. Let $\ell \in \{k-1, \dots, N-1\}$. Since $y_{\ell+1}$ is the solution to (5.2), we have

$$\left(h\beta_k f\left(t_{\ell+1}, y_{\ell+1}\right) - \alpha_k y_{\ell+1}\right) + \sum_{j=1}^{k} \left\{h\beta_{k-j} f\left(t_{\ell+1-j}, y_{\ell+1-j}\right) - \alpha_{k-j} y_{\ell+1-j}\right\} = 0$$

By (5.9) and since $y_{\ell+1-j} = y(t_{\ell+1-j})$ for $j = 1, \dots, k$ we derive

$$\left(h\beta_k f\left(t_{\ell+1}, y_{\ell+1}\right) - \alpha_k y_{\ell+1}\right) + \eta_l(y,h) - \left(h\beta_k f\left(t_{\ell+1}, y(t_{\ell+1})\right) - \alpha_k y(t_{\ell+1})\right) = 0.$$

Hence, using $\alpha_k = 1$

$$\tau_\ell(h) = y(t_{\ell+1}) - y_{\ell+1} = h\beta_k \left\{f\left(t_{\ell+1}, y(t_{\ell+1})\right) - f\left(t_{\ell+1}, y_{\ell+1}\right)\right\} - \eta_\ell(y,h).$$

Lipschitz continuity of f together with the triangle inequality yield the claim

$$\left(1 - h|\beta_k|L\right)\|\tau_\ell(h)\| \le \|\eta_\ell(y,h)\| \le \left(1 + h|\beta_k|L\right)\|\tau_\ell(h)\|.$$

□

Corollary 5.13. *Lemma 5.12 holds for arbitrary k-step methods (α, Φ) as well, if Φ is Lipschitz with respect to the argument $y_{\ell+1}$.*

Proof. The proof is a variant of the last one with minor modifications. □

Definition 5.14 (characteristic polynomials). We define for a linear k-step method of the form (5.2) the characteristic polynomials $\rho, \sigma \in \mathbb{P}_k$ by

$$\rho(\zeta) := \sum_{j=0}^{k} \alpha_j \zeta^j, \qquad \sigma(\zeta) := \sum_{j=0}^{k} \beta_j \zeta^j. \tag{5.11}$$

Theorem 5.15. *Consider a linear k-step method of the form* (5.2) *with sufficiently small mesh-size h. Then the following properties are equivalent:*

(i) *The method has consistency order $p \geq 1$.*

(ii) *The error defined by* (5.9) *vanishes for all polynomials $q \in \mathbb{P}_p$, i.e.*

$$\forall q \in \mathbb{P}_p \quad \forall \ell \in \{k-1, \dots, N-1\} : \qquad \tilde{\eta}_\ell(q, h) = 0.$$

(iii) *There holds $\sum_{j=0}^{k} \alpha_j = 0$ and (using $0^0 = 1$)*

$$\sum_{j=0}^{k} \alpha_j j^i = i \sum_{j=0}^{k} \beta_j j^{i-1}, \qquad i = 1, \dots, p.$$

(iv) $\rho(\exp(h)) - h\sigma(\exp(h)) = \mathcal{O}(h^{p+1})$ *for $h \to 0$.*

(v) $\rho(\zeta)/\ln\zeta - \sigma(\zeta) = \mathcal{O}(|\zeta - 1|^p)$ *for $\zeta \to 1$.*

Proof. **Step 1 ((i)⇔(iii)):** For the solution $y \in C^{p+1}([t_0, T]; \mathbb{R}^n)$ to $y'(t) = f(t, y(t))$ and $y(t_0) = y_0$ the truncation error is given for $\ell = k-1, \dots, N-1$ by (5.9). Inserting for all $j = 0, \dots, k$ the Taylor expansions around $t_{\ell+1-k}$

$$y(t_{\ell+1-k} + jh) = \sum_{i=0}^{p} \frac{y^{(i)}(t_{\ell+1-k})}{i!}(jh)^i + \mathcal{O}(h^{p+1}),$$

$$y'(t_{\ell+1-k} + jh) = \sum_{i=0}^{p-1} \frac{y^{(i+1)}(t_{\ell+1-k})}{i!}(jh)^i + \mathcal{O}(h^p)$$

in (5.9) leads to

$$\eta_\ell(y, h) = \sum_{j=0}^{k} \left\{ \beta_j \sum_{i=1}^{p} \frac{y^{(i)}(t_{\ell+1-k})}{(i-1)!} j^{i-1} h^i - \alpha_j \sum_{i=0}^{p} \frac{y^{(i)}(t_{\ell+1-k})}{i!}(jh)^i \right\} + \mathcal{O}(h^{p+1})$$

$$= -y(t_{\ell+1-k}) \sum_{j=0}^{k} \alpha_j + \sum_{i=1}^{k} \frac{y^{(i)}(t_{\ell+1-k})}{i!} h^i \sum_{j=0}^{k} \left\{ i\beta_j j^{i-1} - \alpha_j j^i \right\} + \mathcal{O}(h^{p+1}).$$

The claim follows with Lemma 5.12.

Step 2 ((ii)⇔(iii)): This is obvious, since for $q \in \mathbb{P}_p$ we have $q^{(p+1)} \equiv 0$. Hence, the error terms $\mathcal{O}(h^{p+1})$ and $\mathcal{O}(h^p)$ in the last step vanish, if q is used instead of y.

Step 3 ((iii)⇔(iv)): Plugging $y(t) = \exp(t)$ with $t_{\ell+1-k} = 0$ into (5.9)leads to

$$\eta_\ell(y, h) = \sum_{j=0}^{k} \{ h\beta_j \exp(jh) - \alpha_j \exp(jh) \} = h\sigma(\exp(h)) - \rho(\exp(h)).$$

The claim follows similar to the first step.

Step 4 ((iv)⇔(v)): Using $z = \exp(h)$ (iv) is equivalent to

$$\rho(z) - \ln(z)\sigma(z) = \mathcal{O}((\ln(z)^{p+1}) \qquad z \to 1.$$

Taylor expansion of ln around 1

$$\ln(z) = (z-1) + \mathcal{O}((z-1)^2) \qquad z \to 1$$

yields the claim. □

COROLLARY 5.16. *A linear multi-step method is consistent, i.e. it has consistency order* $p \geq 1$, *if*

$$\rho(1) = 0 \qquad \text{and } \rho'(1) = \sigma(1). \tag{5.12}$$

PROOF. This is a direct consequence of Theorem 5.15 (iii) with $p = 1$ and Definition 5.14. □

Note, that Theorem 5.15 (iii) represents a linear system with $p+1$ equations for the $2k+1$ unknowns $\alpha_0, \dots, \alpha_{k-1}, \beta_0, \dots, \beta_k$. In principle, this system can be used to construct k-step methods with high consistency order.

EXAMPLE 5.17. Linear, explicit two-step methods are given by

$$\alpha_2 y_{\ell+1} + \alpha_1 y_\ell + \alpha_0 y_{\ell-1} = h\left(\beta_1 f(t_\ell, y_\ell) + \beta_0 f(t_{\ell-1}, y(t_{\ell-1}))\right)$$

with the conventions $\alpha_2 = 1$ and $|\alpha_0| + |\beta_0| \neq 0$. Using Theorem 5.15 (iii) the only possible method of this form with consistency order 3 is

$$y_{\ell+1} + 4y_\ell - 5y_{\ell-1} = h\left(4f(t_\ell, y_\ell) + 2f(t_{\ell-1}, y(t_{\ell-1}))\right).$$

COROLLARY 5.18. *The Adams-Bashforth methods of Example 5.5 have* $k+1$ *steps and consistency order* $k+1$.

PROOF. Adams-Bashforth methods are explicit Adams methods as presented in Section 5.5.2 with $s = 1$ and $r = 0$. Hence, following Remark 5.4 they have $k+1$ steps. Since they are constructed using polynomial interpolation with $k+1$ interpolation nodes applied to the function $f(t, y(t)) = y'(t)$, they are by construction exact if $y \in \mathbb{P}_{k+1}$. Hence, the claim follows with Theorem 5.15 (ii). □

With the same arguments, the following corollary holds true.

COROLLARY 5.19. *The Adams-Moulton methods of Example 5.6 have* $\min\{1, k\}$ *steps and consistency order* $k+1$.

5.4. Linear difference equation

Recall the general form of a linear k-step method in (5.2). For $h \to 0$ (hence $N(h) \to \infty$) the right hand side vanishes and we are led to the linear difference equation

$$\alpha_k y_{\ell+k} + \alpha_{k+1} y_{\ell+k-1} + \cdots + \alpha_0 y_\ell = 0, \qquad \ell \in \mathbb{N}_0, \tag{5.13}$$

for the sequence $(y_\ell)_{\ell \in \mathbb{N}_0} \subset \mathbb{R}^n$. Again, we normalize $\alpha_k = 1$ and claim $\alpha_0 \neq 0$. To simplify the presentation, we assume in this subsection $(y_\ell)_{\ell \in \mathbb{N}_0} \subset \mathbb{C}$ and show, that solutions to (5.13) are determined by the roots of the first characteristic polynomial

$$\rho(\lambda) := \sum_{j=0}^{k} \alpha_j \lambda^j.$$

These solutions will lead us to the so-called root condition, which is necessary to ensure stability of a multi-step method.

LEMMA 5.20. *Each polynomial $p \in \mathbb{P}_k$ of the form $p(t) = \sum_{j=0}^{k-1} a_j t^j + t^k$ is (up to the sign $\pm$) the characteristic polynomial of the companion matrix*

$$A_p := \begin{pmatrix} 0 & \cdots & \cdots & 0 & -a_0 \\ 1 & \ddots & & \vdots & -a_1 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & -a_{k-2} \\ 0 & \cdots & 0 & 1 & -a_{k-1} \end{pmatrix} \in \mathbb{R}^{k \times k}.$$

Hence, the roots of p are the eigenvalues of A_p.

PROOF. This can be shown expanding $\det(A - \lambda \,\mathrm{id})$ iteratively by the last rows. □

REMARK 5.21. Since $\alpha_k = 1$, the solution $(y_\ell)_{\ell \in \mathbb{N}_0} \subset \mathbb{C}$ to (5.13) is uniquely determined, if $y_0, \ldots, y_{k-1}$ are given. Using the linearity of (5.13) the space of solutions is a linear space with dimension k.

LEMMA 5.22. *Let the characteristic polynomial ρ associated with* (5.13) *have $\lambda_1, \ldots, \lambda_k$ piece wise different roots. Then all solutions $(y_\ell)_{\ell \in \mathbb{N}_0} \subset \mathbb{C}$ to* (5.13) *are given by*

$$y_\ell = \sum_{n=1}^{k} c_n \lambda_n^\ell, \qquad \ell \in \mathbb{N}_0, \tag{5.14}$$

with constants $c_1, \ldots, c_k \in \mathbb{C}$.

PROOF. Since $\alpha_0 \neq 0$, $\lambda = 0$ is no root of ρ. Hence, the sequences $(\lambda_n^\ell)_{\ell \in \mathbb{N}_0}$ are for $n = 1, \ldots, k$ well defined and solutions to (5.13):

$$\sum_{j=0}^{k} \alpha_j \lambda_n^{\ell+j} = \lambda_n^\ell \rho(\lambda_n) = 0, \qquad \ell \in \mathbb{N}_0.$$

Using the last remark, we only have to show, that (λ_n^ℓ) for $n = 1, \ldots, k$ are linearly independent. In other words we have to prove, that if

$$\forall \ell \in \mathbb{N}_0 : \quad \sum_{n=1}^{k} c_n \lambda_n^\ell = 0, \tag{5.15}$$

then $c_1 = \cdots = c_k = 0$. This is obvious, since for $\ell = 0, \ldots, k-1$ (5.15) is the linear system of equations

$$\begin{pmatrix} \lambda_1^0 & \ldots & \lambda_1^{k-1} \\ \vdots & \ddots & \vdots \\ \lambda_k^0 & \ldots & \lambda_k^{k-1} \end{pmatrix} \begin{pmatrix} c_1 \\ \vdots \\ c_k \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Note, that the system matrix is the Vandermonde matrix, which is regular for piece wise different nodes $\lambda_1, \ldots, \lambda_k$. □

LEMMA 5.23. *Let the characteristic polynomial ρ associated with* (5.13) *has the root λ with multiplicity $\mu > 1$. Then the sequences $(\lambda_\ell^{(m)})_{\ell \in \mathbb{N}_0}$ defined by*

$$\forall \ell \in \{0, \ldots, m-1\} : \lambda_\ell^{(m)} := 0, \qquad \forall \ell \geq m : \lambda_\ell^{(m)} := \frac{\ell!}{(\ell - m)!} \lambda^{\ell - m}$$

for $m = 0, \ldots, \mu - 1$, are linearly independent solutions to (5.13). *Note, that we use the convention $0! = 1$.*

Proof. The proof is a variant of the preceding one. Since the root λ of ρ has multiplicity $\mu > 1$, i.e. $\rho^{(m)}(\lambda) = 0$ for $m = 0, \ldots, \mu - 1$, the same holds true for the functions

$$\rho_\ell(\lambda) := \lambda^\ell \rho(\lambda) = \sum_{j=0}^{k} \alpha_j \lambda^{\ell+j}.$$

Hence, $(\lambda_\ell^{(m)})_{\ell\in\mathbb{N}_0}$ are for $m = 0, \ldots, \mu - 1$ solutions to (5.13):

$$0 = \rho_\ell^{(m)}(\lambda) = \sum_{j=0}^{k} \alpha_j \frac{(\ell+j)!}{(\ell+j-m)!} \lambda^{\ell+j-m} = \sum_{j=0}^{k} \alpha_j \lambda_{\ell+j}^{(m)}.$$

To show linear independency, we assume

$$\forall \ell \in \mathbb{N}_0 : \quad \sum_{m=0}^{\mu-1} c_m \lambda_\ell^{(m)} = 0$$

and use again $\ell = 0, \ldots, \mu - 1$ leading to the linear system of equations

$$\begin{pmatrix} \lambda^0 & 0 & \dots & \dots & 0 \\ \lambda^1 & 1\lambda^0 & \ddots & & \vdots \\ \lambda^2 & 2\lambda^1 & 2\lambda^0 & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ \lambda^{\mu-1} & (\mu-1)\lambda^{\mu-2} & \dots & & (\mu-1)!\lambda^0 \end{pmatrix} \begin{pmatrix} c_0 \\ \vdots \\ c_{\mu-1} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{5.16}$$

Since $\lambda \neq 0$, the system matrix is regular and $c_0 = \cdots = c_{\mu-1}$. □

Remark 5.24. The system matrix of the last proof is related to Hermite interpolation in the following sense: Let $p \in \mathbb{P}_{\mu-1}$ be the solution to

$$p^{(m)}(\lambda) = f_m, \qquad m = 0, \ldots, \mu - 1.$$

Using the monomial basis of $\mathbb{P}_{\mu-1}$, i.e. $p(z) = \sum_{m=0}^{\mu-1} c_m z^m$, leads to the transpose of the system matrix in (5.16).

Theorem 5.25. *All solutions to* (5.13) *are given by*

$$y_\ell = \sum_{n=1}^{\tilde{k}} \sum_{m=0}^{\mu_n-1} \alpha_{n,m} \lambda_\ell^{(n,m)}, \qquad \ell \in \mathbb{N}_0, \tag{5.17}$$

where $\alpha_{n,m} \in \mathbb{C}$ are arbitrary constants, and $\lambda_1, \ldots, \lambda_{\tilde{k}}$ are the piece wise different roots of ρ with corresponding multiplicities $\mu_1, \ldots, \mu_{\tilde{k}} \geq 1$. Moreover, we define the linearly independent sequences $\left(\lambda_\ell^{(n,m)}\right)_{\ell\in\mathbb{N}_0}$ by

$$\forall \ell \in \{0, \ldots, m-1\} : \lambda_\ell^{(n,m)} := 0, \qquad \forall \ell \geq m : \lambda_\ell^{(n,m)} := \frac{\ell!}{(\ell-m)!} \lambda_n^{\ell-m}.$$

Proof. The theorem is a consequence of the fundamental theorem of algebra and a combination of the preceding proofs. To show linear independency of $\left(\lambda_\ell^{(n,m)}\right)_{\ell\in\mathbb{N}_0}$, we can proceed as in the last proofs. Similar to the last remark, the resulting system matrix is related to the Hermite interpolation for $p \in \mathbb{P}_{k-1}$:

$$\forall n = 1, \ldots, \tilde{k}, \quad \forall m = 0, \ldots, \mu_n - 1 : \qquad p^{(m)}(\lambda_n) = f_{n,m}.$$

Hence, $\left(\lambda_\ell^{(n,m)}\right)_{\ell\in\mathbb{N}_0}$ are linearly independent, since the underlying Hermite interpolation problem can be shown to be uniquely solvable. □

5.5. Stability

Based on the results in the last section, we define the so called root condition to ensure stability of a linear multi-step method.

DEFINITION 5.26 (root condition). The linear k-step method of the form (5.2) satisfies the *root condition*, if all roots λ of the first characteristic polynomial $\rho(\zeta) = \sum_{j=0}^k \alpha_j\zeta^j$ satisfy $|\lambda| \le 1$. Moreover, λ has to be simple if $|\lambda| = 1$.

COROLLARY 5.27. *If the linear k-step method of the form* (5.2) *satisfies the root condition, then all solutions $(y_\ell)_{\ell\in\mathbb{N}_0}$ to the linear difference equation* (5.13) *are bounded, i.e. there exists $C > 0$ such that*

$$\forall \ell \in \mathbb{N}_0 : \quad \|y_\ell\| \le C.$$

Moreover, if the linear k-step method violates the root condition, then there exist initial values $y_0, \dots, y_{k-1}$ such that

$$\lim_{\ell\to\infty} \|y_\ell\| = \infty,$$

whereas $(y_\ell)_{\ell\in\mathbb{N}_0}$ is the unique solution to (5.13) *with these initial values.*

PROOF. The corollary is a direct consequence of the last theorem. □

Be aware, that the constants $\alpha_0, \dots, \alpha_k$ are independent of the differential equation to solve. Of course, the argument $h \to 0$ in the beginning of the last subsection does not replace a rigorous proof. But at least heuristically, using a linear multi-step method we have to expect for all different right hand sides increasing solutions, if the method violates the root condition.

EXAMPLE 5.28. Consider the explicit, linear 2-step method of Example 5.17

$$y_{\ell+1} + 4y_\ell - 5y_{\ell-1} = h\left(4f(t_\ell, y_\ell) + 2f(t_{\ell-1}, y(t_{\ell-1}))\right). \tag{5.18}$$

It violates the root condition, since the root of the first characteristic polynomial are $\lambda_1 = 1$ and $\lambda_2 = -5$. If we apply this method to the simple differential equation $y' = \lambda y$ with $\lambda \in \mathbb{R}$, then we get the linear difference equation

$$y_{\ell+1} + 4(1-\lambda h)y_\ell - (5 + 2\lambda h)y_{\ell-1} = 0$$

The roots of this equation are

$$\lambda_1 = 1 + \lambda h + \mathcal{O}((\lambda h)^2), \quad \lambda_2 = -5 - \lambda h + \mathcal{O}((\lambda h)^2) \qquad \lambda h \to 0.$$

Hence, for $\lambda \le 0$ the first solution component $\left(\lambda_1^\ell\right)_{\ell\in\mathbb{N}_0}$ is bounded for sufficiently small h. But the absolute value of the second component $\left(\lambda_2^\ell\right)_{\ell\in\mathbb{N}_0}$ increases exponentially even for small h. There is no chance of convergence for this method.

LEMMA 5.29. *The Adams methods of Section 5.5.2 satisfy the root condition Def. 5.26.*

PROOF. In Remark 5.4 we have shown, that the first characteristic polynomial of a m-step Adams method is given by

$$\rho(\zeta) = \zeta^{m+1} - \zeta^{m-1-r} = \zeta^{m-1-r}\left(\zeta^{r+1} - 1\right).$$

Hence, the roots are $\lambda = 0$ with multiplicity $m-1-r$ and the $r+1$ simple roots of unity $\lambda_k = \exp\left(\frac{2k\pi\mathrm{i}}{r+1}\right)$ for $k = 0, \dots, r$. □

Theorem 5.30 (First Dahlquist barrier). *If a linear k-step method satisfies the root condition Def. 5.26, then its consistency order p satisfies*

(i) $p \leq k$ *if* ${}^{\beta_k}/_{\alpha_k} \leq 0$,
(ii) $p \leq k+1$ *if k is odd,*
(iii) $p \leq k+2$ *if k is even.*

Proof. The proof can be found in [**HNW93**, Chapter III.3, Theorem 3.5]. □

Note, that by Cor. 5.18 and Lemma 5.29 the Adams-Bashforth methods of Example 5.5 are in this sense optimal.

5.6. Convergence

In this section we prove, that a multi-step methods converges with order p if and only if it has consistency order p and satisfies the root condition. But first, we need some auxiliary results.

Definition 5.31 (Convergence of multi-step methods). We say that a k-step method as defined in Def. 5.1 has convergence order p, if for all $f \in C^p([t_0,T] \times \mathbb{R}^n; \mathbb{R}^n)$ and hence all solutions $y \in C^{p+1}([t_0,T];\mathbb{R}^n)$ to $y'(t) = f(t,y(t))$ with $y(t_0) = y_0$, and all initial data $y_1, \dots, y_{k-1}$, there exist constants $C, h_0 > 0$ such that for all $h \in (0,h_0)$, and all $\ell = k-1, \dots, N$ there exist unique solutions $y_{\ell+1}$ to (5.1) and

$$\max_{\ell=0,\dots,N} \|y_\ell - y(t_\ell)\| \leq C \left(h^p + \max_{\ell=0,\dots,k-1} \|y_\ell - y(t_\ell)\| \right). \tag{5.19}$$

Remark 5.32. For the following lemma we need amongst others the Jordan canonical form of a matrix $A \in \mathbb{R}^{n\times n}$. With $\sigma(A) \subset \mathbb{C}$ we denote the set of all eigenvalues, and with

$$\rho(A) := \max_{\lambda\in\sigma(A)} |\lambda|$$

the spectral radius. Moreover, an eigenvalue is called semi-simple, if the algebraic and geometric multiplicity coincide. Now, let $\lambda_1, \dots, \lambda_r \in \sigma(A)$ be the semi-simple eigenvalues of A and $\lambda_{r+1}, \dots, \lambda_m \in \sigma(A)$ be the other eigenvalues. Then there exists a regular matrix $T \in \mathbb{C}^{n\times n}$ such that $A = T^{-1}JT$ and

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix}$$

with the Jordan blocks

$$\forall \ell = 1,\dots,r : J_\ell := \begin{pmatrix} \lambda_\ell & & \\ & \ddots & \\ & & \lambda_\ell \end{pmatrix} \quad \text{and} \quad \forall \ell = r{+}1,\dots,m : J_\ell := \begin{pmatrix} \lambda_\ell & 1 & & \\ & \lambda_\ell & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_\ell \end{pmatrix}.$$

Moreover, we recall that the row sum norm of A is the operator norm induced by the supremum norm, i.e.

$$\|A\|_\infty := \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{i=1,\dots,n} \sum_{j=1}^n |A_{ij}|.$$

Finally, for all operator norms $\|\cdot\|$ there holds

$$\forall A, B \in \mathbb{R}^{n\times n} : \qquad \|AB\| \leq \|A\| \, \|B\|.$$

Hence, for all $k \in \mathbb{N}$ using $A^k = T^{-1}J^kT$ and $J^k = TA^kT^{-1}$ there holds

$$C^{-1}\|J^k\| \leq \left\|A^k\right\| \leq C\|J^k\|$$

with $C := \|T^{-1}\|\,\|T\| > 0$.

LEMMA 5.33. *For $A \in \mathbb{R}^{n\times n}$ the following statements are equivalent:*

(i) $\sup\limits_{k\in\mathbb{N}} \left\|A^k\right\| < \infty$

(ii) $\rho(A) \leq 1$ *and all* $\lambda \in \sigma(A)$ *with* $|\lambda| = 1$ *are semi-simple.*

PROOF. We use the notations of the preceding remark.

Step 1 ((ii)⇒(i)): We chose $\epsilon > 0$ such that $|\lambda_\ell|+\epsilon \leq 1$ for all $\ell = r+1, \ldots, m$. Using the diagonal matrix $D := \operatorname{diag}(1, \epsilon, \ldots, \epsilon^{n-1}) \in \mathbb{R}^{n\times n}$ we define $\hat{J} := D^{-1}JD$, which is still a block diagonal matrix with the blocks $\hat{J}_\ell$ given by

$$\forall \ell = 1, \ldots, r : \hat{J}_\ell = J_\ell \quad \text{and} \quad \forall \ell = r+1, \ldots, m : \hat{J}_\ell = \begin{pmatrix} \lambda_\ell & \epsilon & & \\ & \lambda_\ell & \ddots & \\ & & \ddots & \epsilon \\ & & & \lambda_\ell \end{pmatrix}.$$

Hence, $\|\hat{J}\|_\infty = \max_{\ell=1,\ldots,m} \|\hat{J}_\ell\|_\infty \leq 1$ leading to

$$\left\|A^k\right\|_\infty = \left\|\left(T^{-1}D\hat{J}D^{-1}T\right)^k\right\|_\infty \leq C\underbrace{\|D\|_\infty}_{\leq 1}\underbrace{\left\|D^{-1}\right\|_\infty}_{\leq \epsilon^{-n+1}}\left\|\hat{J}^k\right\|_\infty \leq \frac{C}{\epsilon^{n-1}} < \infty$$

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

Step 2 ((i)⇒(ii)): We prove by contraposition

$$(A \Rightarrow B \wedge C) \Leftrightarrow (\neg(B \wedge C) \Rightarrow \neg A) \Leftrightarrow ((\neg B \Rightarrow \neg A) \vee (\neg C \Rightarrow \neg A)).$$

Step 2a ($\rho(A) > 1 \Rightarrow \lim\limits_{k\to\infty} \|A^k\| = \infty$): There exists a $\lambda_{\hat{\ell}}$ with $|\lambda_{\hat{\ell}}| > 1$. If $\hat{\ell} \leq r$, then $\|J^k_{\hat{\ell}}\|_\infty = |\lambda_{\hat{\ell}}|^k$. Otherwise, we use $y := (1, 0, \ldots, 0)^\top$ with $J^k_{\hat{\ell}}y = (\lambda^k_{\hat{\ell}}, 0, \ldots, 0)^\top$ to compute

$$\left\|J^k_{\hat{\ell}}\right\|_\infty = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\left\|J^k_{\hat{\ell}}x\right\|_\infty}{\|x\|_\infty} \geq \frac{\left\|J^k_{\hat{\ell}}y\right\|_\infty}{\|y\|_\infty} = |\lambda_{\hat{\ell}}|^k \to \infty \quad \text{for } k \to \infty.$$

Hence, the claim follows with

$$C^{-1}\left\|J^k_{\hat{\ell}}\right\|_\infty \leq C^{-1}\max_{\ell=1,\ldots,m}\left\|J^k_\ell\right\|_\infty = C^{-1}\left\|J^k\right\|_\infty \leq \left\|A^k\right\|_\infty.$$

Step 2b ($\exists \lambda \in \sigma(A)$ with $|\lambda| = 1$, which is not semi-simple $\Rightarrow \lim\limits_{k\to\infty} \|A^k\| = \infty$): There exists a $\lambda_{\hat{\ell}}$ with $|\lambda_{\hat{\ell}}| = 1$ and $\hat{\ell} > r$. By induction for $k \in \mathbb{N}$ it can be shown that $J^k_{\hat{\ell}}y = (k\lambda^{k-1}_{\hat{\ell}}, \lambda^k_{\hat{\ell}}, 0, \ldots, 0)^\top$ for $y := (0, 1, 0, \ldots, 0)^\top$. Hence, the claim follows with the same argument as in the step **Step 2a**, since $\|J^k_{\hat{\ell}}\|_\infty \geq k \to \infty$ for $k \to \infty$. □

Compare the following lemma with Lemma 2.11,

LEMMA 5.34 (Gronwall). *Let $A \geq 0$, and $a_\ell \geq 0$ such that*

$$\forall \ell \in \mathbb{N}_0 : \quad a_\ell \leq A\sum_{j=0}^{\ell-1} a_j + |a_0|.$$

Then

$$\forall \ell \in \mathbb{N}_0 : \quad a_\ell \leq (1+A)^\ell |a_0| \leq \exp(A\ell)|a_0|.$$

PROOF. The second inequality is a consequence of $1 + A \leq \exp(A)$. The first inequality is trivial for $A = 0$. For $A > 0$, the inequality can be shown by induction using

$$\forall \ell \in \mathbb{N}_0 : \qquad \sum_{j=0}^{\ell-1} (1+A)^j = \frac{(1+A)^\ell - 1}{A}.$$

□

THEOREM 5.35. *A k-step method as in Def. 5.1 converges with order p (see Def. 5.31), if the following properties hold:*

(i) *it has consistency order p (see Def. 5.9),*
(ii) *it satisfies the root condition Def. 5.26,*
(iii) *Φ is Lipschitz with respect to all arguments $y_{\ell+1-j}$ for $j = 0, \ldots, k$, i.e.*

$$\|\Phi(t, y_k, \ldots, y_0, h) - \Phi(t, \tilde{y}_k, \ldots, \tilde{y}_0, h)\| \leq L \sum_{j=0}^{k} \|y_j - \tilde{y}_j\|\,.$$

(iii) *is always satisfied for linear k-step methods.*

PROOF. To simplify the notation, we present the proof only for the scalar case $n = 1$. Let $f \in C^p([t_0, T] \times \mathbb{R}; \mathbb{R})$ and hence the solutions $y \in C^{p+1}([t_0, T]; \mathbb{R})$ to $y'(t) = f(t, y(t))$ with $y(t_0) = y_0$. By Rem. 5.2 there exist $\tilde{h}_0 > 0$ such that for all $y_1, \ldots, y_{k-1} \in \mathbb{R}$, all $h \leq \tilde{h}_0$, and all $\ell = k-1, \ldots, N-1$ there exist unique solutions $y_{\ell+1}$ to (5.1).

Step 1: We rewrite the error $e_\ell := y(t_\ell) - y_\ell$ for $\ell = 0, \ldots, N$ of the scalar k-step method as an error

$$E_\ell := \begin{pmatrix} e_{\ell-k+1} \\ \vdots \\ e_\ell \end{pmatrix} \in \mathbb{R}^k, \qquad \ell = k-1, \ldots, N, \tag{5.20}$$

of a vectorial one step method. Using the definition of the truncation error (5.8) we compute for $\ell \geq k-1$

$$\sum_{j=0}^{k} \alpha_{k-j} e_{\ell+1-j} = h\delta_\ell - \eta_\ell(y, h)$$

with

$$\delta_\ell := \Phi(t_\ell, y\left(t_{\ell+1}\right), \ldots, y\left(t_{\ell+1-k}\right), h) - \Phi(t_\ell, y_{\ell+1}, \ldots, y_{\ell+1-k}, h).$$

Hence, since $\alpha_k = 1$,

$$e_{\ell+1} = h\delta_\ell - \eta_\ell(y, h) - \sum_{j=1}^{k} \alpha_{k-j} e_{\ell+1-j},$$

leads to

$$E_{\ell+1} = A_\rho^\top E_\ell + F_\ell, , \qquad \ell = k-1, \ldots, N-1, \tag{5.21}$$

with the transpose of the companion matrix

$$A_\rho := \begin{pmatrix} 0 & \cdots & \cdots & 0 & -\alpha_0 \\ 1 & \ddots & & \vdots & -\alpha_1 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & -\alpha_{k-2} \\ 0 & \cdots & 0 & 1 & -\alpha_{k-1} \end{pmatrix} \in \mathbb{R}^{k\times k}$$

of Lemma 5.20 to the first characteristic polynomial ρ, and

$$F_\ell := (0,\dots,0,h\delta_\ell - \eta_\ell(y,h))^\top \in \mathbb{R}^k.$$

(5.21) is an inhomogeneous, linear difference equation. By induction over ℓ it can be shown that

$$E_{\ell+k-1} = \left(A_\rho^\top\right)^\ell E_{k-1} + \sum_{j=0}^{\ell-1} \left(A_\rho^\top\right)^{\ell-j-1} F_{j+k-1}, \qquad \ell = 0,\dots,N-k+1. \tag{5.22}$$

Step 2 (Bound for $\|F_{\ell+k-1}\|_\infty$): Due to (i), (iii), and Cor. 5.13, there exists a constant $\tilde{C} > 0$ such that

$$|\eta_{\ell+k-1}(y,h)| \le \tilde{C} h^{p+1}.$$

Moreover, (iii) leads to

$$|\delta_{\ell+k-1}| \le L \sum_{j=0}^{k} |e_{\ell+k-j}| \le Lk \underbrace{\max\{|e_{\ell+k-1}|,\dots,|e_\ell|\}}_{\|E_{\ell+k-1}\|_\infty} + L \underbrace{|e_{\ell+k}|}_{\le\|E_{\ell+k}\|_\infty}.$$

Hence

$$\|F_{\ell+k-1}\|_\infty = |h\delta_{\ell+k-1} - \eta_{\ell+k-1}(y,h)| \le \tilde{C}h^{p+1} + hL\left(k\|E_{\ell+k-1}\|_\infty + \|E_{\ell+k}\|_\infty\right).$$

Step 3 (Recursive bound for $\|E_\ell\|_\infty$): Due to (ii) together with Lemma 5.20 and Lemma 5.33, there holds

$$M := \sup_{n\in\mathbb{N}_0} \left\|\left(A_\rho^\top\right)^n\right\| < \infty.$$

Hence, 5.22 yields

$$\|E_{\ell+k-1}\|_\infty \le M\left(\|E_{k-1}\|_\infty + \sum_{j=0}^{\ell-1}\|F_{j+k-1}\|_\infty\right), \qquad \ell = 0,\dots,N-k+1.$$

By Step 2

$$\begin{aligned}\sum_{j=0}^{\ell-1}\|F_{j+k-1}\|_\infty &\le \sum_{j=0}^{\ell-1}\left\{\tilde{C}h^{p+1} + hL\left(k\|E_{j+k-1}\|_\infty + \|E_{j+k}\|_\infty\right)\right\} \\ &\le \tilde{C}\underbrace{\frac{\ell}{N}}_{\le 1}(T-t_0)h^p + hL\left\{\|E_{\ell+k-1}\|_\infty + (k+1)\sum_{j=0}^{\ell-1}\|E_{j+k-1}\|_\infty\right\}.\end{aligned}$$

There exists an $h_0 \le \tilde{h}_0$ such that $1 - hLM \ge 1/2$ for all $h \le h_0$. Thus, for all $\ell = 0,\dots,N-k+1$

$$\|E_{\ell+k-1}\|_\infty \le 2M\left\{\tilde{C}(T-t_0)h^p + \|E_{k-1}\|_\infty\right\} + 2ML(k+1)h\sum_{j=0}^{\ell-1}\|E_{j+k-1}\|_\infty.$$

Step 4 (Bound for $|e_\ell|$): Using Lemma 5.34 with $A := 2ML(k+1)h$, and

$$a_0 := 2M\left\{\tilde{C}(T-t_0)h^p + \|E_{k-1}\|_\infty\right\}, \quad \forall \ell \in \{1,\dots,N-k+1\} : a_\ell := \|E_{\ell+k-1}\|_\infty$$

yields for $\ell = 0,\dots,N-k+1$

$$|y(t_{\ell+k-1}) - y_{\ell+k-1}| = |e_{\ell+k-1}| \le \|E_{\ell+k-1}\|_\infty \le \exp(A\ell)\,|a_0|.$$

Note, that

$$\exp(A\ell) = \exp((2ML(k+1)h\ell) \le \exp(2ML(k+1)(T-t_0))$$

can be bounded uniformly in ℓ. Finally, with

$$\|E_{k-1}\|_\infty = \max_{\ell=0,\dots,k-1} |e_\ell| = \max_{\ell=1,\dots,k-1} \|y_\ell - y(t_\ell)\|$$

we have (5.19) with

$$C := \max\{(2ML(k+1)(T-t_0)),1\}\max\left\{2M\tilde{C}(T-t_0),1\right\}.$$

□

Corollary 5.36. *The Adams-Bashforth methods of Example 5.5 and the Adams-Moulton methods of Example 5.6 have convergence order $k+1$.*

Proof. Cor. 5.18 (Cor. 5.19) with Lemma 5.29 and Theorem 5.35. □

Theorem 5.37. *If a linear k-step method of the form* (5.2) *converges in the sense of Def. 5.31 for the initial value problems $y' = 0$ and $\tilde{y}' = 1$ with $y(0) = \tilde{y}(0) = 0$, then*

(i) *it satisfies the root condition Def. 5.26,*
(ii) *and it has at least the consistency order* 1 *(see Def. 5.9).*

Proof. The unique solution to the initial value problems are $y(t) = 0$ and $\tilde{y}(t) = t$.

Claim (i): Confer with Cor. 5.27. For $f \equiv 0$ a linear k-step method results almost into a linear difference method. Let us assume that there exists a root λ of the first characteristic polynomial with $|\lambda| = 1$ and multiplicity $\mu \ge 2$. Then by Lemma 5.23

$$y_\ell = \sqrt{h}\ell\lambda^{\ell-1}$$

is the solution of (5.2) with starting values $y_0,\dots,y_{k-1}$. Since

$$\lim_{h\to 0}\max_{\ell=1,\dots,k-1} \|y_\ell - y(t_\ell)\| = 0$$

(5.19) implies

$$\lim_{h\to 0}\|y_N - y(T)\| = 0.$$

Note, that $N = {}^{T}\!/_{h}$. This is a contradiction, since

$$|y_N| = \frac{T}{\sqrt{h}} \to \infty \qquad \text{for } h \to 0.$$

Claim (ii): If we apply the method to $f \equiv 1$, we get

$$\sum_{j=0}^{k} \alpha_{k-j} y_{\ell+1-j} = h\sigma(1), \qquad \ell = k-1,\dots,N-1 \tag{5.23}$$

with the second characteristic polynomial $\sigma(\zeta) := \sum_{j=0}^{k}\beta_j\zeta^j$. W.l.o.g. we assume $T = 1$. If the starting values $y_0,\dots,y_{k-1}$ converge, then for all $\epsilon > 0$ by (5.19) there exists a $N_0 \in \mathbb{N}$ with $h_0 := {}^{1}\!/_{N_0}$ such that

$$\forall N \ge N_0 \quad \forall j = 0,\dots,k : \quad |y_{N-j} - 1| < \epsilon,$$

since $t_{N-j} \to 1$ for $h \to 0$. Hence, using

$$\left|\sum_{j=0}^{k} \alpha_j \left(y_{N-j} - 1\right)\right| \leq \left((k+1) \max_{j=0,\dots,k} |\alpha_j|\right) \epsilon$$

leads to

$$|\rho(1)| \leq h\,|\sigma(1)| + |\rho(1) - h\sigma(1)| \leq h\,|\sigma(1)| + \left((k+1) \max_{j=0,\dots,k} |\alpha_j|\right) \epsilon.$$

Since ϵ can be chosen arbitrary small, $\lambda = 1$ is a root of ρ.

From the first step we already know, that $\lambda = 1$ is simple, i.e. $\rho'(1) \neq 0$. The sequence $(y_\ell)_{\ell \in \mathbb{N}_0}$ with

$$y_\ell := h \frac{\sigma(1)}{\rho'(1)} \ell$$

is a solution to (5.23), since

$$\sum_{j=0}^{k} \alpha_{k-j} \left(\ell + 1 - j\right) = (\ell + 1 - k) \sum_{j=0}^{k} \alpha_j + \sum_{j=0}^{k} \alpha_j j = (\ell + 1 - k)\rho(1) + \rho'(1) = \rho'(1).$$

Since the starting values $y_0, \dots, y_{k-1}$ of this sequence converge, we have

$$1 = \lim_{h \to 0} y_N = \lim_{h \to 0} \frac{\sigma(1)}{\rho'(1)} \frac{N}{h} = \frac{\sigma(1)}{\rho'(1)}.$$

Finally, Cor. 5.16 yields the claim. □

In this sense, we have the famous short hand notation

$$\text{Convergence} = \text{Consistency} + \text{Stability}. \tag{5.24}$$

5.7. Remarks to linear multi-step methods

We collect in this section several remarks to linear multi-step methods.

5.7.1. Computation of initial values. As indicated in Remark 5.3, for a k-step method in addition to the initial value y_0 the values $y_1, \dots, y_{k-1}$ are needed. In order to guarantee convergence, by (5.19) these values should converge to $y(t_1), \dots, y(t_{k-1})$ for $h \to 0$. Since we expect an error of the form $\mathcal{O}(h^p)$ and since we need only $k-1$ steps, a one-step method of order $p-1$ would be fine. Such methods provide a local consistency error of the form $\mathcal{O}(h^p)$. For $k-1$ steps, this would lead to an additional error of the same order, i.e. $(k-1)\mathcal{O}(h^p) = \mathcal{O}(h^p)$.

5.7.2. Adaptivity and non-uniform meshes. One main advantage of one-step methods are their flexibility. Different mesh-sizes can be used without any additional effort. For multi-step methods, we have seen for Adams methods in Sec. 5.2 that the computation of the coefficients of the linear multi-step method becomes independent of the mesh-size h and the step index ℓ, if we use uniform meshes. This allows to use such coefficients from textbooks without any problems. Moreover, the presented convergence theory relies on uniform meshes and fixed coefficients.

For non-uniform meshes, in each step the local mesh-sizes have to be used to compute the coefficients in the multi-step method. Moreover, we would need to extend the convergence analysis and in particular a stability condition replacing the root condition. We refer to [**HNW93**, Chap. III.5] for multi-step methods with non-uniform meshes and note, that they are not as common.

Concerning adaptivity, in Sec. 2.6 of Chapter 2 embedded Runge-Kutta methods were introduced. These methods provide with low additional costs two one-step methods with order p and $p+1$, which can be used to compute a local error estimator. This estimator is used to optimize the local mesh-size leading to black box algorithms. The user only has to provide the problem and the desired tolerance.

In principle, the Adams-Bashforth methods of Example 5.5 and the Adams-Moulton methods of Example 5.6 provide explicit and implicit multi-step methods of arbitrary convergence order. Hence, it is easy to construct an error estimator of the form $(p, p+1)$. Nevertheless, if the error estimator indicates that the error is too large, we would need to use a finer mesh. Hence, non-uniform meshes or at least a restart of the multi-step method with a smaller, uniform mesh-size would be necessary.

5.7.3. A and A(α) stability. In Chapter 4 we introduced the concept of A-stability for stiff problems. It relies on linear differential equations. After diagonalization we arrive at the model problem

$$y' = \lambda y, \qquad y(0) = y_0,$$

with $\mathrm{Re}(\lambda) \leq 0$. The solution is given by $y(t) = y_0 \exp(\lambda t)$. Using a linear k-step method of the form (5.2) for this problem leads us to

$$\sum_{j=0}^{k} \alpha_{k-j} y_{\ell+1-j} = z \sum_{j=0}^{k} \beta_{k-j} y_{\ell+1-j}, \qquad \ell = k-1, \ldots, N-1,$$

with $z := \lambda h$.

For $\xi \in \mathbb{C} \setminus \{0\}$ we insert the Ansatz $y_\ell := \xi^\ell$ into the linear difference equation:

$$\xi^{\ell+1-k} \Bigg(\underbrace{\sum_{j=0}^{k} \alpha_{k-j} \xi^{k-j}}_{=\rho(\xi)} - z \underbrace{\sum_{j=0}^{k} \beta_{k-j} \xi^{k-j}}_{=\sigma(\xi)} \Bigg) = 0.$$

Hence, for fixed $z = \lambda h$ the constant ξ_z has to be a root of

$$\rho(\xi_z) = z\sigma(\xi_z).$$

In order to guarantee, that the discrete solution stay bounded, we are led to a similar condition as the root condition of Def 5.26.

Definition 5.38 (Stability domain of a linear multi-step method). Let $\rho, \sigma \in \mathbb{P}_k$ be as in Def 5.14 the characteristic polynomials of a linear k-step method of the form (5.2). Moreover, for $z \in \mathbb{C}$ let ξ_z be the roots of the polynomial $\rho_z \in \mathbb{P}_k$ defined by

$$\rho_z(\zeta) := \rho(\zeta) - z\sigma(\zeta).$$

Then the stability domain $S \subset \mathbb{C}$ of the linear k-step method is defined by

$$S := \{z \in \mathbb{C} : |\xi_z| \leq 1 \quad \wedge \quad \xi_z \text{ is simple if } |\xi_z| = 1\}.$$

Note, that the root condition Def 5.26 is equivalent to $0 \in S$.

Definition 5.39 (A and A(α) stability). A linear multi-step method is called *A stable*, if $\mathbb{C}_{\mathrm{Re}\leq 0} \subset S$. It is called *A($\alpha$) stable*, if

$$\{z \in \mathbb{C} : \arg(z) \in (\pi - \alpha, \pi + \alpha)\} \subset S.$$

A(α) stability implies that the method is stable in a sector of $\mathbb{C}_{\mathrm{Re}\leq 0}$. In practice, this is often enough. E.g. for the discretization of the heat equation in Ex. 4.2 of Chapter 4, all eigenvalues λ are real. So A(α) stability with $\alpha > 0$ would guarantee, that the discrete solutions to this problem stay bounded.

The reason for introducing A(α) stability is simple: It can be shown, that there exist no A-stable linear multi-step method with order $p \geq 3$. But there exist A(α)-stable linear multi-step methods of higher order.

5.7.4. Implicit multi-step methods. Recall, that for implicit methods we have to solve a possibly non-linear system of equations. This could be done with Newton's method. To this end, we define for fixed ℓ, h, and $y_\ell, \dots, y_{\ell+1-k}$, the function $\Psi : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\Psi(y_{\ell+1}) := \sum_{j=0}^{k} \alpha_{k-j} y_{\ell+1-j} - h\Phi(t_\ell, y_{\ell+1}, \dots, y_{\ell+1-k}, h).$$

Newton iteration leads to the sequence $\left(y_{\ell+1}^{(i)}\right)_{i\in\mathbb{N}_0}$ with initial value $y_{\ell+1}^{(0)}$ and

$$y_{\ell+1}^{(i+1)} = y_{\ell+1}^{(i)} - \left(D\Psi(y_{\ell+1}^{(i)})\right)^{-1} \Psi(y_{\ell+1}^{(i)}), \qquad i \in \mathbb{N}.$$

The initial value might be computed with an explicit method or by just using the preceding value y_ℓ. Note, that we have a function evaluation $\Psi(y_{\ell+1}^{(i)})$ and an evaluation of the Jacobi matrix $D\Psi(y_{\ell+1}^{(i)})$ in each step of the Newton iteration. Hence, the computational costs increase with the number of Newton iterations.

Sometimes, to reduce these costs so called predictor-corrector methods are used. These methods consists of three different steps:

- **P** In the predictor step, an explicit method such as an Adams-Bashforth method (see Ex. 5.5) is used to compute the initial value $y_{\ell+1}^{(0)}$.
- **E** In the evaluation step, one function evaluation of the form $f\left(y_{\ell+1}^{(i)}\right)$ is performed.
- **C** In the corrector step, only one step of a fixed point iteration with an (implicit) Adams-Moulton method (see Ex. 5.6) is performed.

The predictor and the corrector step need no function evaluation and of course no evaluation of a Jacobi matrix. For $m \in \mathbb{N}$ these steps can be combined to the predictor-corrector method

$$\mathbf{P}\left(\mathbf{EC}\right)^m \mathbf{E}.$$

Note, that in practice only small values of m such as 1 or 2 are used. Moreover, such a method leads to nested function evaluations. Hence, these methods have some similarities to Runge-Kutta methods.

Exercise 5.40. Let $m^{(P)}$ be the order of the predictor method and $m^{(C)}$ be the order of the corrector method. Then, for $m \in \mathbb{N}$ the order of the predictor-corrector method $\mathbf{P}\left(\mathbf{EC}\right)^m \mathbf{E}$ is

$$p_m := \min\left\{m^{(C)}, m^{(P)} + m\right\}.$$

Note, that implicit methods are typically used for stiff problems. Hence, apart from the order of a predictor-corrector method, m has to be sufficiently large such that the discrete solution stays bounded.

CHAPTER 6

Structure-preserving integrators

The aim in this chapter is to derive numerical methods for autonomous initial value problems

$$y'(t) = f(y(t)), \qquad y(0) = y_0, \tag{6.1}$$

such that some properties of the continuous system are carried over the discrete system. With the concepts of A- and B-stability we have already seen, that this can be useful. Note, that this approach differs from the usual way of developing methods, such that the discrete solution converges in a most efficient way to the continuous solution.

Structure-preserving integrators might be useful, if we are not interested in the solution itself but in some derived quantities such as the energy of the system. Moreover, if for the continuous system we have mass or energy conservation, discrete approximations not preserving the mass or energy can be useless in practice.

6.1. Hamiltonian systems

A family of differential equations with valuable structure are Hamiltonian systems.

Definition 6.1 (Hamiltonian system). Let $H : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and (p, q) with $p, q : \mathbb{R} \to \mathbb{R}^d$ be a solution to the system of differential equations

$$p' = -\nabla_q H(p, q), \tag{6.2a}$$

$$q' = \nabla_p H(p, q). \tag{6.2b}$$

Then we call (6.2) a *Hamiltonian system*, H the *Hamiltonian*, q the (generalized) coordinates, and p the (generalized) momenta.

Remark 6.2. (6.2) is equivalent to

$$\begin{pmatrix} p \\ q \end{pmatrix}'(t) = J^{-1}\nabla_{(p,q)} H(p(t), q(t)) \qquad \text{with } J := \begin{pmatrix} 0 & \mathrm{id}_{d\times d} \\ -\mathrm{id}_{d\times d} & 0 \end{pmatrix}.$$

Note that,

$$J^2 = -\mathrm{id}_{2d\times 2d}, \quad J^\top = -J, \quad J^{-1} = -J, \quad J^{-\top} = J, \quad \det J = 1. \tag{6.3}$$

Proposition 6.3. *Let* $(p, q) \in C^1([0, T], \mathbb{R}^d) \times C^1([0, T], \mathbb{R}^d)$ *be a solution to* (6.2) *with* $H \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R})$. *Then the Hamiltonian is preserved, i.e.*

$$\forall t \in [0, T] : \quad H(p(t), q(t)) = H(p(0), q(0)).$$

Proof. Using the chain rule we compute

$$\begin{aligned} \partial_t H(p(t), q(t)) &= \begin{pmatrix} \nabla_p H(p(t), q(t)) \\ \nabla_q H(p(t), q(t)) \end{pmatrix} \cdot \begin{pmatrix} p'(t) \\ q'(t) \end{pmatrix} \\ &= \begin{pmatrix} \nabla_p H(p(t), q(t)) \\ \nabla_q H(p(t), q(t)) \end{pmatrix} \cdot \begin{pmatrix} -\nabla_q H(p(t), q(t)) \\ \nabla_p H(p(t), q(t)) \end{pmatrix} = 0. \end{aligned}$$

□

Example 6.4 (mathematical pendulum). Mathematically, the motion of a pendulum for small-angle oscillations can be described by

$$mq''(t) + m\frac{g}{l}\sin(q(t)) = 0,$$

whereas q describes the angular displacement, m the mass of the pendulum, g is acceleration due to gravity, l is the length of the pendulum, and t the time. We introduce $p := q'$ and define the energy

$$E(p,q) := \underbrace{\frac{m}{2}l^2p^2}_{\text{kinetic energy}} - \underbrace{mgl\cos q}_{\text{potential energy}}$$

and the Hamiltonian

$$H(p,q) := \frac{1}{ml^2}E(p,q) = \frac{1}{2}p^2 - \frac{g}{l}\cos q.$$

Obviously, (p,q) satisfy (6.2). The last proposition proves, that the energy E is conserved.

Example 6.5 (N-body problem). For $N \in \mathbb{N}$ with $N \geq 2$ we have N bodies with mass $m_i \in \mathbb{R}$, position $q_i \in \mathbb{R}^3$ and momentum $p_i := m_i q_i' \in \mathbb{R}^3$ for $i = 1, \dots, N$. Newton's law of universal gravitation describes the interaction of the gravitational forces leading to the equation of motion

$$\forall i = 1, \dots, N: \quad p_i' = \sum_{\substack{j=1 \\ j \neq i}}^{N} g m_i m_j \frac{q_j - q_i}{\|q_j - q_i\|_2^3},$$

where g is the gravitational constant.

We define $\mathbf{q} := (q_1, \dots, q_N)^\top$, $\mathbf{p} := (p_1, \dots, p_N)^\top$, and the Hamiltonian

$$H(\mathbf{p}, \mathbf{q}) := \frac{1}{2}\sum_{i=1}^{N} \frac{\|p_i\|_2^2}{m_i} - \sum_{i=2}^{N}\sum_{j=1}^{i-1} g\frac{m_i m_j}{\|q_i - q_j\|_2}.$$

A straightforward calculation shows, that $(\mathbf{p}, \mathbf{q})$ satisfy (6.2). Again, the Hamiltonian is preserved due to last proposition.

Exercise 6.6. Show, that not only the Hamiltonian but the total momentum $\sum_{i=1}^{N} p_i$ and the total angular momentum $\sum_{i=1}^{N} q_i \times p_i$ are preserved as well in a N-body problem of the form Ex. 6.5.

Remark 6.7. In both examples, the Hamiltonian is of the form

$$H(p,q) = T(p) + U(q),$$

where T is the kinetic energy, which depends only on the momentum p, and U is the potential energy, which depends only on the coordinates q. Indeed, several Hamiltonian systems have this structure.

In the following we show that energy conservation it not the only useful property of Hamiltonian systems. Even more specific for Hamiltonian systems is that the flow of such a system is a symplectic mapping.

Definition 6.8. For $f \in C([0,T] \times \mathbb{R}^n; \mathbb{R}^n)$ we define the *flow* or *evolution* $\Phi^t : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\forall t \in [0,T] \quad \forall y_0 \in \mathbb{R}^n : \qquad \Phi^t(y_0) := y(t),$$

where y is the solution to the initial value problem $y'(t) = f(t,y)$ with $y(0) = y_0$.

Remark 6.9. Recall, that Rem 3.28 guarantees that $\Phi^t \in C^p(\mathbb{R}^n;\mathbb{R}^n)$ for all $t \in [0,T]$ if $f \in C^p([0,T]\times\mathbb{R}^n;\mathbb{R}^n)$. For $p \geq 1$ let $W(t) := D_{y_0}\Phi^t(y_0)$ be the Wronski matrix of the flow. There holds

$$W'(t) = \partial_t D_{y_0}\Phi^t(y_0) = D_{y_0}y'(t) = D_{y_0}f(y(t)) = D_{y_0}f\left(\Phi^t(y_0)\right) = (Df)(\Phi^t(y_0))W(t),$$

and $W(0) = \mathrm{id}_{n\times n}$. In other words, W is a solution to the initial value problem

$$W'(t) = (Df)(\Phi^t(y_0))W(t), \qquad W(0) = \mathrm{id}_{n\times n}\,. \tag{6.4}$$

In particular, $W(h) = \mathrm{id}_{n\times n} + \mathcal{O}(h)$ for $h \to 0$.

Definition 6.10 (symplectic mapping). Let $J := \begin{pmatrix} 0 & \mathrm{id}_{d\times d} \\ -\mathrm{id}_{d\times d} & 0 \end{pmatrix} \subset \mathbb{R}^{2d\times 2d}$ be as in Remark 6.2. A matrix $A \in \mathbb{R}^{2d\times 2d}$ is called *symplectic*, if

$$A^\top JA = J.$$

A linear mapping $A : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is called *symplectic*, if the matrix representation is symplectic. A mapping $\Phi \in C^1(\mathbb{R}^{2d},\mathbb{R}^{2d})$ is called *symplectic*, if for all $z \in \mathbb{R}^{2d}$ the Jacobian $D\Phi(z) \in \mathbb{R}^{2d\times 2d}$ is symplectic, i.e.

$$\forall z \in \mathbb{R}^{2d}: \quad D\Phi(z)^\top JD\Phi(z) = J.$$

Remark 6.11. If we define the bilinear form $\omega : \mathbb{R}^{2d}\times\mathbb{R}^{2d} \to \mathbb{R}$ by

$$\omega(u,v) := u^\top Jv, \qquad u,v \in \mathbb{R}^{2d},$$

then linear, symplectic mappings A preserve ω, i.e.

$$\forall u,v \in \mathbb{R}^{2d}: \quad \omega(Au,Av) = u^\top A^\top JAv = u^\top Jv = \omega(u,v). \tag{6.5}$$

Exercise 6.12. The set $\mathrm{Sp}_{2d} := \{A \in \mathbb{R}^{2d\times 2d} : A^\top JA = J\}$ is a group under the operation of matrix multiplication, i.e.

(i) $\mathrm{id} \in \mathrm{Sp}_{2d}$,
(ii) $AB \in \mathrm{Sp}_{2d}$ if $A,B \in \mathrm{Sp}_{2d}$, and
(iii) $A^{-1} \in \mathrm{Sp}_{2d}$ if $A \in \mathrm{Sp}_{2d}$.

Sp_{2d} is called *symplectic group*.

Exercise 6.13. Show that, if $\Phi_1,\Phi_2 \in C^1(\mathbb{R}^{2d},\mathbb{R}^{2d})$ are symplectic, then $\Phi_1\circ\Phi_2$ is symplectic.

Theorem 6.14 (Poincare). *Let $H \in C^2(\mathbb{R}^d\times\mathbb{R}^d;\mathbb{R})$ be as in Def. 6.1 the Hamiltonian and $\Phi^t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ the associated flow, i.e. $\Phi^t(y_0) = y(t)$ and $y \in C^2([0,T];\mathbb{R}^{2d})$ is the solution to*

$$y' = J^{-1}\nabla H(y), \qquad y(0) = y_0.$$

Then for all $t \in [0,T]$ the flow $\Phi^t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is symplectic.

Proof. For fixed $y_0 \in \mathbb{R}^{2d}$ we define the mapping $t \mapsto F(t)$ by

$$F(t) := \left(D_{y_0}\Phi^t(y_0)\right)^\top JD_{y_0}\Phi^t(y_0).$$

Obviously, $F(0) = J$, since $\Phi^0(y_0) = y_0$ and therefore $D\Phi^t(y_0) = \mathrm{id}_{2d\times 2d}$. Hence, the claim is proven, if F is constant, i.e. if we can show $F'(t) = 0$.

Since

$$\partial_t\Phi^t(y_0) = y'(t) = J^{-1}\nabla H(y(t)) = J^{-1}\nabla H\left(\Phi^t(y_0)\right),$$

we have by symmetry of second derivatives

$$\partial_t D_{y_0}\Phi^t(y_0) = D_{y_0}\partial_t\Phi^t(y_0) = D_{y_0}\left\{J^{-1}\nabla H\left(\Phi^t(y_0)\right)\right\} = J^{-1}D^2H(\Phi^t(y_0))D_{y_0}\Phi^t(y_0),$$

where D^2H denotes the Hessian matrix of H. To shorten notation, we write in the following D instead of D_{y_0}:

$$\begin{aligned}
F'(t) &= \partial_t \left\{ \left(D\Phi^t(y_0)\right)^\top JD\Phi^t(y_0) \right\} \\
&= \partial_t \left\{ \left(D\Phi^t(y_0)\right)^\top \right\} JD\Phi^t(y_0) + \left(D\Phi^t(y_0)\right)^\top J\partial_t \left\{D\Phi^t(y_0)\right\} \\
&= \left(J^{-1}D^2H(\Phi^t(y_0))D\Phi^t(y_0)\right)^\top JD\Phi^t(y_0) + \left(D\Phi^t(y_0)\right)^\top JJ^{-1}D^2H(\Phi^t(y_0))D\Phi^t(y_0) \\
&= \left(D\Phi^t(y_0)\right)^\top D^2H(\Phi^t(y_0)) \underbrace{J^{-\top}J}_{=-\,\mathrm{id}} D\Phi^t(y_0) + \left(D\Phi^t(y_0)\right)^\top \underbrace{JJ^{-1}}_{=\mathrm{id}} D^2H(\Phi^t(y_0)D\Phi^t(y_0) \\
&= 0.
\end{aligned}$$

□

Indeed, the flow is only symplectic for Hamiltonian systems. So this is a very specific property of such systems. But first, we have to prove the next lemma.

Lemma 6.15 (integrability lemma). *Let $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$. If for all $y \in \mathbb{R}^n$ the Jacobian $Df(y)$ is symmetric, then there exists a $H \in C^2(\mathbb{R}^n;\mathbb{R})$ such that $f = \nabla H$.*

Proof. We define

$$\forall y \in \mathbb{R}^n : \quad H(y) := \int_0^1 f(ty) \cdot y \,\mathrm{d}t.$$

Using $\partial_i f_j = \partial_j f_i$ we compute

$$\begin{aligned}
\partial_i H(y) &= \int_0^1 \partial_i \left\{ \sum_{j=1}^n y_j f_j(ty) \right\} \mathrm{d}t = \int_0^1 \left(f_i(ty) + \sum_{j=1}^n t y_j (\partial_i f_j)(ty) \right) \mathrm{d}t \\
&= \int_0^1 \left(f_i(ty) + \sum_{j=1}^n t y_j (\partial_j f_i)(ty) \right) \mathrm{d}t = \int_0^1 \partial_t \left\{ t f_i(ty) \right\} \mathrm{d}t = f_i(y).
\end{aligned}$$

□

Theorem 6.16. *Let $f \in C^1(\mathbb{R}^{2d};\mathbb{R}^{2d})$ and $y' = f(y)$. Then, the following properties are equivalent:*

(i) *There exists a Hamiltonian $H \in C^2(\mathbb{R}^d \times \mathbb{R}^d;\mathbb{R})$ such that $f := J^{-1}\nabla H$, i.e. we have a Hamiltonian system.*

(ii) *For all $t \in [0,T]$, the flow $\Phi^t : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is symplectic.*

Proof. The first part of the Theorem is already proven in Theorem 6.14. For the other direction let $W(t) := D_{y_0}\Phi^t(y_0)$ be the Wronski matrix of a symplectic flow. Since Φ^t is symplectic for all $t \in [0,T]$, we have for all $y_0 \in \mathbb{R}^{2d}$

$$W(t)^\top JW(t) = J.$$

In particular, the mapping $t \mapsto W(t)^\top JW(t)$ is constant, i.e., with Remark 6.9 there holds

$$\begin{aligned}
0 = \partial_t \left\{ W(t)^\top JW(t) \right\} &= W'(t)^\top JW(t) + W(t)^\top JW'(t) \\
&= W(t)^\top (Df(\Phi^t(y_0)))^\top JW(t) + W(t)^\top JDf(\Phi^t(y_0))W(t).
\end{aligned}$$

Using $W(0) = \mathrm{id}$ and $\Phi^0(y_0) = y_0$, this implies for $t = 0$

$$Df(y_0)^\top J + JDf(y_0) = 0.$$

$J = -J^\top$ and $JDf(y_0) = D(Jf(y_0))$ leads to

$$(D(Jf(y_0)))^\top = D(Jf(y_0)),$$

i.e., the function $y_0 \mapsto Jf(y_0)$ has a symmetric Jacobian $D(Jf(y_0))$. Hence, by Lemma 6.15 there exists $H \in C^2(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R})$ such that $Jf = \nabla H$ and the claim is proven. □

LEMMA 6.17. *If $\Phi \in C^1(\mathbb{R}^{2d}, \mathbb{R}^{2d})$ is symplectic, then*

$$\forall z \in \mathbb{R}^{2d}: \quad |\det D\Phi(z)| = 1.$$

Moreover, if for some $\Omega \subset \mathbb{R}^{2d}$ the mapping $\Phi|_\Omega : \Omega \to \Phi(\Omega) \subset \mathbb{R}^{2d}$ is injective (hence bijective), then

$$|\Omega| := \int_\Omega 1 \,\mathrm{d}x = \int_{\Phi(\Omega)} 1 \,\mathrm{d}x =: |\Phi(\Omega)|,$$

i.e., Φ is area preserving.

PROOF. In (6.3) we have seen that $\det J = 1$. Hence, if Φ is symplectic,

$$1 = \det J = \det\left(D\Phi(z)^\top J D\Phi(z)\right) = \det D\Phi(z)^\top \det J \det D\Phi(z) = (\det D\Phi(z))^2$$

yields the first claim. If Φ is symplectic and injective on Ω, then by the inverse function theorem there exists the inverse $\Phi^{-1} \in C^1(\Phi(\Omega); \Omega)$ and the claim follows by substitution

$$\int_{\Phi(\Omega)} f(x) \,\mathrm{d}x = \int_\Omega f \circ \Phi(x) |\det D\Phi(x)| \,\mathrm{d}x$$

with $f \equiv 1$. □

6.2. Invariants

As we have seen, the Hamiltonian of a Hamiltonian system is preserved in time. Moreover, the total momentum and the total angular momentum of an N-body problem (see Ex. 6.6) are preserved as well. These are examples of so called invariants. Other examples of invariants might be energy or mass conservation.

DEFINITION 6.18. A function $I : \mathbb{R}^n \to \mathbb{R}$ is called *invariant* of the initial value problem

$$\forall t \in [0, T] : y'(t) = f(y(t)), \qquad y(0) = y_0, \tag{6.6}$$

if there holds

$$\forall y_0 \in \mathbb{R}^n \quad \forall t \in [0, T] : I(y_{y_0}(t)) = I(y_0),$$

where $y_{y_0} \in C^1([0, T], \mathbb{R}^n)$ is the solution to (6.6).

PROPOSITION 6.19. *Let $I \in C^1(\mathbb{R}^n, \mathbb{R})$ and $f \in C(\mathbb{R}^n, \mathbb{R}^n)$ such that for all $y_0 \in \mathbb{R}^n$ there exists a (unique) solution to* (6.6)*. Then, I is an invariant if and only if*

$$\forall y \in \mathbb{R}^n : \quad \nabla I(y) \cdot f(y) = 0.$$

PROOF.

$$\partial_t I(y(t)) = \nabla_y I(y(t)) \cdot y'(t) = \nabla_y I(y(t)) \cdot f(y(t)).$$

□

DEFINITION 6.20. An invariant I is called *linear invariant*, if there exist $v \in \mathbb{R}^n$ and $w \in \mathbb{R}$ such that

$$I(y) = v \cdot y + w.$$

An invariant I is called *quadratic invariant*, if there exist $Q \in \mathbb{R}^{n\times n}$, $v \in \mathbb{R}^n$, and $w \in \mathbb{R}$ such that

$$I(y) = \frac{1}{2} y^\top Q y + v \cdot y + w.$$

Examples of linear and quadratic invariants are the total momentum and the total angular momentum of an N-body problem (see Ex. 6.6).

PROPOSITION 6.21.

(i) *All Runge-Kutta methods preserve linear invariants.*
(ii) *Gauss methods as defined in Def. 4.30 preserve quadratic invariants.*

PROOF. **Step** (i)**:** By Prop. 6.19 there holds for linear invariants $v \cdot f(y) = 0$ for all $y \in \mathbb{R}^n$. For autonomous systems, Runge-Kutta methods as defined in Def. 3.9 are given by $y_{\ell+1} := y_\ell + h_\ell \sum_{j=1}^m b_j k_j$ with

$$k_j = f\left(y_\ell + h\sum_{i=1}^m A_{ji} k_i\right), \qquad j = 1, \ldots, m.$$

Hence,

$$I\left(y_{\ell+1}\right) = v \cdot y_\ell + h_\ell \sum_{j=1}^m b_j \underbrace{v \cdot f(\ldots)}_{=0} + w = v \cdot y_\ell + w = I\left(y_\ell\right), \qquad \ell = 0, \ldots, N-1.$$

Step (ii)**:** The proof is a variant of the one to Theorem 4.31 concerning the B-stability of Gauss methods. Let $q \in \mathbb{P}_m$ be the collocation polynomial as defined in (3.25). Then $u : [0,1] \to \mathbb{R}$ defined by $u(t) := I(q(t_\ell + th_\ell))$ is a polynomial of degree $2m$, if I is a quadratic invariant. Since $u' \in \mathbb{P}_{2m-1}$ and since Gauss quadrature formulas with m quadrature nodes are exact for polynomials of degree $2m-1$, we have

$$I(y_{\ell+1}) = u(1) = u(0) + \int_0^1 u'(\tau)\,\mathrm{d}\tau = I(y_\ell) + \sum_{j=1}^m \alpha_j u'(c_j) = I(y_\ell),$$

since

$$u'(c_j) = h_\ell \nabla I(q(t_\ell + c_j h_\ell)) \cdot q'(t_\ell + c_j h_\ell) = h_\ell \nabla I(q(t_\ell + c_j h_\ell)) \cdot f(q(t_\ell + c_j h_\ell)) = 0.$$

□

The last proof indicates that Runge-Kutta methods may not preserve other invariants such as the Hamiltonian in the Examples 6.4 and 6.5. But e.g. for the mathematical pendulum for small-angle oscillations, i.e., small q, the energy is given by

$$E(p,q) = \tilde{E}(p,q) + \mathcal{O}(q^4), \qquad \tilde{E}(p,q) := \frac{m}{2} l^2 p^2 - mgl\left(1 - \frac{1}{2}q^2\right).$$

$\tilde{E}$ is a quadratic functional, which is preserved by Gauss methods. Hence, we expect that Gauss methods almost conserve the energy for long time intervals. Moreover, the proof of (ii) suggests that the Gauss methods are the only collocation methods preserving all quadratic invariants.

6.3. Symplectic integrators

As we have seen, for Hamiltonian systems Def. 6.1 the flow is a symplectic mapping. Hence, we try to construct one-step methods preserving this structure.

Definition 6.22 (discrete flow). For a given one-step method with incremental function Φ we define the *discrete flow* or *discrete evolution* $\Psi^h : \mathbb{R}^n \to \mathbb{R}^n$ by

$$\Psi^h(y_0) := y_1, \tag{6.7}$$

where y_1 is the solution to

$$y_1 = y_0 + h\Phi(t_0, y_0, y_1, h).$$

Thereby, we assume that Φ is Lipschitz continuous with respect to y_1 and h is sufficiently small.

Please compare this definition to Def. 6.8.

Definition 6.23 (symplectic integrator). A one-step method with discrete flow $\Psi^h : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ is called a *symplectic integrator*, if applied to any Hamiltonian system as defined in Def. 6.1 the discrete flow is symplectic, i.e., there exists an $H > 0$ such that

$$\forall h \in (0, H) \quad \forall y_0 \in \mathbb{R}^{2d} : \quad D\Psi^h(y_0)^\top J D\Psi^h(y_0) = J.$$

Since the discrete flow of a symplectic integrator should be symplectic for all Hamiltonian systems, the most simple case $d = 1$ can often be used to show that an integrator is not symplectic. For $d = 1$, Lemma 6.17 can be improved.

Lemma 6.24. *$\Phi \in C^1(\mathbb{R}^2; \mathbb{R}^2)$ is symplectic if and only if $\det D\Phi(z) = 1$ for all $z \in \mathbb{R}^2$.*

Proof. A straightforward calculation shows

$$\begin{pmatrix} 0 & \det D\Phi(z) \\ -\det D\Phi(z) & 0 \end{pmatrix} = (D\Phi(z))^\top J D\Phi(z) \overset{!}{=} J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

□

Proposition 6.25. *The explicit Euler method is not symplectic.*

Proof. For $d = 1$ we consider the Hamiltonian system corresponding to the Hamiltonian $H \in C^2(\mathbb{R}^2; \mathbb{R}^2)$ defined by

$$H(p, q) := \frac{1}{2}p^2 + \frac{1}{2}q^2.$$

Hence, the explicit Euler method applied to the system

$$\begin{aligned} p'(t) &= -q(t), \\ q'(t) &= p(t) \end{aligned} \tag{6.8}$$

leads to

$$\begin{aligned} p_1 &= p_0 - hq_0, \\ q_1 &= q_0 + hp_0, \end{aligned}$$

i.e.

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \Psi^h\left(\left(\begin{smallmatrix} p_0 \\ q_0 \end{smallmatrix}\right)\right) \qquad \text{with} \qquad D\Psi^h = \begin{pmatrix} 1 & -h \\ h & 1 \end{pmatrix}.$$

Since $\det \Psi^h = 1 + h^2$, the last lemma yields the claim. □

Note that the Hamiltonian should be constant for solutions to the Hamiltonian system. But using the explicit Euler method leads to a growth, since

$$H(p_1, q_1) = \frac{1}{2}\left((p_0 - hq_0)^2 + (q_0 + hp_0)^2\right) = (1 + h^2)H(p_0, q_0).$$

For stiff problems, using implicit method was the remedy. But even the implicit Euler method is not symplectic.

Proposition 6.26. *The implicit Euler method is not symplectic.*

Proof. The implicit Euler method applied to the same Hamiltonian system (6.8) as in the last proof leads to

$$\begin{pmatrix} p_1 \\ q_1 \end{pmatrix} = \Psi^h\left(\left(\begin{smallmatrix} p_0 \\ q_0 \end{smallmatrix}\right)\right) := \frac{1}{1+h^2}\begin{pmatrix} p_0 - hq_0 \\ q_0 + hp_0 \end{pmatrix}$$

with $D\Psi^h = \frac{1}{1+h^2}\left(\begin{smallmatrix} 1 & -h \\ h & 1 \end{smallmatrix}\right)$ and $\det D\Psi^h = (1 + h^2)^{-1}$. Hence, Lemma 6.24 yields the claim. □

For the implicit Euler method, we compute

$$H(p_1, q_1) = \frac{1}{1 + h^2}H(p_0, q_0),$$

i.e., the Hamiltonian decays exponentially for the discrete solutions.

Theorem 6.27. *All Runge-Kutta methods, which preserve quadratic invariants, are symplectic. In particular, due to Prop. 6.21 Gauss methods are symplectic.*

Proof. We consider a Hamiltonian system $y' = f(y)$ with (continuous) flow Φ^t such that $\left(D\Phi^t(y_0)\right)^\top JD\Phi^t(y_0) = J$ for all $y_0 \in \mathbb{R}^{2d}$. Moreover, we use an m-stage Runge-Kutta method with discrete flow Ψ^h such that

$$y_1 = \Psi^h(y_0) := y_0 + h\sum_{j=1}^m b_j k_j(y_0)$$

with stages

$$k_i(y_0) = f\left(y_0 + h\sum_{j=1}^m A_{ij}k_j(y_0)\right) \quad \text{for all } i = 1, \dots, m.$$

We have to show that for sufficiently small h

$$\forall y_0 \in \mathbb{R}^{2d}: \quad \left(D_{y_0}\Psi^h(y_0)\right)^\top JD_{y_0}\Psi^h(y_0) = J.$$

Step 1: We first apply the Runge-Kutta method to the extended system for $x := (y, W)^\top : [0, T] \to \mathbb{R}^{2d} \times \mathbb{R}^{2d\times 2d}$ and

$$x' = \begin{pmatrix} y' \\ W' \end{pmatrix} = \begin{pmatrix} f(y) \\ (Df)(y)W \end{pmatrix}, \qquad x(0) = \begin{pmatrix} y_0 \\ \mathrm{id}_{2d\times 2d} \end{pmatrix}.$$

Note that $W(t) = D_{y_0}\Phi^t(y_0)$ is the Wronski matrix of the flow Φ^t as in the proof of Theorem 6.16. To enhance readability, we drop in the following the dependency on y_0 and apply the Runge-Kutta method to this extended system leading to

$$\begin{pmatrix} \hat{y}_1 \\ W_1 \end{pmatrix} = \Psi^h\left(\left(\begin{smallmatrix} y_0 \\ \mathrm{id} \end{smallmatrix}\right)\right) = \begin{pmatrix} y_0 + h\sum_{j=1}^m b_j k_j^y \\ \mathrm{id} + h\sum_{j=1}^m b_j k_j^W \end{pmatrix}$$

with stages $\left(k_j^y, k_j^W\right)^\top$ satisfying for all $i = 1, \dots, m$

$$\begin{pmatrix} k_i^y \\ k_i^W \end{pmatrix} = \begin{pmatrix} f\left(y_0 + h\,\sum_{j=1}^m A_{ij} k_j^y\right) \\ (Df)(y_0 + h\,\sum_{j=1}^m A_{ij} k_j^y)\left(\mathrm{id} + \sum_{j=1}^m A_{ij} k_j^W\right) \end{pmatrix}. \tag{6.9}$$

The stages k_i^y are exactly the same as the ones for the original problem $y' = f(y)$, i.e., $k_i^y = k_i$ and $y_1 = \hat{y}_1$. Since the second row in (6.9) is the derivative of the first one with respect to y_0, there holds $k_j^W(y_0) = Dk_j^y(y_0)$ for $j = 1, \dots, m$. Hence,

$$Dy_1(y_0) = D\hat{y}_1(y_0) = D\left\{ y_0 + h \sum_{j=1}^m b_j k_j^y(y_0) \right\} = \mathrm{id} + h \sum_{j=1}^m b_j D k_j^y(y_0) = W_1(y_0).$$

Step 2: Up to now, we have not used that the (continuous) flow is symplectic and that the Runge-Kutta method preserve quadratic invariants. The first property proves that the quadratic functional I defined by

$$I\left(\left(\begin{smallmatrix} y(t) \\ W(t) \end{smallmatrix}\right)\right) := W(t)^\top J W(t)$$

is an invariant of the extended system, since $W(t) = D\Phi^t(y_0)$. Moreover, $I(t) = J$ for all $t \in [0, T]$. Since this invariant is preserved, the claim follows by

$$J = I\left(\left(\begin{smallmatrix} \hat{y}_0 \\ \mathrm{id} \end{smallmatrix}\right)\right) = I\left(\left(\begin{smallmatrix} \hat{y}_1 \\ W_1 \end{smallmatrix}\right)\right) = W_1^\top J W_1 = (Dy_1)^\top J Dy_1 = \left(D\Psi^h(y_0)\right)^\top J D\Psi^h(y_0).$$

□

Note that the requirement on Runge-Kutta methods, that they should preserve quadratic invariants, is quite restrictive. Therefore, often a different type of one-step method is used.

Definition 6.28 (Partitioned Runge-Kutta methods). Let $\begin{array}{c|c} c^{(p)} & A^{(p)} \\ \hline & b^{(p)\top} \end{array}$ and $\begin{array}{c|c} c^{(q)} & A^{(q)} \\ \hline & b^{(q)\top} \end{array}$ be the Butcher tableau of two m-stage Runge-Kutta methods and $(p, q)^\top$ with $p, q \in \mathbb{R}^d$ be the solution to the initial value problem

$$\begin{pmatrix} p \\ q \end{pmatrix}' = f(p, q) = \begin{pmatrix} f_1(p, q) \\ f_2(p, q) \end{pmatrix}, \qquad \begin{pmatrix} p \\ q \end{pmatrix}(0) = \begin{pmatrix} p_0 \\ q_0 \end{pmatrix}$$

with $f_1, f_2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, and $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{2d}$. Then we define one step of the *partitioned Runge-Kutta method* by

$$p_1 = p_0 + h \sum_{j=1}^m b_j^{(p)} k_j^{(p)}, \qquad q_1 = q_0 + h \sum_{j=1}^m b_j^{(q)} k_j^{(q)},$$

where the stages $k_j^{(p)}$ and $k_j^{(q)}$ satisfy for $i = 1, \dots, m$

$$\begin{pmatrix} k_i^{(p)} \\ k_i^{(q)} \end{pmatrix} = f\left(p_0 + h \sum_{j=1}^m A_{ij}^{(p)} k_i^{(p)}, q_0 + h \sum_{j=1}^m A_{ij}^{(q)} k_i^{(q)} \right).$$

Note that this is the first time in this lecture that we use different methods for the different components of a differential equation. Of course, the reason for this is the specific structure of Hamiltonian systems defined in Def. 6.1.

Example 6.29 (symplectic Euler). The symplectic Euler method is a partitioned Runge-Kutta method with Butcher tableau $\begin{array}{c|c} 0 & 0 \\ \hline & 1^\top \end{array}$ (explicit Euler) and $\begin{array}{c|c} 1 & 1 \\ \hline & 1^\top \end{array}$ (implicit Euler) leading to

$$\begin{pmatrix} p_{\ell+1} \\ q_{\ell+1} \end{pmatrix} = \begin{pmatrix} p_\ell + hf_1\left(p_\ell, q_{\ell+1}\right) \\ q_\ell + hf_2\left(p_\ell, q_{\ell+1}\right) \end{pmatrix}.$$

We can interchange the Butcher tableau to a second version of the symplectic Euler method

$$\begin{pmatrix} p_{\ell+1} \\ q_{\ell+1} \end{pmatrix} = \begin{pmatrix} p_\ell + hf_1\left(p_{\ell+1}, q_\ell\right) \\ q_\ell + hf_2\left(p_{\ell+1}, q_\ell\right) \end{pmatrix}. \tag{6.10}$$

Proposition 6.30. *The symplectic Euler methods are symplectic with convergence order* 1.

Proof. The convergence order follows as usual with Taylor expansions. Here, we only show that the second symplectic Euler method is symplectic. To this end, we need the derivative

$$D\Psi^h = \begin{pmatrix} \partial_{p_0} p_1 & \partial_{q_0} p_1 \\ \partial_{p_0} q_1 & \partial_{q_0} q_1 \end{pmatrix}$$

of the discrete flow Ψ^h. Note that we have defined symplectic integrators only for Hamiltonian systems, i.e. $f_1(p,q) = -\nabla_q H(p,q)$ and $f_2(p,q) = \nabla_p H(p,q)$ for a Hamiltonian $H \in C^2(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R})$. Building the derivative of (6.10) for $\ell = 0$ with respect to p_0 and q_0 leads to

$$\begin{pmatrix} A & 0_{d\times d} \\ B & \mathrm{id}_{d\times d} \end{pmatrix} D\Psi^h = \begin{pmatrix} \mathrm{id}_{d\times d} & C \\ 0_{d\times d} & A \end{pmatrix}$$

with the symmetric matrices $A, B, C \in \mathbb{R}^{d\times d}$ given by

$$A := \mathrm{id}_{d\times d} + hH_{pq}, \qquad B := -hH_{pp}, \qquad C := -hH_{qq},$$

where $\left(\begin{smallmatrix} H_{pp} & H_{pq} \\ H_{pq} & H_{qq} \end{smallmatrix}\right)$ denotes the Hessian matrix of H at the point (p_1, q_0). Hence,

$$D\Psi^h = \begin{pmatrix} A^{-1} & 0 \\ -BA^{-1} & \mathrm{id} \end{pmatrix} \begin{pmatrix} \mathrm{id} & C \\ 0 & A \end{pmatrix} = \begin{pmatrix} A^{-1} & A^{-1}C \\ -BA^{-1} & A - BA^{-1}C \end{pmatrix}.$$

Finally, we compute

$$\begin{aligned} D\Psi^{h\top} J D\Psi^h &= \begin{pmatrix} A^{-1} & -A^{-1}B \\ CA^{-1} & A - CA^{-1}B \end{pmatrix} \begin{pmatrix} 0 & \mathrm{id} \\ -\mathrm{id} & 0 \end{pmatrix} \begin{pmatrix} A^{-1} & A^{-1}C \\ -BA^{-1} & A - BA^{-1}C \end{pmatrix} \\ &= \begin{pmatrix} A^{-1}B & A^{-1} \\ CA^{-1}B - A & CA^{-1} \end{pmatrix} \begin{pmatrix} A^{-1} & A^{-1}C \\ -BA^{-1} & A - BA^{-1}C \end{pmatrix} \\ &= \begin{pmatrix} 0 & \mathrm{id} \\ -\mathrm{id} & 0 \end{pmatrix} = J. \end{aligned}$$

□

Compare the next proposition with Rem. 6.7. Many Hamiltonians have the following form.

Proposition 6.31. *If the Hamiltonian is of the form*

$$H(p,q) = T(p) + U(q),$$

then both versions of the symplectic Euler method are explicit. Moreover, if in addition $T(p) = p^\top M p$ with some matrix $M \in \mathbb{R}^{d\times d}$, then the symplectic Euler methods preserve all quadratic invariants with the special form

$$I(p,q) = p^\top C q + c^\top p + d^\top q + w$$

with arbitrary $C \in \mathbb{R}^{d\times d}$, $c, d \in \mathbb{R}^d$, and $w \in \mathbb{R}$. This includes all possible linear invariants.

Proof. For the first version of the symplectic Euler method, the method becomes

$$\begin{pmatrix} p_{\ell+1} \\ q_{\ell+1} \end{pmatrix} = \begin{pmatrix} p_\ell - h\nabla U\left(q_{\ell+1}\right) \\ q_\ell + h\nabla T\left(p_\ell\right) \end{pmatrix}. \tag{6.11}$$

Hence, from the second equation $q_{\ell+1}$ can be computed explicitly. Then, the known value $q_{\ell+1}$ can be used in the first equation to compute $p_{\ell+1}$ explicitly.

Since I should be an invariant, Prop.6.19 leads to

$$0 = \begin{pmatrix} \nabla_p I(p,q) \\ \nabla_q I(p,q) \end{pmatrix} \cdot \begin{pmatrix} -\nabla U(q) \\ \nabla T(p) \end{pmatrix} = -\left(q^\top C^\top + c^\top\right) \nabla U(q) + \left(p^\top C + d^\top\right) \nabla T(p)$$

for all (p,q). Hence, $\left(q^\top C^\top + c^\top\right) \nabla U(q) = \left(p^\top C + d^\top\right) \nabla T(p) = \text{const}$. Since $\nabla T(p) = Mp$, this constant has to be zero. Thus, if I is an invariant, then $\left(q^\top C^\top + c^\top\right) \nabla U(q) = \left(p^\top C + d^\top\right) \nabla T(p) = 0$ for all p and q. Using (6.11) the claim follows by

$$\begin{aligned}
I(p_1,q_1) &= \left(p_0 - h\nabla U(q_1)\right)^\top (Cq_1 + c) + d^\top q_1 + w \\
&= p_0^\top (Cq_1 + c) - h\underbrace{\nabla U(q_1)^\top (Cq_1 + c)}_{=0} + d^\top q_1 + w \\
&= \left(p_0^\top C + d^\top\right) q_1 + p_0^\top c + w \\
&= \left(p_0^\top C + d^\top\right) (q_0 + h\nabla T(p_0)) + p_0^\top c + w \\
&= \left(p_0^\top C + d^\top\right) q_0 + h\underbrace{\left(p_0^\top C + d^\top\right) \nabla T(p_0)}_{=0} + p_0^\top c + w \\
&= p_0^\top C q_0 + p_0^\top c + d^\top q_0 + w = I(p_0,q_0).
\end{aligned}$$

□

Corollary 6.32. *The total momentum $\sum_{i=1}^N p_i$ and the total angular momentum $\sum_{i=1}^N q_i \times p_i$ of the N-body problem of the form Ex. 6.5 are preserved by the symplectic Euler methods due to Ex. 6.6 and the last proposition.*

Example 6.33 (Störmer-Verlet method). The Störmer-Verlet method is a partitioned Runge-Kutta method with Butcher tableau

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

For a Hamiltonian system, this leads to the stages

$$\begin{aligned}
k_1^{(p)} &= -\nabla_q H\left(p_\ell, q_\ell + \frac{h}{2}k_1^{(q)}\right),\\
k_2^{(p)} &= -\nabla_q H\left(p_\ell + \frac{h}{2}\left(k_1^{(p)} + k_2^{(p)}\right), q_\ell + \frac{h}{2}k_1^{(q)}\right),\\
k_1^{(q)} &= \nabla_p H\left(p_\ell, q_\ell + \frac{h}{2}k_1^{(q)}\right),\\
k_2^{(q)} &= \nabla_p H\left(p_\ell + \frac{h}{2}\left(k_1^{(p)} + k_2^{(p)}\right), q_\ell + \frac{h}{2}k_1^{(q)}\right).
\end{aligned}$$

This can be simplified. We define $q_{\ell+1/2} := q_\ell + \frac{h}{2}k_1^{(q)}$. Hence, we first solve

$$q_{\ell+1/2} = q_\ell + \frac{h}{2}\nabla_p H\left(p_\ell, q_{\ell+1/2}\right) \tag{6.12a}$$

for $q_{\ell+1/2}$. Second, we solve

$$p_{\ell+1} = p_\ell + \frac{h}{2}\left(-\nabla_q H\left(p_\ell, q_{\ell+1/2}\right) - \nabla_q H\left(p_{\ell+1}, q_{\ell+1/2}\right)\right) \tag{6.12b}$$

for $p_{\ell+1}$. Then, $q_{\ell+1}$ is explicitly given by

$$q_{\ell+1} = q_{\ell+1/2} + \frac{h}{2}\nabla_p H\left(p_{\ell+1}, q_{\ell+1/2}\right). \tag{6.12c}$$

As for the symplectic Euler methods, of course a second version is possible, if the Butcher tableau are interchanged. (6.12a) and (6.12b) and thus the whole method is explicit, if $H(p,q) = T(p) + U(q)$.

Proposition 6.34. *The Störmer-Verlet methods are symplectic with convergence order* 2.

Proof. A closer look to (6.12) shows, that the Störmer-Verlet methods consists of two steps with mesh size $h/2$ of the two symplectic Euler methods. Hence, Exercise 6.13 shows that the methods are symplectic. The convergence order can be derived in the usual way by Taylor expansions. □

Prop. 6.31 holds true for the Störmer-Verlet methods as well, since they are a composition of the symplectic Euler methods. Note that, in general, symplectic integrators do not preserve the Hamiltonian.

6.4. Adjoint integrators

The proof of Prop. 6.34 uses the fact, that the Störmer-Verlet method consists of two steps of symplectic Euler methods. In principle, such combinations can be extended. Since for those adjoint integrators will be useful, we first discuss them.

Remark 6.35. Recall that implicit one-step methods are locally explicit (see Prop. 3.8), i.e., the discrete flow reads for autonomous systems

$$\Psi^h(y_0) = y_0 + h\Phi(y_0, \Psi^h(y_0), h) =: y_0 + h\tilde{\Phi}(y_0, h).$$

Hence, $D\Psi^h(y_0) = \mathrm{id} + hD\tilde{\Phi}(y_0, h)$, and $D\Psi^h(y_0)$ is invertible for sufficiently small $|h|$. The inverse functions theorem yields that $\left(\Psi^h\right)^{-1}$ is locally well-defined for sufficiently small $|h|$.

Definition 6.36 (adjoint integrator). Given the flow Ψ^h of a one-step method, we define the *adjoint integrator* by the *adjoint flow*

$$\left(\Psi^h\right)^* := \left(\Psi^{-h}\right)^{-1}. \tag{6.13}$$

An integrator is called *symmetric* or *reversible*, if $\left(\Psi^h\right)^* = \Psi^h$.

Proposition 6.37. *Let $\Phi^t : \mathbb{R}^n \to \mathbb{R}^n$ be the in Def. 6.8 defined (continuous) flow of an autonomous initial value problem with right hand side $f \in C^1(\mathbb{R}^n;\mathbb{R}^n)$. Then*

$$\forall t \in \mathbb{R}: \quad \Phi^{-t} \circ \Phi^t = \mathrm{id},$$

i.e., Φ^t is reversible (in the sense of the preceding definition).

Proof. By definition of the flow we have $z_0 := \Phi^s(y_0) = y(s)$ for the solution y to the initial value problem $y' = f(y)$ with $y(0) = y_0$. In the same way, $\Phi^{-s}(z_0) = z(-s)$ for the solution z to $z' = f(z)$ with $z(0) = z_0$. Since the solution to both initial value problems are locally unique, the claim follows with $z(t) = y(s+t)$, i.e.,

$$\Phi^{-s} \circ \Phi^s(y_0) = \Phi^{-s}(z_0) = z(-s) = y(0) = y_0.$$

□

As we have seen, the continuous flow of an autonomous system is always reversible. Hence, it seems reasonable to ask for reversible integrators as well. In practice, it is easy to check whether a one-step method is reversible or not. We just have to interchange y_0 with y_1 and h with $-h$ and check if this is the same method.

Example 6.38. The implicit midpoint rule is reversible, since

$$\begin{aligned} y_1 &= \Psi^h(y_0) = y_0 + hf\left(\frac{y_0+y_1}{2}\right), \\ y_0 &= \Psi^{-h}(y_1) = y_1 - hf\left(\frac{y_0+y_1}{2}\right) \quad \Rightarrow y_1 = y_0 + hf\left(\frac{y_0+y_1}{2}\right). \end{aligned}$$

Example 6.39. The adjoint of the explicit Euler method is the implicit Euler method and vice versa, since

$$\begin{aligned} y_1 &= \Psi^h(y_0) = y_0 + hf(y_0), \\ y_0 &= \Psi^{-h}(y_1) = y_1 - hf(y_1) \quad \Rightarrow y_1 = y_0 + hf(y_1). \end{aligned}$$

Hence, explicit and implicit Euler methods are not reversible.

Exercise 6.40. Show that the two symplectic Euler methods of Ex. 6.29 are adjoint to each other.

Proposition 6.41. *An m-stage Runge-Kutta method* $\begin{array}{c|c} c & A \\ \hline & b^\top \end{array}$ *with*

$$A_{m+1-i,m+1-j} + A_{ij} = b_j, \qquad i,j = 1,\dots,m,$$

is reversible.

Proof. We have

$$\begin{aligned} y_1 &= y_0 + h\sum_{j=1}^m b_j k_j^{(1)} \qquad \text{with } k_i^{(1)} = f\left(y_0 + h\sum_{j=1}^m A_{ij}k_j^{(1)}\right), \\ \tilde{y}_0 &= y_1 - h\sum_{j=1}^m b_j k_j^{(2)} \qquad \text{with } k_i^{(2)} = f\left(y_1 - h\sum_{j=1}^m A_{ij}k_j^{(2)}\right). \end{aligned}$$

Since

$$k_i^{(1)} = f\Big(y_1 + h\sum_{j=1}^{m} \underbrace{(A_{ij} - b_j)}_{=-A_{m+1-i,m+1-j}} \, k_j^{(1)}\Big) = k_{m+1-i}^{(2)},$$

and $b_j = b_{m+1-j}$ by assumption, the claim follows with

$$\tilde{y}_0 = y_1 - h\sum_{j=1}^{m} b_{m+1-j}k_{m+1-j}^{(2)} = y_1 - h\sum_{j=1}^{m} b_j k_j^{(1)} = y_0.$$

□

Exercise 6.42. Show that consistent, explicit Runge-Kutta methods are not reversible. To this end, apply the method to the differential equation $y' = y$ and show that the discrete flow $\Psi^h\,(y_0)$ is a polynomial in h.

Proposition 6.43. *Collocation methods as defined in Def. 3.21 with collocation nodes* $0 \le c_1 \le \dots \le c_m \le 1$ *are reversible, if the nodes are symmetric with respect to* $1/2$, *i.e., if*

$$\forall j = 1, \dots, m: \quad \frac{1}{2} - c_j = c_{m+1-j} - \frac{1}{2}.$$

Proof. By definition, $y_1 = q(h)$ for the solution $q \in \mathbb{P}_m$ of

$$q(0) = y_0, \qquad q'(c_j h) = f\,(q(c_j h)) \quad \text{for all } j = 1, \dots, m.$$

Interchanging y_0 with y_1 and h with $-h$ gives $\tilde{y}_0 = \tilde{q}(-h)$ with the solution $\tilde{q} \in \mathbb{P}_m$ of

$$\tilde{q}(0) = y_1, \qquad \tilde{q}'(-c_j h) = f\,(\tilde{q}(-c_j h)) \quad \text{for all } j = 1, \dots, m.$$

The polynomial $\hat{q}$ defined by $\hat{q}(t) := \tilde{q}(t-h)$ satisfies $\hat{q}(h) = y_1$, and for all $j = 1, \dots, m$

$$\hat{q}'(h - c_j h) = \tilde{q}'(-c_j h) = f\,(\tilde{q}(-c_j h)) = f\,(\hat{q}(h - c_j h))\,.$$

Since $h - c_j h = h(1 - c_j) = c_{m+1-j}h$ by assumption, uniqueness of the collocation polynomial for sufficiently small h yields $\hat{q} \equiv q$. Thus, the claim follows with

$$\tilde{y}_0 = \tilde{q}(-h) = \hat{q}(0) = q(0) = y_0.$$

□

Exercise 6.44. Show that the Gauss collocation methods are reversible.

Lemma 6.45. *Let* Ψ^h *and* $\hat{\Psi}^h$ *be the discrete flows of two numerical methods. Then, for sufficiently small* h *there holds*

$$\Big(\big(\Psi^h\big)^*\Big)^* = \Psi^h,$$

$$\Big(\Psi^h \circ \hat{\Psi}^h\Big)^* = \Big(\hat{\Psi}^h\Big)^* \circ \big(\Psi^h\big)^*.$$

Proof. For the first equation, we use the definition of the adjoint flow $\tilde{\Psi}^h := \big(\Psi^h\big)^*$ in (6.13):

$$y_1 = \Big(\big(\Psi^h\big)^*\Big)^*(y_0) = \Big(\tilde{\Psi}^{-h}\Big)^{-1}(y_0) \Rightarrow y_0 = \tilde{\Psi}^{-h}y_1 = \big(\Psi^{-h}\big)^*(y_1) = \big(\Psi^h\big)^{-1}(y_1) \Rightarrow y_1 = \Psi^h\,(y_0)\,.$$

The second equation follows directly from the definition, since

$$\Big(\Psi^h \circ \hat{\Psi}^h\Big)^* = \Big(\Psi^{-h} \circ \hat{\Psi}^{-h}\Big)^{-1} = \Big(\hat{\Psi}^{-h}\Big)^{-1} \circ \big(\Psi^{-h}\big)^{-1} = \Big(\hat{\Psi}^h\Big)^* \circ \big(\Psi^h\big)^*.$$

□

COROLLARY 6.46. *Let Ψ^h be the discrete flow of a one-step method. Then*

$$\Psi_1^h := \Psi^{h/2} \circ \left(\Psi^{h/2}\right)^* \qquad \text{and } \Psi_2^h := \left(\Psi^{h/2}\right)^* \circ \Psi^{h/2}$$

are the discrete flow of reversible one-step methods.

PROOF. This is a direct consequence of the preceding lemma. □

COROLLARY 6.47. *The Störmer-Verlet methods of Ex. 6.33 are reversible.*

PROOF. In the proof of Prop. 6.34 it was shown that the Störmer-Verlet methods are a combination of the two symplectic Euler methods of Ex. 6.29. Hence, Exercise 6.40 together with the last corollary yields the claim. □

With Cor. 6.46 is it straightforward to construct reversible integrators. But be aware, that even explicit method may lead to implicit methods, as Example 6.39 for the explicit and implicit Euler method show. So the computational effort might grow substantially.

Nevertheless, as we have seen in the proof of Prop. 6.34 the Störmer-Verlet method is a combination of two one-step methods of order one but it has order two. So the convergence order might increase. In the following we show, that the convergence order of reversible integrators is always even.

LEMMA 6.48. *Let Ψ^h and $\hat{\Psi}^h$ be the discrete flows of two one-step methods of order p. Moreover, we assume that the incremental function Φ of $\Psi^h y_0 = y_0 + h\Phi(y_0, h)$ is at least C^1 w.r.t. y_0. Then $\Psi^h \circ \hat{\Psi}^h$ is the flow of a one-step method of order at least p.*

PROOF. Let $y \in C^{p+1}([0,T];\mathbb{R}^n)$ be the solution to $y' = f(y)$ with $y(0) = y_0$. Since

$$y(h) - \hat{\Psi}^h(y_0) = \mathcal{O}\left(h^{p+1}\right),$$

Taylor expansion of Φ yields the claim:

$$\begin{aligned}\Psi^h \circ \hat{\Psi}^h(y_0) &= \Psi^h\left(y(h) - \mathcal{O}\left(h^{p+1}\right)\right)\\ &= y(h) - \mathcal{O}\left(h^{p+1}\right) + h\Phi\left(y(h) - \mathcal{O}\left(h^{p+1}\right), h\right)\\ &= y(h) + h\Phi(y(h), h) + \mathcal{O}\left(h^{p+1}\right)\\ &= \Psi^h\left(y(h)\right) + \mathcal{O}\left(h^{p+1}\right) = y(2h) + \mathcal{O}\left(h^{p+1}\right).\end{aligned}$$

□

Note that implicit methods are locally explicit. Hence, the incremental function Φ can be chosen as stated even for implicit methods.

LEMMA 6.49. *Let $\Phi^t \in C(\mathbb{R}^n;\mathbb{R}^n)$ for all t be Lipschitz continuous. Moreover, we use a one-step method with discrete flow Ψ^h such that*

$$\forall y_0 \in \mathbb{R}^n: \quad \Psi^h(y_0) - \Phi^h(y_0) = \mathcal{O}(h^{p+1}) \qquad \text{for } |h| \to 0. \tag{6.14}$$

Then,

$$\forall y_0 \in \mathbb{R}^n: \quad \left(\Psi^h\right)^*(y_0) - \Phi^h(y_0) = \mathcal{O}(h^{p+1}) \qquad \text{for } |h| \to 0.$$

PROOF. $y_1 := \left(\Psi^h\right)^*(y_0)$ is by definition of the adjoint flow for sufficiently small $|h|$ equivalent to $\Psi^{-h}(y_1) = y_0$. By Lipschitz continuity and with Prop. 6.37 we compute

$$\begin{aligned}\left\|\left(\Psi^h\right)^*(y_0) - \Phi^h(y_0)\right\| &= \left\|\Phi^h\left(\Phi^{-h}(y_1)\right) - \Phi^h\left(\Psi^{-h}(y_1)\right)\right\|\\ &\le L\left\|\Phi^{-h}(y_1) - \Psi^{-h}(y_1)\right\| = \mathcal{O}(h^{p+1}).\end{aligned}$$

□

Typically, (6.14) will hold for one-step methods of order p, if $f \in C^p(\mathbb{R}^n;\mathbb{R}^n)$ and hence $\Phi^t, \Psi^t \in C^p(\mathbb{R}^n;\mathbb{R}^n)$ for $t \in [0,T]$. Note that in Def. 2.7 for the local consistency error we have used only $h > 0$.

Theorem 6.50. *Consider a one-step method with discrete flow Ψ^h. Moreover, we assume a local error of the form*

$$\forall y_0 \in \mathbb{R}^n : \quad \Psi^h(y_0) - \Phi^h(y_0) = C(y_0)h^{p+1} + \mathcal{O}(h^{p+2}) \qquad \text{for } |h| \to 0, \tag{6.15a}$$

with a flow $\Phi^h \in C^2(\mathbb{R}^n;\mathbb{R}^n)$ of an ODE and a function $C \in C^1(\mathbb{R}^n;\mathbb{R}^n)$. Then, the adjoint flow $\left(\Psi^h\right)^*$ satisfies

$$\forall y_0 \in \mathbb{R}^n : \left(\Psi^h\right)^*(y_0) - \Phi^h(y_0) = (-1)^p C(y_0)h^{p+1} + \mathcal{O}(h^{p+2}) \quad \text{for } |h| \to 0. \tag{6.15b}$$

Proof. **Step 1:** We define for all sufficiently small $|h|$ and all $y_0 \in \mathbb{R}^n$ the local errors

$$\tau(y_0,h) := \Psi^h(y_0) - \Phi^h(y_0), \qquad \tau^*(y_0,h) := \left(\Psi^h\right)^*(y_0) - \Phi^h(y_0).$$

By assumption the last lemma implies

$$\tau(y_0,h) = C(y_0)h^{p+1} + \mathcal{O}(h^{p+2}), \qquad \tau^*(y_0,h) = \mathcal{O}\left(h^{p+1}\right).$$

In the following we have to show the special form of $\tau^*(y_0,h)$.

Step 2: Let $y_1 := \Phi^h(y_0)$. Since $\Phi^{-h}(y_1) = y_0$ by Prop. 6.37 and by definition of the adjoint flow $\Psi^{-h}\left(\left(\Psi^h\right)^*(y_0)\right) = y_0$, we have

$$\tau(y_1,-h) = \Psi^{-h}(y_1) - \Psi^{-h}\left(\left(\Psi^h\right)^*(y_0)\right) = \Psi^{-h}(y_1) - \Psi^{-h}(y_1 + \tau^*(y_0,h)).$$

By assumption,

$$\begin{aligned}\Psi^{-h}(y_1) =& \Phi^{-h}(y_1) + C(y_1)(-h)^{p+1} + \mathcal{O}\left(h^{p+2}\right),\\ \Psi^{-h}(y_1 + \tau^*(y_0,h)) =& \Phi^{-h}(y_1 + \tau^*(y_0,h)) + C(y_1 + \tau^*(y_0,h))(-h)^{p+1} + \mathcal{O}\left(h^{p+2}\right).\end{aligned}$$

Recall, that due to Rem. 6.9 $D\Phi^{-h}(y_1) = \mathrm{id} + \mathcal{O}(h)$. Thus, by Taylor expansion

$$\Phi^{-h}(y_1 + \tau^*(y_0,h)) = \Phi^{-h}(y_1) + \tau^*(y_0,h) + \mathcal{O}\left(h^{p+2}\right),$$

since $\|\tau^*(y_0,h)\| = \mathcal{O}\left(h^{p+1}\right)$. Similarly, Taylor expansion of C leads to

$$C(y_1 + \tau^*(y_0,h)) = C(y_1) + \mathcal{O}\left(h^{p+2}\right).$$

All together, we have

$$\tau(y_1,-h) = -\tau^*(y_0,h) + \mathcal{O}\left(h^{p+2}\right).$$

Finally, $C(y_1) = C(y_0) + \mathcal{O}(h)$ yields

$$\begin{aligned}\tau^*(y_0,h) &= -\tau(y_1,-h) + \mathcal{O}\left(h^{p+2}\right) = -C(y_1)(-h)^{p+1} + \mathcal{O}\left(h^{p+2}\right)\\ &= (-1)^p C(y_0)h^{p+1} + \mathcal{O}\left(h^{p+2}\right).\end{aligned}$$

□

Corollary 6.51. *In addition to the assumptions of the last theorem let Ψ^h be reversible. Then the convergence order p is even.*

Proof. This is a consequence of the fact that for reversible Ψ^h there holds $\tau(y_0,h) = \tau^*(y_h,h)$. Hence, the lowest order terms

$$C(y_0)h^{p+1} = (-1)^p C(y_0)h^{p+1}$$

yield $C(y_0) = 0$ if p is odd. □

The last corollary shows, that the Störmer-Verlet methods of Ex. 6.33 have at least convergence order 2, since they are reversible by Cor. 6.47.

6.5. Splitting methods

We have seen in Prop. 6.31 that a specific form of the right hand side of a differential equation might be quite advantageous for a numerical solver. In this section we consider autonomous ODE's of the form

$$y' = f(y) + g(y). \tag{6.16}$$

Of course, all autonomous ODE's can be written in this form. So in addition we assume, that the continuous flows Φ_f^t and Φ_g^t to the ODE's

$$y' = f(y) \qquad \text{and } y' = g(y) \tag{6.17}$$

are computable.

Definition 6.52 (Lie-Trotter splitting). A discrete flow Ψ^h for the ODE (6.16) can be constructed by

$$\Psi^h := \Phi_f^h \circ \Phi_g^h. \tag{6.18}$$

The adjoint flow $\left(\Psi^h\right)^*$ is due to Prop. 6.37 and Lemma 6.45 given by $\left(\Psi^h\right)^* = \Phi_g^h \circ \Phi_f^h$. These methods are called *Lie-Trotter splitting methods.*

Example 6.53. Consider a Hamiltonian system (see Def. 6.1) of the form $H(p,q) = T(p) + U(q)$ as in Prop. 6.31. Then, the system has the form (6.16) with $y = (p,q)^\top$ and

$$f\left(\begin{pmatrix} p \\ q \end{pmatrix}\right) := \begin{pmatrix} -\nabla U(q) \\ 0 \end{pmatrix}, \qquad g\left(\begin{pmatrix} p \\ q \end{pmatrix}\right) := \begin{pmatrix} 0 \\ \nabla T(p) \end{pmatrix}.$$

The continuous flows to (6.17) are

$$\Phi_f^t\left(\begin{pmatrix} p_0 \\ q_0 \end{pmatrix}\right) = \begin{pmatrix} p_0 - t\nabla U(q_0) \\ q_0 \end{pmatrix}, \qquad \Phi_g^t\left(\begin{pmatrix} p_0 \\ q_0 \end{pmatrix}\right) = \begin{pmatrix} p_0 \\ q_0 + t\nabla T(p_0) \end{pmatrix}.$$

A straightforward calculation shows that the Lie-Trotter splittings Ψ^h and $\left(\Psi^h\right)^*$ as defined in Def. 6.52 are the two symplectic Euler methods.

Proposition 6.54. *The Lie-Trotter splitting has at least the consistency order $p = 1$, if f, g are sufficiently smooth.*

Proof. Let y, w, z be the solutions to

$$\begin{aligned} y' &= f(y) + g(y), & y(0) &= y_0, \\ w' &= g(w), & w(0) &= y_0, \\ z' &= f(z), & z(0) &= w(h). \end{aligned}$$

Then,

$$\begin{aligned} \Phi^h(y_0) &= y(h) = y_0 + \int_0^h f(y(\tau))\,\mathrm{d}\tau + \int_0^h g(y(\tau))\,\mathrm{d}\tau, \\ \Phi_g^h(y_0) &= w(h) = y_0 + \int_0^h g(w(\tau))\,\mathrm{d}\tau, \\ \Psi^h(y_0) &= \Phi_f^h\left(\Phi_g^h(y_0)\right) = y_0 + \int_0^h g(w(\tau))\,\mathrm{d}\tau + \int_0^h f(z(\tau))\,\mathrm{d}\tau. \end{aligned}$$

Hence,

$$\Phi^h(y_0) - \Psi^h(y_0) = \int_0^h \left(f(y(\tau)) - f(z(\tau))\right)\,\mathrm{d}\tau + \int_0^h \left(g(y(\tau)) - g(w(\tau))\right)\,\mathrm{d}\tau.$$

If we can show $\|f(y(t)) - f(z(t))\| = \mathcal{O}(h)$ and $\|g(y(t)) - g(w(t))\| = \mathcal{O}(h)$ for all $t \in [0,h]$, then the claim follows with

$$\left\|\Phi^h(y_0) - \Psi^h(y_0)\right\| \leq h \left(\sup_{\tau\in[0,h]} \|f(y(\tau)) - f(z(\tau))\| + \sup_{\tau\in[0,h]} \|g(y(\tau)) - g(w(\tau))\| \right).$$

We only show $\|f(y(t)) - f(z(t))\| = \mathcal{O}(h)$:

$$\begin{aligned} y(t) - z(t) &= y(0) - z(0) + \int_0^h (y'(\tau) - z'(\tau))\,\mathrm{d}\tau \\ &= y_0 - \left(y_0 + \int_0^h w'(\tau)\,\mathrm{d}\tau\right) + \int_0^h (y'(\tau) - z'(\tau))\,\mathrm{d}\tau \\ &= \int_0^h (y'(\tau) - z'(\tau) - w'(\tau))\,\mathrm{d}\tau \end{aligned}$$

Hence,

$$\|f(y(t)) - f(z(t))\|_\infty \leq L_f \|y(t) - z(t)\|_\infty \leq L_f h \max\{\|y'\|_\infty, \|z'\|_\infty, \|w'\|_\infty\}$$

yields the claim. □

Definition 6.55 (Strang splitting). The discrete flows Ψ^h for the ODE (6.16) defined by

$$\Psi^h := \Phi_f^{h/2} \circ \Phi_g^h \circ \Phi_f^{h/2}.$$

or by

$$\Psi^h := \Phi_g^{h/2} \circ \Phi_f^h \circ \Phi_g^{h/2}.$$

are called *Strang splittings.*

Obviously, Ψ^h is reversible. Moreover, it is the combination of two Lie-Trotter splittings

$$\Psi^h_{\text{strang}} = \Psi^{h/2}_{\text{LT}} \circ \left(\Psi^{h/2}_{\text{LT}}\right)^*.$$

Hence, the convergence order is at least $p = 2$. Applied to separable Hamiltonian systems as in Ex. 6.53 results into the Störmer-Verlet methods.

CHAPTER 7

Boundary value problems

Up to now, we considered numerical solvers for initial value problems. In this chapter we introduce solvers for boundary value problems.

7.1. Comparison to initial value problems

To fix some of the main differences between initial value problems and boundary value problems, we consider solutions $y \in C^2([0,T];\mathbb{R})$ to

$$\forall x \in [0,T]: \quad y''(x) + y(x) = 0 \tag{7.1}$$

with $T > 0$. It's straightforward to show, that all solutions are given by

$$y(x) = C_1 \cos(x) + C_2 \sin(x) \tag{7.2}$$

with constants $C_1, C_2 \in \mathbb{R}$. Since up to now we considered systems of differential equations with order one, we reformulate (7.1) to the equivalent system:

$$Y' = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} Y \tag{7.3}$$

with $Y := (y, y')^\top \in C^1([0,T];\mathbb{R}^2)$. Solutions to (7.3) are

$$Y(x) = C_1 \begin{pmatrix} \cos(x) \\ -\sin(x) \end{pmatrix} + C_2 \begin{pmatrix} \sin(x) \\ \cos(x) \end{pmatrix}.$$

For an initial value problem we might solve for solutions to (7.3) such that

$$Y(0) = \begin{pmatrix} Y_1^{(0)} \\ Y_2^{(0)} \end{pmatrix}. \tag{7.4}$$

Hence, the initial value problem (7.3) together with (7.4) is uniquely solvable with solution

$$Y(x) = Y_1^{(0)} \begin{pmatrix} \cos(x) \\ -\sin(x) \end{pmatrix} + Y_2^{(0)} \begin{pmatrix} \sin(x) \\ \cos(x) \end{pmatrix}.$$

For a boundary value problem we might ask for solutions to (7.1) such that

$$y(0) = y_0 \qquad \text{and} \qquad y(T) = y_T. \tag{7.5}$$

Using (7.2) this leads to the linear system of equations

$$\begin{pmatrix} 1 & 0 \\ \cos(T) & \sin(T) \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_T \end{pmatrix} \tag{7.6}$$

for the coefficients C_1 and C_2. Obviously, (7.6) is uniquely solvable if and only if

$$\sin(T) \neq 0 \qquad \Leftrightarrow \qquad T \notin \{k\pi, k \in \mathbb{N}\}.$$

Hence, while for initial value problems the Theorem of Picard-Lindelöf guarantees unique solvability, we cannot expect this for boundary value problems.

Exercise 7.1. Show, that for all $T > 0$ there exists a κ such that the boundary value problem

$$\forall x \in [0,T]: \quad y''(x) + \kappa^2 y(x) = 0, \qquad y(0) = y(T) = 0$$

is not uniquely solvable.

Remark 7.2. (7.5) are so called *Dirichlet boundary conditions*. If we use *Neumann boundary conditions* of the form

$$y'(0) = y'_0 \qquad \text{and} \qquad y'(T) = y'_T,$$

then all functions $y \equiv C$ with a constant $C \in \mathbb{R}$ are solutions to the corresponding boundary value problem with the differential equation $u''(x) = 0$. Hence, for all T the solution is not unique.

We will not discuss in the following the solvability of the boundary value problems under consideration in this chapter. We always assume, that the problems are uniquely solvable.

But there is a big difference in the numerical methods as well. For an initial value problem, we have all necessary information for a computation at one point available. This gives rise to time stepping: We compute the numerical approximations of the solution one after the other. For boundary value problems, we have the necessary information at different points. Hence, we cannot expect to construct approximations by time stepping. We have to solve all approximations in the whole interval at the same time.

Typically, this will lead to large systems of possibly non-linear equations. In this chapter, we mainly confine ourselves to linear differential equations leading to linear systems of equations.

7.2. Shooting methods

The most simple and most flexible approach for numerical solvers of a boundary value problem in one dimension are shooting methods. We introduce this method for the following model problem: For a given interval $[a,b]$, given boundary values $y_a, y_b \in \mathbb{R}$, and given right hand side $f \in C([a,b] \times \mathbb{R} \times \mathbb{R})$ we ask for solutions $y \in C^2([a,b];\mathbb{R})$ to

$$\forall x \in [a,b]: \quad y''(x) = f\left(x, y(x), y'(x)\right), \tag{7.7a}$$

$$y(a) = y_a, \qquad y(b) = y_b. \tag{7.7b}$$

We reformulate this scalar second order equation into the first order system

$$Y'(x) = \begin{pmatrix} {y^{(1)}}'(x) \\ {y^{(2)}}'(x) \end{pmatrix} = F\left(x, Y(x)\right) := \begin{pmatrix} y^{(2)}(x) \\ f\left(x, y^{(1)}(x), y^{(2)}(x)\right) \end{pmatrix}. \tag{7.8a}$$

Moreover, we denote by $Y_s = \left(y_s^{(1)}, y_s^{(2)}\right)^\top$ for $s \in \mathbb{R}$ the solution to the initial value problem for this system combined with the initial values

$$Y_s(a) = \begin{pmatrix} y_a \\ s \end{pmatrix}. \tag{7.8b}$$

Note, that we have chosen the correct initial value $y_s^{(1)}(a) = y(a) = y_a$, and an arbitrary value $y_s^{(2)}(a) = y'(a) = s$. Our problem (7.7) can be reformulated in the following way:

$$\text{find } \tilde{s} \in \mathbb{R}: \qquad G(\tilde{s}) := y_{\tilde{s}}^{(1)}(b) - y_b = 0. \tag{7.9}$$

(7.9) is a typically nonlinear equation, which involves the solution of an initial value problem. We can solve this problem numerically with Newton's method, if we are able to compute the derivative $G'(s) = \partial_s y_s^{(1)}(b)$.

Lemma 7.3. *If y_s is the solution to the initial value problem*

$$\forall x \in [a,b]: \quad y''(x) = f\left(x, y(x), y'(x)\right), \qquad y(a) = y_a, \quad y'(a) = s, \tag{7.10}$$

then the function $v_s := \partial_s y_s$ is the solution to the initial value problem

$$v''(x) = c(x)v(x) + d(x)v'(x), \qquad v(a) = 0, \quad v'(a) = 1, \tag{7.11}$$

with

$$c(x) := f_y\left(x, y_s(x), y_s'(x)\right), \qquad d(x) := f_{y'}\left(x, y_s(x), y_s'(x)\right).$$

Proof. Differentiation of (7.10) with respect to s and interchanging the order of differentiation, e.g. $\partial_s y_s'(x) = \partial_s \partial_x y_s(x) = \partial_x \partial_s y_s(x) = v_s'(x)$, yields the result. □

The last lemma can be used for the shooting algorithm.

Algorithm 7.4 (Plain shooting method).
Input: problem description (f, a, b, y_a, y_b) for (7.7), starting value s_0.

```
1: $n := 0$
2: repeat
3:     Compute solution $y_{s_n}$ of (7.10) (numerically)
4:     Compute solution $v_{s_n}$ of (7.11) (numerically)
5:     $s_{n+1} = s_n - \frac{y_{s_n}(b) - y_b}{v_{s_n}(b)}$
6:     $n = n + 1$
7: until Stopping criteria of Newton's method
```

Output: (approximation) $y_{s_{n-1}}$ to the solution of (7.7)

As stopping criteria in line 7 one might give a tolerance TOL and use

- the residuum $\left|y_{s_{n-1}}(b) - y_b\right| \leq \text{TOL}$, or
- the update $\left|\frac{y_{s_{n-1}}(b) - y_b}{v_{s_{n-1}}(b)}\right| \leq \text{TOL}$.

Remark 7.5. If an adaptive Runge-Kutta method as described in Sec. 2.6 is used in line 3, then the discrete approximation to y_{s_n} is typically only known on the adaptive mesh. Hence, in the numerical simulation of (7.11) only this mesh can be used, since y_{s_n} and y_{s_n}' are needed for the coefficients c_{s_n} and d_{s_n}.

In order to simplify the usage of standard ODE solver packages, it can be reasonable to use a solver for the system of ODE's consisting of both problems, (7.10) and (7.11).

In Theo. 1.3 we have seen for the continuous problem, that errors might increase exponentially with the factor $\exp(L|b-a|)$, where L is the Lipschitz constant of the right hand side. The same holds true for discretizations, see e.g., (2.15). Hence, we should expect severe stability issues in Alg. 7.4, if $L|b-a|$ is large.

Example 7.6. Consider the boundary value problem

$$y'' = 35y + 2y', \qquad y(0) = y(b) = 1$$

for $b > 0$. It is straightforward to show, that

$$y(x) = C\exp(7x) + D\exp(-5x)$$

for arbitrary constants $C, D \in \mathbb{R}$ is a solution of the differential equation. The solution to the boundary value problem is given by the constants

$$C = \frac{1 - \exp(-5b)}{\exp(7b) - \exp(-5b)}, \qquad D = \frac{\exp(7b) - 1}{\exp(7b) - \exp(-5b)}.$$

We compute

$$\tilde{s} := y'(0) = 7C - 5D = \frac{12 - 7\exp(-5b) - 5\exp(7b)}{\exp(7b) - \exp(-5b)} = -5 + \frac{12\,(1 - \exp(-5b))}{\exp(7b) - \exp(-5b)}.$$

Hence, $\tilde{s} \approx -5$ for large values of b.

The solution to the initial value problem

$$y'' = 35y + 2y', \qquad y(0) = 1, \quad y'(0) = s$$

is given by

$$y_s(x) = \frac{s+5}{12}\exp(7x) + \frac{7-s}{12}\exp(-5x).$$

A perturbation $s = \tilde{s} + \epsilon$ leads to an error

$$y_{\tilde{s}+\epsilon}(b) - 1 = \frac{\exp(7b) - \exp(-5b)}{12}\epsilon.$$

Already $b = 6$ and $|\epsilon| \approx 10^{-15}$ (maybe a rounding error) leads to the very large error ≈ 145.

As we have seen in the last example, the shooting method is extremely sensitive to errors, if $L|b - a|$ is large. One remedy is to split the interval $[a, b]$ into small intervals. Let $a = x_0 < x_1 < \cdots < x_N = b$ be such a mesh, and $y_i := y(x_i)$ for $i = 0, \ldots, N$. Since $y_0 = y_a$ and $y_N = y_b$ are known, we have $N - 1$ unknowns. Additionally, we introduce the N unknowns $s_i := y'(x_i)$ for $i = 0, \ldots, N - 1$.

Let us denote with y_{x_i,y_i,s_i} the solution to the initial value problems for each interval $[x_i, x_{i+1}]$, i.e., for $i = 0, \ldots, N - 1$

$$\forall x \in [x_i, x_{i+1}] : \quad y''(x) = f\left(x, y(x), y'(x)\right), \qquad y(x_i) = y_i, \quad y'(x_i) = s_i.$$

In order to guarantee that the solutions are at least continuously differentiable on $[a, b]$, we formulate the (typically non-linear) system of equations:

$$y_{x_i,y_i,s_i}(x_{i+1}) = y_{i+1}, \qquad i = 0, \ldots, N - 1 \tag{7.12a}$$

$$y'_{x_i,y_i,s_i}(x_{i+1}) = s_{i+1}, \qquad i = 0, \ldots, N - 2. \tag{7.12b}$$

(7.12) consists of $2N - 1$ non-linear equations for the $2N - 1$ unknowns $y_1, \ldots, y_{N-1}$, and $s_0, \ldots, s_{N-1}$. This system can be solved with Newton's method. The Jacobian can be computed as in Lem 7.3 by implicit differentiation.

7.3. Finite difference method

The finite difference method is one of the most popular numerical methods for linear differential equations.

7.3.1. One-dimensional problems. We are looking for solutions $y \in C^2([a,b])$ to the model problem

$$\forall x \in [a,b]: \quad \alpha y''(x) + \beta y'(x) + \gamma y(x) = f(x), \tag{7.13a}$$
$$y(a) = y_a, \tag{7.13b}$$
$$y'(b) = y_b, \tag{7.13c}$$

for given constants $\beta, \gamma, y_a, y_b \in \mathbb{R}$, a constant $\alpha \neq 0$, a bounded interval $[a,b] \subset \mathbb{R}$, and a function $f \in C([a,b])$.

LEMMA 7.7. *For $N \in \mathbb{N}$ let $x_j := a + jh$, $j = 0, \ldots, N$, with mesh-size $h := \frac{b-a}{N}$ be a uniform grid of the interval $[a,b]$. Then for $j = 1, \ldots, N-1$ and $y \in C^4([a,b])$ there holds*

$$y'(x_j) = \frac{y(x_{j+1}) - y(x_{j-1})}{2h} + \mathcal{O}(h^2), \tag{7.14a}$$
$$y''(x_j) = \frac{y(x_{j+1}) - 2y(x_j) + y(x_{j-1})}{h^2} + \mathcal{O}(h^2). \tag{7.14b}$$

Moreover, there holds

$$y'(b) - \frac{h}{2}y''(b) = \frac{y(b) - y(b-h)}{h} + \mathcal{O}(h^2), \tag{7.14c}$$
$$y'(a) + \frac{h}{2}y''(a) = \frac{y(a+h) - y(a)}{h} + \mathcal{O}(h^2). \tag{7.14d}$$

PROOF. Taylor expansion of y yields the result, since

$$y(x_{j+1}) = y(x_j + h) = y(x_j) + hy'(x_j) + \frac{1}{2}h^2 y''(x_j) + \frac{1}{6}h^3 y'''(x_j) + \mathcal{O}(h^4),$$
$$y(x_{j-1}) = y(x_j - h) = y(x_j) - hy'(x_j) + \frac{1}{2}h^2 y''(x_j) - \frac{1}{6}h^3 y'''(x_j) + \mathcal{O}(h^4).$$

□

For the finite difference method, we replace in (7.13a) the derivatives by their approximations (7.14). The approximations $y_j \approx y(x_j)$, $j = 0, \ldots, N$, are the solution to the linear system of equations

$$\forall j = 1, \ldots, N-1: \quad \alpha \frac{y_{j+1} - 2y_j + y_{j-1}}{h^2} + \beta \frac{y_{j+1} - y_{j-1}}{2h} + \gamma y_j = f(x_j). \tag{7.15a}$$

(7.13b) leads to the equation

$$y_0 = y_a. \tag{7.15b}$$

For the Neumann boundary condition (7.13c), the differential equation leads to

$$y'(b) - \frac{h}{2}y''(b) = \left(1 + h\frac{\beta}{2\alpha}\right) y'(b) + h\frac{\gamma}{2\alpha} y(b) - h\frac{1}{2\alpha} f(b).$$

(7.14c) can be used to construct the second order approximation

$$\left(1 + h\frac{\beta}{2\alpha}\right) y_b + h\frac{\gamma}{2\alpha} y_N - h\frac{1}{2\alpha} f(b) = \frac{y_N - y_{N-1}}{h}. \tag{7.15c}$$

All together, the Equations (7.15) lead to the system of N linear equations for the N unknowns $y_1, \ldots, y_N$:

$$\left(-\frac{2\alpha}{h^2} + \gamma\right) y_1 + \left(\frac{\alpha}{h^2} + \frac{\beta}{2h}\right) y_2 = f(x_1) - \left(\frac{\alpha}{h^2} + \frac{\beta}{2h}\right) y_a, \tag{7.16a}$$

for all $j = 2, \dots, N-1$

$$\left(\frac{\alpha}{h^2} + \frac{\beta}{2h}\right) y_{j-1} + \left(-\frac{2\alpha}{h^2} + \gamma\right) y_j + \left(\frac{\alpha}{h^2} + \frac{\beta}{2h}\right) y_{j+1} = f(x_j), \tag{7.16b}$$

and

$$-\frac{1}{h} y_{N-1} + \left(\frac{1}{h} - h\frac{\gamma}{2\alpha}\right) y_N = \left(1 + h\frac{\beta}{2\alpha}\right) y_b - h\frac{1}{2\alpha} f(b). \tag{7.16c}$$

EXAMPLE 7.8 (1D Poisson problem). The choice $\alpha = -1$ and $\beta = \gamma = 0$ leads to the one-dimensional Poisson problem

$$\forall x \in [a,b] : -y''(x) = f(x), \qquad y(a) = y_a \quad y'(b) = y_b.$$

The discretization with finite differences (7.16) becomes

$$A_h y_h = g_h \tag{7.17}$$

with

$$\begin{aligned} A_h &:= \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{N\times N} \\ y_h &:= (y_1, \cdots, y_N)^\top \in \mathbb{R}^N, \\ g_h &:= \left(f(x_1) + \frac{1}{h^2} y_a, f(x_2), \cdots, f(x_{N-1}), \frac{1}{h} y_b + \frac{1}{2} f(b)\right)^\top \in \mathbb{R}^N. \end{aligned} \tag{7.18}$$

Note, that A_h is except for the last row and the sign the same matrix as in Example 4.2 (heat equation). The difference to the matrix $\mathbf{M}_h$ in Example 4.2 arises due to the Neumann bounday condition at the right boundary.

REMARK 7.9. Lemma 7.7 does not guarantee convergence of $y_j - y(x_j) \to 0$ for $h \to \infty$, since we have to solve a linear system of equations with a system matrix depending on h. We have to show, that the inverse of this matrix can be uniformly bounded in a suitable norm. We postpone this to Section 7.3.3.

7.3.2. Finite Differences on rectangular grids. Let us assume for simplicity, that we are looking for solutions y to the Poisson problem with homogeneous Dirichlet boundary conditions

$$\forall x \in \Omega : \quad -\Delta y(x) = f(x), \tag{7.19a}$$

$$\forall x \in \partial\Omega : \quad y(x) = 0, \tag{7.19b}$$

with a rectangular domain $\Omega = [a_1, b_1] \times [a_2, b_2]$ and a given function $f \in C(\Omega)$. We introduce a tensor product mesh $\{x_{jk}, j = 0, \dots, N_1, k = 0, \dots, N_2\}$ of Ω with

$$x_{jk} := \begin{pmatrix} a_1 + jh_1 \\ a_2 + kh_2 \end{pmatrix}, \qquad h_1 := \frac{b_1 - a_1}{N_1}, \quad h_2 := \frac{b_2 - a_2}{N_2}.$$

For the Laplace operator $-\Delta = -\partial_{x_1}^2 - \partial_{x_2}^2$ we use again an approximation with finite differences, i.e.

$$\begin{aligned} -\Delta y(x_{jk}) =& \frac{1}{h_1^2} \left(-y(x_{j-1,k}) + 2y(x_{j,k}) - y(x_{j+1,k})\right) + \mathcal{O}(h_1^2) \\ &+ \frac{1}{h_2^2} \left(-y(x_{j,k-1}) + 2y(x_{j,k}) - y(x_{j,k+1})\right) + \mathcal{O}(h_2^2). \end{aligned} \tag{7.20}$$

Due to the homogeneous Dirichlet boundary condition this leads to the $(N_1-1)(N_2-1)$ linear equations for the $(N_1-1)(N_2-1)$ unknowns $y_{jk} \approx y(x_{jk})$ for $j = 1, \dots, N_1 - 1$ and $k = 1, \dots, N_2 - 1$

$$-\frac{1}{h_1^2} y_{j-1,k} - \frac{1}{h_1^2} y_{j+1,k} - \frac{1}{h_2^2} y_{j,k-1} - \frac{1}{h_2^2} y_{j,k+1} + \left(\frac{2}{h_1^2} + \frac{2}{h_2^2} \right) y_{jk} = f(x_{jk}). \tag{7.21}$$

If $h_1 = h_2 = h$, the term on the left hand side of often noted by the 5 point stencil

$$-\Delta_h y(x_j, x_k) := \frac{1}{h^2} \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} y(x_j, x_k).$$

Remark 7.10. Non-homogeneous Dirichlet boundary conditions as well as Neumman boundary conditions can be used with minor but technical modifications.

It is straightforward to generalize this construction to three dimensions. Note, that the resulting linear system of equations becomes quite large. On the other hand, the system matrix is sparse and at least for the Poisson problem positive definite. Hence, one might use the conjugate gradient method as an iterative solver.

7.3.3. Convergence of one-dimensional finite difference methods. We present the convergence analysis of the finite difference method only for the one-dimensional Poisson problem 7.8. We notice in the end of this section, how the analysis can be generalized.

First, we put the finite difference discretization of the Poisson problem into an abstract framework. To this end, we introduce

- a normed solution space $(X, \|\cdot\|_X)$,
- a normed data space $(Y, \|\cdot\|_Y)$,
- a normed discrete solution space $(X_h, \|\cdot\|_{X_h})$,
- a normed discrete data space $(Y_h, \|\cdot\|_{Y_h})$,
- a bounded, linear operator $A : X \to Y$,
- a discrete linear operator $A_h : X_h \to Y_h$,
- an interpolation operator $I_h^X : X \to X_h$, and
- an interpolation operator $I_h^Y : Y \to Y_h$.

For the finite difference discretization of the Poisson problem, the solution space for the function y is $X = C^2([a,b])$ with a suitable norm. As input data, we have the function f and the boundary values y_a, y_b. Hence, the data space is $Y = C([a,b]) \times \mathbb{R} \times \mathbb{R}$ with a suitable norm. The linear operator $A : X \to Y$ is given by

$$X \ni y \mapsto Ay := (-y'', y(a), y'(b)) \in Y. \tag{7.22}$$

The discrete solution and data spaces are $X_h = Y_h = \mathbb{R}^N$. Note, that we will introduce later on the norms for X_h and Y_h. The discrete operator A_h is given by its matrix representation in (7.18). The interpolation operators can be derived from (7.18) as well:

$$X \ni y \mapsto I_h^X y := (y(x_1), \dots, y(x_N))^\top \in X_h, \tag{7.23a}$$

$$Y \ni (f, y_a, y_b) \mapsto I_h^Y(f, y_a, y_b) := \left(f(x_1) + \frac{1}{h^2} y_a, f(x_2), \cdots, f(x_{N-1}), \frac{1}{h} y_b + \frac{1}{2} f(b) \right)^\top \in Y_h. \tag{7.23b}$$

We can denote all definitions in the approximation diagram

$$\begin{tikzcd} X \arrow[r, "A"] \arrow[d, "I_h^X"'] & Y \arrow[d, "I_h^Y"] \\ X_h \arrow[r, "A_h"] & Y_h \end{tikzcd}.$$

Lemma 7.11. *With the preceding abstract framework let A and A_h have bounded inverses $A^{-1} : Y \to X$ and $A_h^{-1} : Y_h \to X_h$. Moreover let $g \in Y$.*

Then there holds for the solutions $y \in X$ to $Ay = g$ and $y_h \in X_h$ to $A_h y_h = I_h^Y g$

$$\left\| I_h^X y - y_h \right\|_{X_h} \le \left\| A_h^{-1} \right\|_{Y_h \to X_h} \left\| \left(A_h I_h^X - I_h^Y A \right) y \right\|_{Y_h}. \tag{7.24}$$

Proof. The result is a consequence of

$$A_h y_h = I_h^Y g = I_h^Y A y,$$

since

$$A_h \left(I_h^X y - y_h \right) = A_h I_h^X y - I_h^Y A y. \tag{7.25}$$

□

(7.25) describes the discrete residuum. The interpolation $I_h^X y$ of the exact solution y is put into the discrete equation $A_h y_h = I_h^Y g$. For multi-step methods, we have defined the truncation error in Def. 5.10 in the same way. Later on in Lemma 5.12 we have shown, that the truncation error is equivalent to the consistency error. Hence, we might call the second term

$$\left\| \left(A_h I_h^X - I_h^Y A \right) y \right\|_{Y_h}$$

on the right hand side of (7.24) the consistency error of the finite difference method applied to the Poisson problem. Consequently, the first term $\left\| A_h^{-1} \right\|_{Y_h \to X_h}$ should be called stability term.

Definition 7.12. We call the finite difference method stable, if there exists constants $C, h_0 > 0$ such that

$$\forall h \in (0, h_0) : \quad \left\| A_h^{-1} \right\|_{Y_h \to X_h} \le C. \tag{7.26}$$

Note, that up to now the norms of X_h and Y_h are not fixed. We have to find proper norms, such that

- the finite difference method is stable, and
- such that the consistency error vanishes for $h \to 0$.

Before we introduce these norms, we motivate our choice with the weak setting of the continuous setting.

Remark 7.13 (weak form of the Poisson problem). We multiply the differential equation of Example 7.8 by a test function w, integrate over the domain $[a, b]$, and use partial integration to obtain

$$\int_a^b f(x) w(x) \,\mathrm{d}x = \int_a^b -y''(x) w(x) \,\mathrm{d}x = \int_a^b y'(x) w'(x) \,\mathrm{d}x - y'(b) w(b) + y'(a) w(a).$$

Using the boundary conditions, we can rewrite the continuous problem in the following form: Find y with $y(a) = y_a$ such that

$$\int_a^b y'(x) w'(x) \,\mathrm{d}x = \int_a^b f(x) w(x) \,\mathrm{d}x + y_b w(b) \tag{7.27}$$

for all functions w with $w(a) = 0$.

The weak form indicates to chose the Sobolev spaces $L_2((a,b))$ with scalar product

$$(f,g)_{L_2((a,b))} := \int_a^b f(x)g(x)\,\mathrm{d}x$$

for the right hand side f, and $H^1((a,b))$ with scalar product

$$(f,g)_{H^1((a,b))} := (f,g)_{L_2((a,b))} + (f',g')_{L_2((a,b))}$$

for the functions y, w. If we skip the L^2-term in the scalar product of H^1, we get the semi-norm

$$|f|_{H^1((a,b))} := \left(\int_a^b |f'(x)|^2\,\mathrm{d}x\right)^{\frac{1}{2}} = \|f'\|_{L_2((a,b))}.$$

It's only a semi-norm, because it vanishes for all constant functions. Hence, it is not positive definite in general. But since we use a Dirichlet boundary condition at the interval boundary a, this can be fixed by the Friedrich's inequality.

Lemma 7.14 (Friedrich's inequality). *Let $w \in C^1([a,b])$ with $w(a) = 0$. Then*

$$\|w\|_{L_2((a,b))} \le \frac{b-a}{\sqrt{2}} |w|_{H^1((a,b))}.$$

Proof. The Theorem of Cauchy-Schwartz yields

$$|w(x)| \le \underbrace{|w(a)|}_{=0} + \left|\int_a^x w'(x)\,\mathrm{d}x\right| \le \left(\int_a^x 1^2\,\mathrm{d}s\right)^{\frac{1}{2}} \left(\int_a^x w'(s)^2\,\mathrm{d}s\right)^{\frac{1}{2}} \le \sqrt{x-a}\,\|w'\|_{L^2((a,b))}.$$

Hence, the claim follows with

$$\|w\|^2_{L_2((a,b))} = \int_a^b w(x)^2\,\mathrm{d}x \le \|w'\|^2_{L^2((a,b))} \int_a^b (x-a)\,\mathrm{d}x = \|w'\|^2_{L^2((a,b))} \frac{(b-a)^2}{2}.$$

□

Remark 7.15. The Friedrich's inequality can be used to show, that the weak form (7.27) of the Poisson equation is uniquely solvable. In particular, it can be shown that the bilinear form on the left hand side is continuous (trivial by the definition of the norms), and positive definite for functions w with $w(a) = 0$ and $w \not\equiv 0$, since

$$\int_a^b (w'(x))^2\,\mathrm{d}x = |w|^2_{H^1((a,b))} \ge \frac{2}{(b-a)^2} \|w\|^2_{L_2((a,b))} > 0.$$

The last remark gives hope, that we can prove stability of our method as defined in Def. 7.12, if we introduce the discrete scalar products for $v_h = (v_1, \dots, x_N)^\top, w^h = (w_1, \dots, w_N)^\top \in \mathbb{R}^N$

$$(v_h, w_h)_{0,h} := h \sum_{j=1}^N v_j w_j, \tag{7.28a}$$

$$(v_h, w_h)_{1,h} := \frac{1}{h}\left(v_1 w_1 + \sum_{j=2}^N (v_j - v_{j-1})(w_j - w_{j-1})\right). \tag{7.28b}$$

Note, that the discrete scalar product $(v_h, w_h)_{0,h}$ is an approximation to the $L_2((a,b))$ scalar product by Riemann sums

$$(v,w)_{L^2} = \int_a^b v(x)w(x)\,\mathrm{d}x \approx h\sum_{j=1}^N v(x_j)w(x_j) = \left(I_h^X v, I_h^X w\right)_{0,h}.$$

The same holds true for the discrete scalar product $(v_h, w_h)_{1,h}$ and the $H^1((a,b))$ semi-definite bilinear form, if the derivative is replaced by the finite difference quotient and if $v(a) = w(a) = 0$ is used

$$\begin{aligned}\int_a^b v'(x)w'(x)\,\mathrm{d}x &\approx h\sum_{j=1}^N v'(x_j)w'(x_j)\\ &\approx h\sum_{j=1}^N \frac{v(x_j)-v(x_j-h)}{h}\frac{w(x_j)-w(x_j-h)}{h}\\ &= \left(I_h^X v, I_h^X w\right)_{1,h}.\end{aligned}$$

Lemma 7.16. *Let A_h be the matrix defined in* (7.18). *Then*

$$\forall v_h, w_h \in \mathbb{R}^N : \quad (A_h v_h, w_h)_{0,h} = (v_h, w_h)_{1,h}.$$

Proof. For $v_h = (v_1, \dots, x_N)^\top, w^h = (w_1, \dots, w_N)^\top \in \mathbb{R}^N$ there holds

$$\begin{aligned}(A_h v_h, w_h)_{0,h} =& h\sum_{j=1}^N (A_h v_h)_j\, w_j\\ =& \frac{1}{h}\left\{(2v_1 - v_2)\,w_1 + \sum_{j=2}^{N-1}(-v_{j-1} + 2v_j - v_{j+1})\,w_j + (-v_{N-1} + v_N)\,w_N\right\}\\ =& \frac{1}{h}\left\{v_1 w_1 + \sum_{j=2}^N (v_j - v_{j-1})\,w_j + \sum_{j=1}^{N-1}(v_j - v_{j+1})\,w_j\right\}\\ =& \frac{1}{h}\left\{v_1 w_1 + \sum_{j=2}^N (v_j - v_{j-1})\,(w_j - w_{j-1})\right\}\\ =& (v_h, w_h)_{1,h}.\end{aligned}$$

□

Lemma 7.17 (discrete Friedrich's inequality). *There holds*

$$\forall w_h \in \mathbb{R}^N : \qquad \|w_h\|_{0,h} \le (b-a)\,\|w_h\|_{1,h}. \tag{7.29}$$

Proof. Let $w^h = (w_1, \dots, w_N)^\top \in \mathbb{R}^N$ and for simplicity $w_0 := 0$. Then for all $k = 1, \dots, N$ the Cauchy-Schwartz inequality implies

$$\begin{aligned}|w_k| \le \left|\sum_{j=1}^k (w_j - w_{j-1})\right| &= \left(\sum_{j=1}^k 1^2\right)^{\frac{1}{2}}\left(\sum_{j=1}^k (w_j - w_{j-1})^2\right)^{\frac{1}{2}}\\ &= \sqrt{kh}\left(\frac{1}{h}\sum_{j=1}^k (w_j - w_{j-1})^2\right)^{\frac{1}{2}} \le \sqrt{kh}\,\|w_h\|_{1,h}.\end{aligned}$$

Since $hN = b-a$ and $h \leq b-a$, the claim follows with

$$\|w_h\|_{0,h}^2 = h\sum_{k=1}^N w_k^2 \leq h^2 \underbrace{\sum_{k=1}^N k}_{=\frac{N(N+1)}{2}} \|w_h\|_{1,h}^2 = (b-a)\frac{(b-a)+h}{2}\|w_h\|_{1,h}^2 \leq (b-a)^2\|w_h\|_{1,h}^2 .$$

□

Note, that the homogeneous Dirichlet boundary condition $w_0 = w(0) = 0$ was already included in the definition of the scalar products.

LEMMA 7.18 (stability). *Let $A_h \in \mathbb{R}^{N\times N}$ be the matrix defined in* (7.18). *Then A_h is symmetric and positive definite. Moreover,*

$$\forall h > 0: \qquad \left\|A_h^{-1}\right\|_{Y_h\to X_h} \leq (b-a)^2, \tag{7.30}$$

if X_h and Y_h are equipped with the norm $\|\cdot\|_{0,h}$.

PROOF. Obviously, the matrix A_h is symmetric. Using the last two lemmata, we have

$$w_h^\top A_h w_h = \frac{1}{h}\left(A_h w_h, w_h\right)_{0,h} = \frac{1}{h}\left(w_h, w_h\right)_{1,h} \geq \frac{1}{(b-a)^2 h}\|w_h\|_{0,h}^2 .$$

Hence, the matrix is positive definite. This implies, that the matrix is also regular and that there exists a unique solution y_h to $A_h y_h = g_h$ for all $g_h \in Y_h$. Finally,

$$\frac{1}{(b-a)^2}\|y_h\|_{0,h}^2 \leq \|y_h\|_{1,h}^2 = \left(A_h y_h, y_h\right)_{0,h} = \left(g_h, y_h\right)_{0,h} \leq \|g_h\|_{0,h}\|y_h\|_{0,h}$$

yields

$$\left\|A_h^{-1}\right\|_{Y_h\to X_h} = \sup_{g_h\in Y_h\setminus\{0\}} \frac{\left\|A_h^{-1}g_h\right\|_{0,h}}{\|g_h\|_{0,h}} \leq (b-a)^2.$$

□

In other words, the finite difference method applied to the Poisson problem is stable in the sense of Def. 7.12 for all $h > 0$ and with stability constant $C = (b-a)^2$. We are left with the task to bound the consistency error with respect to these norms.

REMARK 7.19. In Ex. 4.2 we have used Dirichlet boundary condition leading to a similar matrix $\mathbf{M}_h$. The eigenvalues of $\mathbf{M}_h$ are given by $\lambda_j = \frac{2}{h^2}\left(-1+\cos(\frac{j\pi}{N})\right)$. $-\mathbf{M}_h$ is symmetric and positive definite as well. Moreover,

$$\left\|\mathbf{M}_h^{-1}\right\|_{Y_h\to X_h} = -\frac{1}{\lambda_1} = -\frac{h^2}{-1+1-\frac{1}{2}\left(\frac{h}{b-a}\pi\right)^2+\mathcal{O}(h^4)} \approx \frac{2}{\pi^2}(b-a)^2.$$

Hence, the finite difference method is stable for the Poisson problem with pure Dirichlet boundary conditions as well.

LEMMA 7.20 (consistency). *With the definitions of this section there holds*

$$\forall y \in C^4([a,b]) \quad \exists C > 0: \qquad \left\|A_h I_h^X - I_h^Y A u\right\|_{Y_h} \leq C h^{\frac{3}{2}} \tag{7.31}$$

for all sufficiently small h.

Proof. Using the definition of the operator A in (7.22), the discrete operator A_h in (7.18), and of the interpolation operators in (7.23), we have

$$\left(I_h^Y Ay\right)_j = \begin{cases} -y''(x_1) + \frac{1}{h^2}y(x_0), & j = 1, \\ -y''(x_j), & j = 2, \dots, N-1, , \\ \frac{1}{h}y'(x_N) - \frac{1}{2}y''(x_N), & j = N \end{cases}$$

$$\left(A_h I_h^X y\right)_j = \begin{cases} \frac{2}{h^2}y(x_1) - \frac{1}{h^2}y(x_2), & j = 1, \\ -\frac{1}{h^2}y(x_{j-1}) + \frac{2}{h^2}y(x_j) - \frac{1}{h^2}y(x_{j+1}), & j = 2, \dots, N-1, . \\ -\frac{1}{h^2}y(x_{N-1}) + \frac{1}{h^2}y(x_N), & j = N \end{cases}$$

Due to Lemma 7.7 there exists a constant $\tilde{C} > 0$ such that

$$\left|\left(I_h^Y Ay\right)_j - \left(A_h I_h^X y\right)_j\right| \le \begin{cases} \tilde{C}h^2, & j = 1, \dots, N-1, \\ \tilde{C}h, & j = N. \end{cases}$$

Note, that we lose one power of h for $j = N$, since (7.14c) is divided by h. Finally, $h(N-1) \le b - a$ yields

$$\begin{aligned} \left\|A_h I_h^X - I_h^Y Au\right\|_{Y_h}^2 &= \left\|A_h I_h^X - I_h^Y Au\right\|_{0,h}^2 = h \sum_{j=1}^{N} \left|\left(I_h^Y Ay\right)_j - \left(A_h I_h^X y\right)_j\right|^2 \\ &\le \tilde{C}h\left((N-1)h^4 + h^2\right) \le \tilde{C}\left((b-a)h^4 + h^3\right). \end{aligned}$$

The claim follows for $h \le 1$ with $C := \sqrt{\tilde{C}(b-a+1)}$. □

Note, that for pure Dirichlet boundary conditions we would have an error $\mathcal{O}(h^2)$.

Theorem 7.21 (convergence). *Let y be the unique solution to the one-dimensional Poisson problem 7.8, and $y_h \in \mathbb{R}^N$ the unique solution to the finite difference discretization of the Poisson problem. Moreover, let $y \in C^4([a,b])$. Then there exists a constant C, which depends on y, such that for all sufficiently small $h > 0$*

$$\left\|y_h - I_h^X y\right\|_{0,h} \le Ch^{\frac{3}{2}}. \tag{7.32}$$

In other words, for $y_h = (y_1, \dots, y_N)^\top$ there holds for the absolute error

$$\left(\sum_{j=1}^{N} |y_j - y(x_j)|^2\right)^{\frac{1}{2}} \le Ch.$$

Proof. The results is a consequence of the abstract convergence Lemma 7.11, the stability Lemma 7.18, and the consistency Lemma 7.20. □

Remark 7.22. One might ask, whether $\|\cdot\|_{0,h} = h\|\cdot\|_2$ or the standard Euclidean norm $\|\cdot\|_2$ is a meaningful norm to measure the error. We have shown after the definition of the norm $\|\cdot\|_{0,h}$ in (7.28), that this norm is an approximation for the L_2 norm. Moreover, for $y \equiv 1$ we compute $\left\|I_h^X y\right\|_2 = \sqrt{N} = \sqrt{\frac{b-a}{h}}$. Hence, the error measured in $\|\cdot\|_{0,h}$ is a relative error, i.e.,

$$\left\|y_h - I_h^X y\right\|_{0,h} \sim \frac{\left\|y_h - I_h^X y\right\|_2}{\left\|I_h^X y\right\|_2}.$$

Therefore, the norm $\|\cdot\|_{0,h}$ is a quite natural norm for measuring the error of the finite difference method.

In constrast to the presented methods for initial value problems, the stability of a numerical scheme for a boundary value problem is strongly related to the equation to solve. We cannot expect to show stability, if we only assume that the differential equation is e.g. Lipschitz. For the Poisson problem with Dirichlet boundary value at least on some parts of the boundary, stability can be derived, since the discretization matrix is positive definite and the lowest eigenvalue does not depend on h, see Rem. 7.19.

The analysis presented in this section can be easily generalized for differential equations of the form

$$-\alpha y''(x) + \gamma y(x) = f(x),$$

whereas the constants α and γ have the same sign. In this case, the Friedrich's inequality would not be needed. The matrix $A_h + \gamma\,\mathrm{id}$ is positive (or negative) definite even for pure Neumann boundary conditions. Nevertheless, in many cases we are interested in solutions to the so called Helmholtz equation

$$-\Delta y - \kappa^2 y = 0,$$

with wavenumber $\kappa > 0$. For these equations, the discretization matrices are indefinite in general. Hence, the convergence analysis for such problems becomes more challenging.

For the convergence analysis of finite difference methods for the Poisson problem in more dimensions, we refer to [**DW11**]. For tensor product meshes, the analysis presented in this section can be easily generalized.

7.4. Finite element method

In this section we give a brief overview over the finite element method. It is the most powerfull and most popular numerical method for boundary value problems in more than one dimension. A full convergence analysis is out of skope for this introduction. But we try to give an idea of some tools from functional analysis, which well be needed for finite element methods.

7.4.1. One-dimensional problems. Similar to the presentation of the finite difference method, we start to develop the method for the one-dimensional problem (7.13). More precisely, we start with the weak form of this problem (compare to Rem. 7.13): Find y with $y(a) = y_a$ such that for all functions w with $w(a) = 0$

$$a(y, w) = l(w), \tag{7.33}$$

with the bilinear form $a(\cdot,\cdot)$ defined by

$$a(v, w) := -\alpha \int_a^b y'(x)w'(x)\,\mathrm{d}x + \beta \int_a^b y'(x)w(x)\,\mathrm{d}x + \gamma \int_a^b y(x)w(x)\,\mathrm{d}x, \tag{7.34a}$$

and the linear form $l(\cdot)$ defined by

$$l(w) := \int_a^b f(x)w(x)\,\mathrm{d}x - \alpha y_b w(b). \tag{7.34b}$$

In order to discretize this problem, let V_h be a function space with dimension $N < \infty$ such that for w_h in V_h there holds $w(a) = 0$. Moreover, let $\tilde{y}$ be a function with $\tilde{y}(a) = y_a$. Note, that $\tilde{y}$ will not satisfy the differential equation in general. Now, the discrete problem reads as folows: Find $y_h \in \tilde{y} + V_h := \{\tilde{y} + v_h,\, v_h \in V_h\}$ such that

$$\forall w_h \in V_h : \qquad a(y_h, w_h) = l(w_h). \tag{7.35}$$

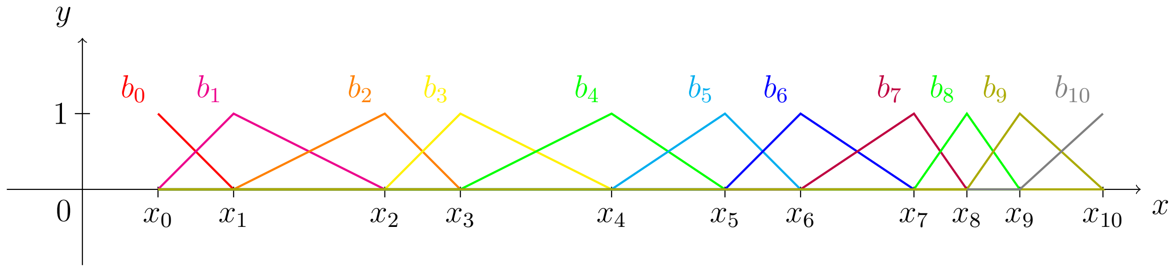


FIGURE 7.4.1. hat functions on non-uniform grids

For a given basis $\{b_1, \dots, b_N\}$ of V_h, it is sufficient to use the test functions $w_h = b_j$ for $j = 1, \dots, N$ in (7.35). Moreover, there exists a coefficient vector $a_h = (a_1, \dots, a_N)^\top \in \mathbb{R}^N$ such that

$$y_h = \tilde{y} + \sum_{k=1}^{N} a_k b_k.$$

Finally, we define the system matrix $A_h = (a(b_k, b_j))_{j,k\in\mathbb{N}} \in \mathbb{R}^{N\times N}$ and the right hand side $g_h = (l(b_1) - a(\tilde{y}, b_1), \dots, l(b_N) - a(\tilde{y}, b_N))^\top \in \mathbb{R}^N$. All together, (7.35) is equivalent to the linear system of equations

$$A_h a_h = g_h. \tag{7.36}$$

We still have to chose V_h, the function $\tilde{y}$, and the basis functions b_j for $j = 1, \dots, N$. We have to define them such that

- the terms $a(b_k, b_j)$, $l(b_k)$ are well defined,
- and we are able to prove convergence $y_h \to y$ in some proper norm.

Moreover, it would be nice to have a discretization, which

- is flexible,
- converges fast,
- leads to sparse matrices A_h,
- and where the linear system of equations can be solved efficiently.

DEFINITION 7.23 (hat functions). Let $a = x_0 < x_1 < \dots < x_N = b$ be a non-uniform mesh on the intervall $[a, b]$. Moreover, we define the mesh-sizes $h_j := x_{j+1} - x_j$ for $j = 0, \dots, N-1$. The so called hat functions sketched in Fig. 7.4.1 are defined by

$$b_0(x) := \begin{cases} \frac{x_1 - x}{h_1}, & x \in [x_0, x_1] \\ 0, & x \geq x_1 \end{cases},$$

$$b_j(x) := \begin{cases} \frac{x - x_{j-1}}{h_{j-1}}, & x \in [x_{j-1}, x_j] \\ \frac{x_{j+1} - x}{h_j}, & x \in [x_j, x_{j+1}] \\ 0, & x \notin [x_{j-1}, x_{j+1}] \end{cases}, \qquad j = 1, \dots, N-1,$$

$$b_N(x) := \begin{cases} \frac{x - x_{N-1}}{h_{N-1}}, & x \in [x_{N-1}, x_N] \\ 0, & x \leq X_{N-1} \end{cases}.$$

REMARK 7.24. For the given mesh $a = x_0 < x_1 < \dots < x_N = b$, the hat function are a basis of the linear spline space

$$\mathbb{S}^1 := \left\{ s \in C([a, b]) : \forall j = 0, \dots, N-1 : s|_{[x_{j-1}, x_j]} \in \mathbb{P}_1 \right\}.$$

More precisely, they build the so called nodal basis, since $b_j(x_j) = 1$ and $b_j(x_k) = 0$ for $j \neq k$. Hence, the solution of the interpolation problem

$$\text{find } s \in \mathbb{S}^1 : \quad s(x_j) = f_j, \qquad j = 0, \ldots, N, \tag{7.37}$$

is given by

$$s = \sum_{j=0}^{N} f_j b_j. \tag{7.38}$$

If $f_j := f(x_j)$ for a function $f \in C^2([a,b])$, the interpolation error is due to (B.4) bounded by

$$\|s - f\|_\infty \leq \frac{\|f''\|_\infty}{8} h^2 \tag{7.39}$$

with the maximal mesh-size $h := \max\{h_0, \ldots, h_{N-1}\}$.

Obviously, the hat functions defined in Def. 7.23 are not differentiable at the meshpoints x_j for $j = 1, \ldots, N-1$. Hence, at first glance the term $a(b_k, b_j)$ seems to be not well defined, if these basis functions are used. On the other hand, we can split the integrals into intergrals over the intervals of the mesh, e.g.,

$$\int_a^b b_k'(x) b_j'(x)\,\mathrm{d}x = \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} b_k'(x) b_j'(x)\,\mathrm{d}x.$$

On each intervall $[x_j, x_{j+1}]$ the basis functions are linear polynomials and therefore the integrals are well defined.

Remark 7.25. Using the L_2-setting, the basis functions b_j are *weakly differentiable*, since the value of b_j' at subdomains with measure 0, in our case at the meshpoints, do not contribute to the value of the integral. More precisely, $b_j \in H^1([a,b])$ for all $j = 0, \ldots, N$. The bilinear form $a(\cdot,\cdot)$ is well defined and continuous on $H^1([a,b]) \times H^1([a,b])$.

In the following we use $\tilde{y} := y_a b_0$ and $V_h := \mathbf{span}\{b_1, \ldots, b_N\}$. The basis functions b_j have only a local support $\operatorname{supp} b_j = [x_{j-1}, x_{j+1}]$ for $j = 1, \ldots, N-1$, and $\operatorname{supp} b_N = [x_{N-1}, x_N]$. Hence, the matrix entries $a(b_k, b_j) = 0$ for $|j - k| > 1$, and A_h becomes a tridiagonal matrix. We could compute the matrix entries $a(b_k, b_j) = 0$ for $|j - k| \leq 1$ by hand using the explicit definitions of the hat functions given in Def. 7.23. Nevertheless, we use an indirect way, since this approach can be generalized easily to more dimensions.

Lemma 7.26. *Let $[x_j, x_{j+1}]$ be an intervall, and $\Psi : [-1,1] \to [x_j, x_{j+1}]$ be the affine mapping defined by*

$$\Psi(\xi) := \frac{x_{j+1} - x_j}{2}\xi + \frac{x_{j+1} + x_j}{2}.$$

Then,

$$\Psi^{-1}(x) = \frac{2x}{x_{j+1} - x_j} - \frac{x_{j+1} + x_j}{x_{j+1} - x_j}.$$

and

$$b_j|_{[x_{j-1},x_j]} = \hat{b}_2 \circ \Psi^{-1}, \qquad b_j|_{[x_j,x_{[j+1]}]} = \hat{b}_1 \circ \Psi^{-1}, \tag{7.40}$$

with

$$\hat{b}_1(\xi) := \frac{1-\xi}{2}, \qquad \hat{b}_2(\xi) := \frac{1+\xi}{2}. \tag{7.41}$$

Moreover, for $\ell = 0, \dots, N-1$ *there holds for the* 2×2 local element matrices

$$\begin{aligned}
M_\ell &:= \left(\int_{x_\ell}^{x_{\ell+1}} b_k(x) b_j(x) \, \mathrm{d}x \right)_{j,k=\ell,\ell+1} = \frac{h_\ell}{2} \hat{M}, \\
D_\ell &:= \left(\int_{x_\ell}^{x_{\ell+1}} b_k'(x) b_j(x) \, \mathrm{d}x \right)_{j,k=\ell,\ell+1} = \hat{D}, \\
S_\ell &:= \left(\int_{x_\ell}^{x_{\ell+1}} b_k'(x) b_j'(x) \, \mathrm{d}x \right)_{j,k=\ell,\ell+1} = \frac{2}{h_\ell} \hat{S},
\end{aligned}$$

with the 2×2 *reference matrices*

$$\hat{M} := \left(\int_{-1}^{1} \hat{b}_k(\xi) \hat{b}_j(\xi) \, \mathrm{d}\xi \right)_{j,k=1,2} = \frac{1}{3} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \tag{7.42a}$$

$$\hat{D} := \left(\int_{-1}^{1} \hat{b}_k'(\xi) \hat{b}_j(\xi) \, \mathrm{d}\xi \right)_{j,k=1,2} = \frac{1}{2} \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}, \tag{7.42b}$$

$$\hat{S} := \left(\int_{-1}^{1} \hat{b}_k'(\xi) \hat{b}_j'(\xi) \, \mathrm{d}\xi \right)_{j,k=1,2} = \frac{1}{2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{7.42c}$$

Proof. Straightforward calculation. For the local element matrices the substitution rule yields the result. □

The discretization matrix A_h can be derived from the preceding Lemma by summing over the intervals, see Alg. 7.27. Note, that for the first interval the reference basis function $\hat{b}_1$ leads to the basis function b_0, which is not part of the basis due to the Dirichlet boundary value.

Algorithm 7.27 (Assembling of finite element system matrix).

```
Input: mesh $a = x_0 < x_1 < \dots < x_N = b$, coefficients $\alpha, \beta, \gamma$
Initializing a sparse matrix $A_h \in \mathbb{R}^{N \times N}$ with zero entries
Compute local mesh-size $h_0 := x_1 - x_0$
Set the first entry
    $(A_h)_{1,1} := \left( -\alpha \frac{2}{h_0} \hat{S} + \beta \hat{D} + \gamma \frac{h_0}{2} \hat{M} \right)_{2,2}$
for $\ell = 1, \dots, N-1$ do
    Compute local mesh-size $h_\ell := x_{\ell+1} - x_\ell$
    Update the $2 \times 2$ submatrix $(A_h)_{j,k=\ell,\ell+1}$ by
        $(A_h)_{j,k=\ell+1,\ell+2} = (A_h)_{j,k=\ell+1,\ell+2} - \alpha \frac{2}{h_\ell} \hat{S} + \beta \hat{D} + \gamma \frac{h_\ell}{2} \hat{M}$
end for
Output: matrix $A_h$
```

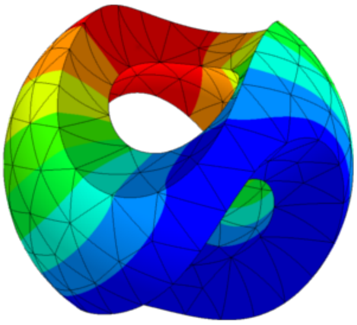
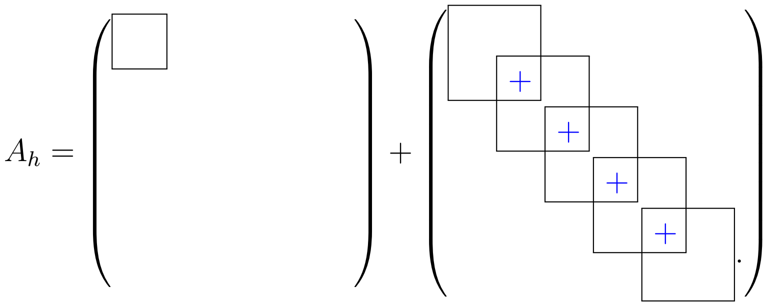


FIGURE 7.4.2. simulation generated by Netgen/NGSolve

We can sketch Alg. 7.27 by



The boxes are the local element matrices

$$-\alpha S_\ell + \beta D_\ell + \gamma M_\ell.$$

The right hand side $g_h = (l(b_1) - a(\tilde{y}, b_1), \ldots, l(b_N) - a(\tilde{y}, b_N))^\top \in \mathbb{R}^N$ can be assembled in the same way. Note, that $\tilde{y} = y_a b_0$. Hence, $a(\tilde{y}, b_j) = 0$ for $j = 2, \ldots, N$ and

$$a(\tilde{y}, b_1) = y_a \left(-\alpha \frac{2}{h_0} \hat{S} + \beta \hat{D} + \gamma \frac{h_0}{2} \hat{M} \right)_{2,1}.$$

For the integrals $\int_a^b f(x) b_k(x) \, \mathrm{d}x$ in the linear form $l(b_k)$, we sum over the local intervals and use a quadrature formula on each intervall.

REMARK 7.28. The finite element method can be easily extended to two and more dimensions with arbitrary shaped domains. The point to discuss is the construction of basis functions for the finite dimensional space V_h. Since we would like to have sparse discretization matrices, we use basis functions with local support. They can e.g. be constructed, if the domain of interest is meshed with triangles (2d) or tetrahedrons (3d). The nodal basis on such meshes can be defined by using functions, which are

- linear on each element (triangle, tetrahedron),
- 1 at one vertex and 0 on all the others,
- continuous on the whole domain.

Fig. 7.4.1 shows a mesh and a numerical simulation generated by the finite element package Netgen/NGSolve (see ngsolve.org). For this mesh, the surfaces of the tetrahedrons are curved in order to obtain good approximations e.g. of spherical boundary parts.

7.4.2. Introduction to the convergence analysis. We give here a brief introduction into the convergence analysis of finite element methods. We confine ourselves to the one dimensional Poisson problem with homogeneous Dirichlet boundary condition at the left boundary, i.e., we use $\alpha = -1$, $\beta = \gamma = 0$, and $y_a = 0$ in the definition of the bilinear and the linear form in (7.34). Moreover, $y_h \in V_h$ should be the solution to the discrete problem (7.35).

We put the problem into the following abstract framework:

- Let V be a Hilbert-space with scalar product $(\cdot,\cdot)_V$. In our case, $V := \{w \in H^1([a,b]) : w(a) = 0\}$ with scalar product $(v,w)_V := (v',w')_{L^2([a,b])}$.
- Let $V_h \subset V$ be a finite dimensional subset of V. In our case, we could use $V_h = \mathbf{span}\{b_1,\dots,b_N\}$ with the hat functions b_j.
- Let $a(\cdot,\cdot)$ be bilinear form, which is *continuous*, *symmetric*, and *coercive*, i.e., there exists constants $\alpha, C > 0$ such that
  - $|a(v,w)| \leq C \|v\|_V \|w\|_V$ for all $v,w \in V$,
  - $a(v,w) = a(w,v)$ for all $v,w \in V$, and
  - $|a(w,w)| \geq \alpha \|w\|_V^2$ for all $w \in V$.

  In our case, $a(v,w) := (v',w')_{L_2([a,b])}$.
- Let $l(\cdot)$ be a continuous linear form on V, i.e., there exists a constant $C_l > 0$ such that $|l(w)| \leq C_l \|w\|_V$ for all $w \in V$. In our case,

$$l(w) := \int_a^b f(x)w(x)\,\mathrm{d}x + y_b w(b).$$

Remark 7.29. Note, that $(v',w')_{L^2([a,b])}$ is a scalar product due to Friedrichs inequality, if Lem. 7.14 is generalized to functions in $H^1([a,b])$. Moreover, the definition of the space V is sloppy, since it makes no sense to define the function value of an L_2 function. More precisely, V can be defined as a completion of functions $w \in C^\infty([a,b])$ with $w(a) = 0$.

Moreover, the bilinear form $a(\cdot,\cdot)$ is continuous due to the Cauchy-Schwartz inequality. The linear form $l(\cdot)$ is continuous if $f \in L_2([a,b])$ due to the Cauchy-Schwarz inequality, and since $|w(b)| \leq \sqrt{b-a}\,\|w\|_V$ for $w \in V$ (see Lem. 7.14).

We start to collect one result from functional analysis.

Theorem 7.30 (Riesz Representation). *Any continuous linear functional l on a Hilbert space V can be represented uniquely as*

$$l(w) = (v_l, w)_V \tag{7.43}$$

with $v_l \in V$. Furthermore, $\|v_l\|_V = \|l\|_{V^*}$, were $\|\cdot\|_{V^*}$ denotes the norm of the dual space V^*.

Proof. The proof can be found in each textbook from functional analysis. Note, that the dual norm is defined by

$$\|l\|_{V^*} = \sup_{w \in V\setminus\{0\}} \frac{|l(w)|}{\|w\|_V}.$$

□

Corollary 7.31. *Let $a(\cdot,\cdot)$ be a continuous bilinear form. Then there exists a linear operator $A : V \to V$ such that*

$$\forall v,w \in V : \qquad a(v,w) = (Av,w)_V\,.$$

Moreover, there holds $\|A\|_{V\to V} = C$, whereas C is the stability constant of the bilinear form. If in addition the bilinear form is symmetric, than A is self-adjoint, i.e., $(Av,w)_V = (v,Aw)_V$ for all $v,w \in V$.

PROOF. For fixed $v \in V$, we can define a linear form $l_v : V \to \mathbb{R}$ by

$$\forall w \in V \quad l_v(w) := a(v,w).$$

Note, that $\|l_v\|_{V^*} = C\,\|v\|_V$. Using the Riesz representation Theorem there exists a $f_v \in V$ with $\|f_v\|_V = C\,\|v\|_V$ such that

$$\forall w \in V \quad a(v,w) = l_v(w) = (f_v,w)_V.$$

Finally, we define the bounded linear operator $A : V \to V$ by $v \mapsto Av := f_v$ and note that

$$\forall v,w \in V : \quad a(v,w) = l_v(w) = (f_v,w)_V = (Av,w)_V.$$

Moreover,

$$\|A\|_{V\to V} = \sup_{v\in V\setminus\{0\}} \frac{\|Av\|_V}{\|v\|_V} = \sup_{v\in V\setminus\{0\}} \frac{\|f_v\|_V}{\|v\|_V} = C.$$

If the bilinear form is symmetric, then

$$(Av,w)_V = a(v,w) = a(w,v) = (Aw,v)_V = (v,Aw)_V\,,$$

since the scalar product $(\cdot,\cdot)_V$ is symmetric. □

COROLLARY 7.32. *Let $a(\cdot,\cdot)$ be a continuous bilinear form, and $l(\cdot)$ be a continuous linear form. Then the variational problem*

$$\textit{find } y \in V : \qquad \forall w \in V \quad a(y,w) = l(w) \tag{7.44}$$

is equivalent to the operator equation

$$\textit{find } y \in V : \qquad Ay = f$$

with the operator A of the last corollary, and f beeing the Riesz representation of l, i.e., $l(w) = (f,w)_V$ for all $w \in V$.

PROOF. Using the definition of the operator A and the right hand side f we have

$$\forall w \in V : a(y,w) - l(w) = (Ay - f, w)_V.$$

Hence, if $Ay = f$ then $a(y,w) = l(w)$ for all $w \in V$. Vice versa, if $Ay \neq f$, then chosing $w = Ay - f$ leads to $a(y,w) \neq l(w)$ for at least this w. □

THEOREM 7.33 (Lax-Milgram). *Let $a(\cdot,\cdot)$ be symmtric, continuous, and coercive with coercivity constant α on the Hilbert space V. Moreover, let l be a continuous linear form on V. Then there exists a unique solution $y \in V$ to the variational problem* (7.44). *Moreover, this solution depends continuously on l:*

$$\|y\|_V \leq \frac{1}{\alpha}\,\|l\|_{V^*}\,. \tag{7.45}$$

PROOF. For $\tau \neq 0$ we define the operator $T : V \to V$ by

$$Tv := v - \tau\,(Av - f)\,,$$

where we use $A : V \to V$ and $f \in V$ as defined in the last corollary. Hence, (7.44) is equivalent to the fixed point problem $y = Ty$. The fixed point problem is uniquely solvable by the Banach fixed point theorem, if we can show, that for at least one τ the operator T is a contraction. To shorten notation, let $v := v_1 - v_2$. Then

$$\begin{aligned}\|Tv_1 - Tv_2\|_V^2 &= \|v - \tau Av\|_V^2 = (v - \tau Av, v - \tau Av)_V \\ &= (v, v)_V - \tau\,(Av, v)_V - \tau\,\underbrace{(v, Av)_V}_{=(Av,v)_V} + \tau^2\,(Av, Av)_V \\ &\le \|v\|_V^2 - 2\tau\alpha\,\|v\|_V^2 + \tau^2\,\|A\|_{V\to V}^2\,\|v\|_V^2 \\ &= \left(1 - \frac{\alpha^2}{\|A\|_{V\to V}^2}\right)\|v_1 - v_2\|_V^2\,,\end{aligned}$$

if $\tau = {}^{\alpha}\!/\!{}_{\|A\|_{V\to V}^2}$. Hence, T is a contraction, and there exists a unique solution $y \in V$ to (7.44). Finally, (7.45) is a consequence of

$$\alpha\,\|y\|_V^2 \le |a(y, y)| = |l(y)| \le \|l\|_{V^*}\,\|y\|_V\,.$$

□

Remark 7.34. The finite dimensional space $V_h \subset V$ is a Hilbert space with respect to the scalar product $(\cdot,\cdot)_V$. Moreover, all properties of the bilinear form $a(\cdot,\cdot)$ and the linear form $l(\cdot)$ hold on V_h as well. Hence, the discrete problem

$$\text{find } y_h \in V_h : \qquad \forall w_h \in V_h \quad a(y_h, w_h) = l(w_h) \tag{7.46}$$

is uniquely solvable for all finite dimensional subspaces $V_h \subset V$.

Theorem 7.35 (Cea). *Let in addition to the assumptions of the Theorem 7.33 $V_h \subset V$ be a finite dimensional subspace. Furthermore, let y be the unique solution to* (7.44)*, y_h be the unique solution to* (7.46)*, α be the coercivity constant of the bilinear form $a(\cdot,\cdot)$, and C be the stability constant of the bilinear form $a(\cdot,\cdot)$.*

Then,

$$\|y - y_h\|_V \le \frac{C}{\alpha} \inf_{w\in V_h} \|y - w_h\|_V\,. \tag{7.47}$$

Proof. The proof is based on the famous Galerkin orthogonality

$$a(y - y_h, w_h) = a(y, w_h) - a(y_h, w_h) = l(w_h) - l(w_h) = 0$$

for all $w_h \in V_h \subset V$. Hence, for all $w_h \in V_h$ there holds

$$\begin{aligned}\|y - y_h\|_V^2 &\le \frac{1}{\alpha}\,|a(y - y_h, y - y_h)| \\ &= \frac{1}{\alpha}\Big|a(y - y_h, y - w_h) + \underbrace{a(y - y_h, \underbrace{v_h - w_h}_{\in V_h})}_{=0}\Big| \\ &\le \frac{C}{\alpha}\,\|y - y_h\|_V\,\|y - w_h\|_V\,.\end{aligned}$$

Dividing by $\|y - y_h\|_V$ yields the claim, since $w_h \in V_h$ was arbitrary. □

Theorem 7.36. *Let $y \in V$ be the solution to* (7.33) *with $\alpha = -1$, $\beta = \gamma = 0$, $y_a = 0$, and $y_h \in V_h$ the solution the the corresponding problem* (7.35) *with the finite dimensional subspace V_h build by the hat functions of Def. 7.23.*

If in addition $y \in C^2([a, b])$, then there exists a constant $c > 0$ independent of y such that

$$\|y - y_h\|_V \le c\,\|y''\|_\infty \max\{h_0, \ldots, h_{N-1}\}\,. \tag{7.48}$$

Proof. The claim follows with the preceding results, if we can bound the best approximation error $\inf_{w\in V_h} \|y - w_h\|_V$. Let $s_h \in \mathbb{S}^1$ be the linear spline of Rem. 7.24 for the interpolation problem with $f = y$. Then, $s_h \in V_h$, and

$$\begin{aligned}
\inf_{w\in V_h} \|y - w_h\|_V^2 \le \|y - s_h\|_V^2 &= \sum_{j=0}^{N-1} \int_{x_j}^{x_{j+1}} \left|y'(x) - s_h'(x)\right|^2 \mathrm{d}x \\
&\le \sum_{j=0}^{N-1} \sup_{x\in[x_j,x_{j+1}]} \left|y'(x) - s_h'(x)\right|^2 h_j \\
&\underset{(B.4)}{\le} \sum_{j=0}^{N-1} h_j^3 \sup_{x\in[x_j,x_{j+1}]} \left|y''(x)\right|^2 \\
&\underset{\sum_{j=0}^{N-1} h_j=(b-a)}{\le} (b-a)\left(\|y''\|_\infty \max\{h_0,\dots,h_{N-1}\}\right)^2 .
\end{aligned}$$

Hence, (7.48) holds with $c := \frac{(b-a)^{5/2}}{2}$. □

Remark 7.37. (7.48) also holds, if y only belongs to $H^2([a,b])$. Furthermore, the scaling argument of Lem. 3.26 leads to

$$\|y - y_h\|_{L^2([a,b])} \le \tilde{c}\,\|y''\|_\infty \left(\max\{h_0,\dots,h_{N-1}\}\right)^2 ,$$

which is the same convergence order as for the finite difference method with pure Dirichlet data. Note, that for the finite difference method we have assumed $y \in C^4([a,b])$.

Corollary 7.38. *Under the assumptions of Theo. 7.36 let $V_h = \{s \in \mathbb{S}_0^p : s(a) = 0\}$ with the spline space*

$$\mathbb{S}_0^p := \left\{s \in C([a,b]) : \forall j = 0,\dots,N-1 : s|_{[x_{j-1},x_j]} \in \mathbb{P}_p\right\}$$

for $p \in \mathbb{N}$. If $y \in C^{p+1}([a,b])$, then

$$\|y - y_h\|_V \le c\,\|y''\|_\infty \left(\max\{h_0,\dots,h_{N-1}\}\right)^p . \tag{7.49}$$

Proof. This is a consequence of the interpolation error representation (B.4). □

Remark 7.39. The last theorems use the finite element space V_h and not the specific basis of V_h. In principle, the error estimates hold for any basis of V_h. But of course, the linear system of equations (7.36), which we have to solve, depends on the basis. Hence, the basis functions should be chosen such that this system is as easy as possible to solve. Moreover, the condition number of the discretization matrix should be small. Otherwise, rounding errors, which will appear during the solution of the linear system of equations, might lead to large errors.

Fig. 7.4.2 shows a suitable polynomial basis of $\mathbb{P}_5$ for the reference interval $[-1,1]$. It consists of the nodal basis functions $\hat{b}_1, \hat{b}_2 \in \mathbb{P}_1$ and so called interior basis functions $\hat{b}_3, \hat{b}_4, \dots$, which vanish on the interval boundaries. The latter are so called *integrated Legendre Polynomials*. Hence, $\hat{b}_3', \hat{b}_4', \dots$ are L_2-orthogonal to each other leading to well conditioned element matrices. The global basis functions can be constructed out of the local ones using the mapping Ψ defined in Lemma 7.26 in the following way: Let Ψ_j be the mapping $\Psi_j : [-1,1] \to [x_j, x_{j+1}]$. Then the basis of V_h can be defined by

(1) the nodal basis functions $b_1,\dots,b_N$ defined in Def. 7.23,
(2) and for each $j = 0,\dots,N-1$ the basis functions

$$B_{N+j(p-1)+k} := \hat{b}_{k+2} \circ \Psi_j^{-1}, \qquad k = 1,\dots,p-1.$$

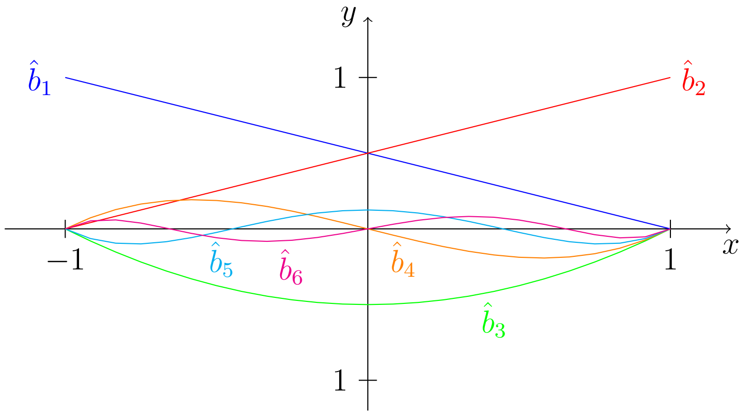


FIGURE 7.4.3. higher order basis functions on the reference element $[-1, 1]$

The corresponding matrix $A_h \in \mathbb{R}^{\text{ndof}\times\text{ndof}}$ with $\text{ndof} = Np$ degrees of freedom can be assembled similar to Alg. 7.27 but with local sub matrices of size $(p+1) \times (p+1)$.

REMARK 7.40. The finite element method can be easily generalized to problems in bounded Lipschitz domains $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$ of the form

$$\text{div}\ (A(x)\nabla y(x)) + b(x) \cdot \nabla y(x) + c(x)y(x) = 0, \qquad x \in \Omega,$$

for matrix functions $x \mapsto A(x) \in \mathbb{R}^{d\times d}$, vector functions $x \mapsto b(x) \in \mathbb{R}^d$, and scalar functions $x \mapsto c(x) \in \mathbb{R}$. Vector valued solution functions $y \mapsto y(x) \in \mathbb{R}^n$ would be possible as well. But the convergence theory presented in this section relies on symmetric and positive definite bilinear forms. Moreover, there are several other issues to be solved including a suitable interpretation of Dirichlet boundary values $y|_{\partial\Omega}$ for functions $y \in H^1(\Omega)$, a Friedrichs inequality, and estimates for the interpolation error on triangles or tetrahedrons.

APPENDIX A

Prerequisites

Theorem A.1 (Banach fixpoint theorem). *Let M be a complete metric space and $\mathrm{A} : M \to M$ be a contraction, i.e., there exists a contraction constant $0 < \kappa < 1$ such that*

$$\forall x, y \in M : \quad d(\mathrm{A}x, \mathrm{A}y) \le \kappa\, d(x, y). \tag{A.1}$$

Then, A has a unique fixpoint $w \in M$, i.e., $\mathrm{A}w = w$. Moreover, for any initial value $w_0 \in M$, the sequence of Banach iterates $w_{n+1} := \mathrm{A}w_n$ converges to w and it holds that

$$d(w, w_n) \le \frac{\kappa}{1-\kappa} d(w_{n+1}, w_n) \le \frac{\kappa^n}{1-\kappa} d(w_1, w_0) \quad \textit{for all } n \in \mathbb{N}_0. \tag{A.2}$$

Remark A.2. Often, the Banach fixpoint theorem is applied in the following setting: Let X be a Banach space and $M \subseteq X$ be a closed subspace. Suppose that $\mathrm{A} : M \to M$ is a contraction with respect to the natural metric, i.e., there exists a contraction constant $0 < \kappa < 1$ such that

$$\forall x, y \in M : \quad \|\mathrm{A}x - \mathrm{A}y\|_X \le \kappa\, \|x - y\|_X. \tag{A.3}$$

Then, the Banach fixpoint theorem applies.

Theorem A.3 (Implicit function theorem). *Let $f \in C^1(\mathbb{R}^m \times \mathbb{R}^n; \mathbb{R}^n)$. Let $(x, y) \in \mathbb{R}^m \times \mathbb{R}^n$ with $f(x, y) = 0$. Suppose that the Jacobi matrix $D_y f(x, y) \in \mathbb{R}^{n \times n}$ is invertible. Then, there exist open sets $U \subset \mathbb{R}^m$ and $V \subset \mathbb{R}^n$ with $(x, y) \in U \times V$ as well as a function $g \in C^1(U; V)$ such that*

$$\forall\, (\widetilde{x}, \widetilde{y}) \in U \times V : \quad \big[\, f(\widetilde{x}, \widetilde{y}) = 0 \quad \Longleftrightarrow \quad \widetilde{y} = g(\widetilde{x}) \,\big] \tag{A.4}$$

Stefan Banach (1892–1945) *was a Polish mathematician. Being the founder of modern functional analysis, he is amongst the most important mathematicians of the 20th century. His major work was the 1932 book* "Théorie des opérations linéaires"*, the first monograph on the general theory of functional analysis. Banach finished his PhD in mathematics at Lviv University in 1922 (the thesis also contained the Banach fixpoint theorem). In the same year, he completed the habilitation and became associate professor at Lviv University. In 1926, he was promoted to full professor. In August 1945, he died of lung cancer.*

APPENDIX B

Interpolation and numerical integration

We repeat in this appendix some results about interpolation and numerical integration.

B.1. Lagrange Interpolation

Let $t_1, \dots, t_s \in [a,b] \subset \mathbb{R}$ be pairwise different interpolation nodes and $g : [a,b] \to \mathbb{R}$ a sufficiently smooth function. Then we are looking for a polynomial $q \in \Pi_{s-1}$ of maximal degree $s-1$ such that

$$q(t_j) = g(t_j), \qquad j = 1, \dots, s. \tag{B.1}$$

For a given basis $\{\Psi_1, \dots, \Psi_s\}$ of Π_{s-1} and $q = \sum_{j=1}^s \alpha_j \Psi_j$ (B.1) is a system of s linear equations for the s unknown coefficients $\alpha_1, \dots, \alpha_s$. The unique solution to (B.1) is given by

$$q = \sum_{j=1}^{s} g(t_j) L_j, \tag{B.2}$$

where $L_j \in \Pi_{s-1}$ are the Lagrange basis polynomials defined by

$$L_j(t) = \prod_{\substack{k=1\\k\neq j}}^{s} \frac{t - t_k}{t_j - t_k}. \tag{B.3}$$

They are defined such that $L_j(t_i) = 1$ for $i = j$ and $L_j(t_i) = 0$ for $i \neq j$.

As shown in Lemma 3.18 for all $k = 0, \dots, s-1$ and all $t \in [a,b]$, there holds the error identity

$$g^{(k)}(t) - q^{(k)}(t) = \frac{g^{(s)}(\xi)}{(s-k)!} \prod_{\ell=1}^{s-k} (t - \zeta_\ell) \tag{B.4}$$

with appropriate scalars $\xi = \xi(k,t), \zeta_\ell = \zeta_\ell(k) \in [a,b]$. In particular, it follows that

$$\|g^{(k)} - q^{(k)}\|_{\infty,[a,b]} \leq \frac{\|g^{(s)}\|_{\infty,[a,b]}}{(s-k)!} |b-a|^{s-k}. \tag{B.5}$$

B.2. Interpolation quadrature

Suppose we want to compute the integral

$$Q(g) := \int_a^b g(\tau) d\tau \tag{B.6}$$

for an at least continuous function $g : [a,b] \to \mathbb{R}$. For an interpolation quadrature we approximate the integrand g by an interpolation polynomial and integrate this polynomial. Using the explicit form (B.2) of the interpolation polynomial, we arrive at

$$Q^s(g) := Q(q) \qquad \text{with } q := \sum_{j=1}^{s} g(t_j) L_j. \tag{B.7}$$

In fact, if the interpolation or quadrature nodes are chosen, the quadrature formula and in particular the quadrature weights $\alpha_1, \ldots, \alpha_s$ of an interpolation quadrature are given by

$$Q^s(g) = \sum_{j=1}^{s} \alpha_j g(t_j) \qquad \text{with } \alpha_j := \int_a^b L_j(\tau) d\tau, \quad j = 1, \ldots, s. \tag{B.8}$$

As shown in Lemma 3.19 using (B.4) with $k = 0$ leads for sufficiently smooth functions g to the error bound

$$|Q(g) - Q^s(g)| \leq \frac{\|g^{(s)}\|_{\infty,[a,b]}}{s!} |b-a|^{s+1}. \tag{B.9}$$

REMARK B.1. Interpolation quadratures of this kind are always exact for polynomials $p \in \Pi_{s-1}$, since $q = p$ if q is the solution to (B.1) with $g = p$. Moreover, each quadrature Q^s with quadrature nodes $t_1, \ldots, t_s$ and quadrature weights $\alpha_1, \ldots, \alpha_s$ is an interpolation quadrature if it is exact for polynomials of maximal degree $s-1$, since the corresponding Lagrange polynomials L_j have maximal degree $s-1$ and therefore

$$\int_a^b L_k(\tau)\,\mathrm{d}\tau = Q(L_k) = Q^s(L_k) = \sum_{j=1}^{s} \alpha_j L_k(t_j) = \alpha_k, \qquad k = 1, \ldots, s.$$

LEMMA B.2. *Let Q^s be a quadrature with s quadrature points $t_1, \ldots, t_s$. Then*

$$\exists\, p \in \Pi_{2s}: \quad Q(p) \neq Q^s(p). \tag{B.10}$$

In other words, the maximum exactness of quadrature formulas is $2s-1$.

PROOF. We use $q(t) := \prod_{k=1}^{s}(t-t_k)^2$ with $q \in \Pi_{2s}$. Since q is non-negative and not the zero function, it holds $Q(q) > 0$. The claim follows with

$$Q^s(q) = \sum_{j=1}^{s} \alpha_j \left(\prod_{k=1}^{s} (t_j - t_k)^2 \right) = 0.$$

□

Hence, interpolation quadratures will be exact for polynomials $p \in \Pi_m$ with $s-1 \leq m \leq 2s-1$. m depends on the quadrature nodes. Note, still we have to prove that $m = 2s-1$ is possible.

LEMMA B.3. *Let the interpolation quadrature Q^s with s quadrature points be exact for polynomials $p \in \Pi_m$. Then for $g \in C^{m+1}([a,b])$ there holds*

$$|Q(g) - Q^s(g)| \leq \frac{\|g^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!} |b-a|^{m+2}. \tag{B.11}$$

PROOF. We add to the s interpolation nodes $t_1, \ldots, t_s$ artificial nodes $t_{s+1}, \ldots, t_{m+1}$ with $t_j \neq t_k$ for $j, k = 1, \ldots, m+1$. Let $q_m \in \Pi_m$ be the solution to

$$q_m(t_j) = g(t_j), \qquad j = 1, \ldots, m+1.$$

Since $Q(q_m) = Q^s(q_m)$ by linearity of Q and Q^s we obtain

$$Q(g) - Q^s(g) = Q(g - q_m) - Q^s(g - q_m).$$

Using

$$Q^s(g - q_m) = \sum_{j=1}^{s} \alpha_j \left(g(t_j) - q_m(t_j) \right) = 0$$

we arrive at

$$|Q(g) - Q^s(g)| = \left| \int_a^b (g(\tau) - q_m(\tau))\, \mathrm{d}\tau \right| \leq |b-a|\, \|g - q_m\|_{\infty,[a,b]}$$

and the assertion follows with the interpolation error (B.5) for the $m+1$ interpolation points $t_1, \dots, t_{m+1}$. □

Example B.4 (Newton-Cotes quadrature). Using equidistant quadrature nodes in the interval $[a,b]$ leads to so called closed and open Newton-Cotes formulas. For the closed formulas the boundaries a and b are quadrature nodes, while for the open not.

The first closed Newton-Cotes formulas are

(1) $s = 1 : Q^1(f) = f(a)$ (left rectangular rule) and $Q^1(f) = (b-a)f(b)$ (right rectangular rule)
(2) $s = 2 : Q^2(f) = \frac{b-a}{2}\left(f(a) + f(b)\right)$ (trapezoidal rule)
(3) $s = 3 : Q^3(f) = \frac{b-a}{6}\big(f(a) + 4f(\frac{a+b}{2}) + f(b)\big)$ (Simpson rule)
(4) $s = 4 : Q^4(f) = \frac{b-a}{8}\big(f(a) + 3f(a + \frac{b-a}{3}) + 3f(a + 2\frac{b-a}{3}) + f(b)\big)$
(5) $s = 5 : Q^5(f) = \frac{b-a}{90}\big(7f(a) + 32f(a + \frac{b-a}{4}) + 12f(a + 2\frac{b-a}{4}) + 32f(a + 3\frac{b-a}{4}) + 7f(b)\big)$.

The first open Newton-Cotes formulas are

(1) $s = 1 : Q^1(f) = (b-a)f(\frac{a+b}{2})$ (midpoint rule)
(2) $s = 2 : Q^2(f) = \frac{b-a}{2}\big(f(a + \frac{b-a}{3}) + f(a + 2\frac{b-a}{3})\big)$
(3) $s = 3 : Q^3(f) = \frac{b-a}{3}\big(2f(a + \frac{b-a}{4}) - f(a + 2\frac{b-a}{4}) + 2f(a + 3\frac{b-a}{4})\big)$.

Exercise B.5. Show, that the Simpson rule is exact for polynomials of degree 3 and that the midpoint rule is exact for polynomials of degree 1.

B.3. Composite quadrature formulas

Newton-Cotes formulas of high order (thereby high accuracy) need high regularity of the integrand. Moreover, as it can be already seen for the open Newton-Cotes formula Q^3, some quadrature weights become negative. This is a severe stability issue due to the loss of significance.

A comparatively simple way to overcome these problems are composite quadrature formulas. For these the interval $[a,b]$ is split into several small intervals, where a low order quadrature formula can be applied. Let $t_0 = a < t_1 < \cdots < t_{n-1} < t_n = b$ be such a mesh. Then

$$\int_a^b g(\tau)\, \mathrm{d}\tau = \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} g(\tau)\, \mathrm{d}\tau \approx \sum_{j=0}^{n-1} Q^s_{[t_j,t_{j+1}]}(g). \tag{B.12}$$

Using the maximal mesh size $h_\Delta := \max_{j=0}^{n-1}(t_{j+1} - t_j)$ and $\sum_{j=0}^{j-1}(t_{j+1} - t_j) = b - a$ (B.11) leads to the error of composite quadrature formulas

$$\left| Q(g) - \sum_{j=0}^{n-1} Q^s_{[t_j,t_{j+1}]}(g) \right| \leq \frac{b-a}{(m+1)!} \|g^{(m+1)}\|_{\infty,[a,b]} h_\Delta^{m+1}, \tag{B.13}$$

if $g \in C^{m+1}([a,b])$ and if $Q^s_{[t_j,t_{j+1}]}$ is for all $j = 0, \dots, n-1$ exact for polynomials of maximal degree m.

Example B.6 (Composite trapezoidal rule). The most popular example is the composite trapezoidal rule, which uses the closed Newton-Cotes formula with $s = 2$ and an equidistant mesh on $[a,b]$ with mesh size $h := \frac{b-a}{n}$:

$$Q_h^2(g) = \frac{h}{2}\left(f(a) + f(b) + 2\sum_{j=1}^{n-1} f(a+jh)\right)$$

Using an improved analysis, there holds the following error estimate:

$$\left|Q(g) - Q_h^2(g)\right| \leq \frac{b-a}{12}\|g''\|_{\infty,[a,b]}h^2.$$

EXAMPLE B.7 (Composite midpoint rule). Using the open Newton-Cotes formula with $s = 1$ with an equidistant mesh on $[a,b]$ with mesh size $h := \frac{b-a}{n}$ leads to

$$Q_h^1(g) = h\sum_{j=0}^{n-1} g\left(a + jh + \frac{h}{2}\right)$$

with the error

$$\left|Q(g) - Q_h^1(g)\right| \leq \frac{b-a}{24}\|g''\|_{\infty,[a,b]}h^2.$$

Note, that the midpoint rule is exact for polynomials of degree one.

B.4. Weighted quadrature formulas

We will generalize the idea of interpolation quadrature using admissible weight functions. Let $\omega : (a,b) \to (0,\infty)$ be a weight function such that $\int_a^b \omega(\tau)\,\mathrm{d}\tau$ exists. We try to find a numerical method to compute

$$Q_\omega(g) := \int_a^b \omega(\tau)g(\tau)\,\mathrm{d}\tau$$

for $g \in C([a,b])$. Using polynomial interpolation with nodes $t_1,\dots,t_s$ with associated Lagrange polynomials $L_1,\dots,L_s \in \Pi_{s-1}$ to approximate the function g leads to the weighted interpolation quadrature

$$Q_\omega^s(g) := \sum_{j=1}^{s} \alpha_j g(t_j) \quad \text{with } \alpha_j := \int_a^b \omega(\tau)L_j(\tau)d\tau, \quad j = 1,\dots,s, \tag{B.14}$$

where the weight function ω is included in the quadrature weights. In principle, the only necessary condition on ω would be, that the integrals for the quadrature weights α_j exist. Nevertheless, for the following we need to assume that ω is positive in the open interval (a,b).

REMARK B.8. The results of Section B.2 can be generalized easily. In particular, Remark B.1, Lemma B.2 and Lemma B.3 still hold. Only B.11 has to be replaced by

$$\left|Q_\omega(g) - Q_\omega^s(g)\right| \leq \frac{\|g^{(m+1)}\|_{\infty,[a,b]}}{(m+1)!}Q_\omega(1)\left|b-a\right|^{m+1}. \tag{B.15}$$

LEMMA B.9. *If the interpolation quadrature Q_ω^s is exact for $p \in \Pi_{2s-2}$, than the quadrature weights are positive.*

PROOF. The claim is a consequence of the property $L_j(t_k) = 0$ for $j \neq k$, $L_j(t_j) = 1$, and $L_j^2 \in \Pi_{2s-2}$:

$$0 < \int_a^b \omega(\tau)\left(L_k(\tau)\right)^2\,\mathrm{d}\tau = Q_\omega^s(L_k^2) = \sum_{j=1}^{s} \alpha_j \left(L_k(t_j)\right)^2 = \alpha_k, \quad k = 1,\dots,s.$$

□

Theorem B.10. *Let (Q^s_ω) be a sequence of interpolation quadratures with $s \in \mathbb{N}$ interpolation points and positive quadrature weights. Then*

$$\forall g \in C([a,b]): \quad \lim_{s\to\infty} Q^s_\omega(g) = Q_\omega(g).$$

Proof. Since for all $s \in \mathbb{N}$ the interpolation quadrature Q^s_ω is exact for constant functions, there holds

$$\forall s \in \mathbb{N}: \quad \sum_{j=1}^{s} \left|\alpha_j^{(s)}\right| = \sum_{j=1}^{s} \alpha_j^{(s)} = Q_\omega(1).$$

Hence, there holds for all $g \in C([a,b])$

$$|Q^s_\omega(g)| = \left|\sum_{j=1}^{s} \alpha_j^{(s)} g(t_j^{(s)})\right| \leq \|g\|_{\infty,[a,b]} \sum_{j=1}^{s} \left|\alpha_j^{(s)}\right| = Q_\omega(1)\|g\|_{\infty,[a,b]}.$$

Note, that the bound is independent of s. Moreover,

$$|Q_\omega(g)| = \left|\int_a^b \omega(\tau) g(\tau)\,\mathrm{d}\tau\right| \leq Q_\omega(1)\|g\|_{\infty,[a,b]}.$$

Finally, using Stone-Weierstrass theorem for all $\epsilon > 0$ there exists a polynomial p with degree $s_p \in \mathbb{N}$ such that $\|g-p\|_{\infty,[a,b]} < \frac{\epsilon}{2Q_\omega(1)}$. Since due to Remark B.1 $Q^s_\omega(p) = Q_\omega(p)$ for all $s > s_p$ the claim follows with

$$\forall s > s_p: |Q^s_\omega(g) - Q_\omega(g)| \leq |Q^s_\omega(g-p)| + |Q_\omega(p-g)| \leq 2Q_\omega(1)\|g-p\|_{\infty,[a,b]} < \epsilon.$$

□

B.5. Gaussian quadrature

In this section we will show, that there exists quadrature nodes such that the interpolation quadrature has maximal exactness $2s-1$.

Definition B.11 (Gauss quadrature). We call a quadrature formula Q^s_ω a Gaussian quadrature if $Q_\omega(p) = Q^s_\omega(p)$ for all $p \in \Pi_{2s-1}$.

In other words a quadrature formula with maximal exactness for polynomials is called a Gaussian quadrature. Of course, the quadrature will depend on the weight function ω. Moreover, we still have to prove, that such a quadrature exists. Nevertheless, convergence of such quadratures follows by Theorem B.10 and Lemma B.9. Note, that due to Remark B.1 Gauss quadratures are always interpolation quadratures.

Lemma B.12. *Let $w : (a,b) \to (0,\infty)$ be an admissible weight function, $t_1, \ldots, t_s \in [a,b]$ pairwise different quadrature nodes of the interpolation quadrature Q^s_ω, and*

$$q_s(t) := \prod_{j=1}^{s} (t - t_j). \tag{B.16}$$

Then Q^s_ω is a Gaussian quadrature if and only if

$$\forall p \in \Pi_{s-1}: \quad \int_a^b \omega(\tau) q_s(\tau) p(\tau)\,\mathrm{d}\tau = 0. \tag{B.17}$$

In other words, $q_s \in \Pi_s$ is orthogonal to all polynomials of degree smaller than s with respect to the weighted L^2 scalar product

$$(f,g)_\omega := \int_a^b \omega(\tau) f(\tau) g(\tau)\,\mathrm{d}\tau. \tag{B.18}$$

PROOF. Since $q_s p \in \Pi_{2s-1}$ for $p \in \Pi_{s-1}$, (B.17) is necessary due to

$$Q_\omega(q_s p) = Q^s_\omega(q_s p) = \sum_{k=1}^{s} \alpha_k \underbrace{q_s(t_k)}_{=0} p(t_k) = 0. \tag{B.19}$$

It is also sufficient, since for all polynomials $p \in \Pi_{2s-1}$ there exists a solution $\tilde{p} \in \Pi_{s-1}$ to the interpolation problem (B.1) for $g = p$. Hence, due to the fundamental theorem of algebra there exists a $q \in \Pi_{s-1}$ such that

$$p - \tilde{p} = q_s q.$$

since $p - \tilde{p}$ vanishes for all $t_1, \ldots, t_s$. Finally, (B.17) together with $Q^s_\omega(q_s q) = 0$ leads to

$$Q_\omega(p) = Q_\omega(\tilde{p}) + Q_\omega(q_s q) = Q_\omega(\tilde{p}) = Q^s_\omega(\tilde{p}) = Q^s_\omega(\tilde{p}) + Q^s_\omega(q_s q) = Q^s_\omega(p),$$

i.e. Q^s_ω is exact for all polynomials of maximal degree $2s - 1$. □

REMARK B.13 (orthogonal polynomials). There exists a sequence $(q_s)_{s\in\mathbb{N}_0}$ of hierarchical, $(\bullet, \bullet)_\omega$-orthogonal polynomials $q_s \in \Pi_s$ such that $(q_s, q_k)_\omega = 0$ for all $k = 0, \ldots, s-1$. It can be constructed with Gram-Schmidt orthogonalization with respect to the scalar product $(\bullet, \bullet)_\omega$ applied to the monomial basis $t \mapsto t^k$ for $k = 0, \ldots, s$ of Π_s.

The sequence of hierarchical, $(\bullet, \bullet)_\omega$-orthogonal polynomials $q_s \in \Pi_s$ is not unique. Moreover, in the preceding lemma q_s has exactly s pairwise different roots in the interval $[a, b]$.

LEMMA B.14. *Let $q_s \in \Pi_s$ be such that $(q_s, q)_\omega = 0$ for all $q \in \Pi_{s-1}$. Then q_s has exactly s simple, pairwise different roots in the interval (a, b).*

PROOF. We prove by contradiction and assume, that there exist $m < s$ roots $t_1, \ldots, t_m$ with odd order in the interval (a, b). $q \in \Pi_m \subset \Pi_{s-1}$ is defined by $q(t) := 0$ for $m = 0$ and $q(t) := \prod_{j=1}^{m}(t - t_j)$ for $m \geq 1$. By assumption $(q_s, q)_\omega = 0$ which is a contradiction, since the function $\omega q_s q$ is not identically zero and does not change sign. □

THEOREM B.15. *For all admissible weight functions ω there exists a unique Gauss quadrature Q^s_ω with $s \in \mathbb{N}$ quadrature nodes, which are the roots of the corresponding orthogonal polynomial $q_s \in \Pi_s$. Moreover, the quadrature weights are positive.*

PROOF. Existence was shown in the preceding lemmas. The quadrature weights are positive due to Lemma B.9. To show uniqueness we assume, that there exists a second Gauss quadrature

$$\tilde{Q}^s_\omega(g) = \sum_{j=1}^{s} \tilde{\alpha}_j g(\tilde{t}_j)$$

with the corresponding Lagrange basis $\tilde{L}_1, \ldots, \tilde{L}_s \in \Pi_{s-1}$. Obviously, $\tilde{L}_k q_s \in \Pi_{2s-1}$. Hence for all $k = 1, \ldots, s$

$$0 = (\tilde{L}_k, q_s)_\omega = Q_\omega(\tilde{L}_k q_s) = \tilde{Q}^s_\omega(\tilde{L}_k q_s) = \sum_{j=1}^{s} \tilde{\alpha}_j \tilde{L}_k(\tilde{t}_j) q_s(\tilde{t}_j) = \tilde{\alpha}_k q_s(\tilde{t}_k).$$

Since $\tilde{\alpha}_k > 0$ by Lemma B.9, $\tilde{t}_k$ is a root of q_s for all $k = 1, \ldots, s$. So the quadrature nodes and therewith the quadrature weights are identical. □

The error of Gauss quadratures can be estimated by (B.15) if the integrand is sufficiently smooth. But even for $g \in C([a, b])$ convergence is guaranteed by Theorem B.10.

Example B.16 (Gauss-Legendre). For $[a, b] = [-1, 1]$ and $\omega(t) \equiv 1$ the hierarchical, L^2-orthogonal polynomials $q_s \in \Pi_s$ of Remark B.13 are called Legendre polynomials. They can be defined by

$$L_s(t) := \frac{1}{2^s s!} \frac{\partial^s}{\partial t^s} \left(t^2 - 1\right)^s, \qquad s \in \mathbb{N}_0.$$

The first Gauss-Legendre quadratures on the intervall $[0, 1]$ are

$$\begin{aligned} Q^1(g) &= g\left(\frac{1}{2}\right) \qquad \text{(midpoint rule)}, \\ Q^2(g) &= \frac{1}{2} g\left(\frac{1}{2} - \frac{\sqrt{3}}{6}\right) + \frac{1}{2} g\left(\frac{1}{2} + \frac{\sqrt{3}}{6}\right), \\ Q^3(g) &= \frac{5}{18} g\left(\frac{1}{2} - \frac{\sqrt{15}}{10}\right) + \frac{4}{9} g\left(\frac{1}{2}\right) + \frac{5}{18} g\left(\frac{1}{2} + \frac{\sqrt{15}}{10}\right). \end{aligned}$$

Example B.17 (Gauss-Chebyshev). For $[a, b] = [-1, 1]$ and $\omega(t) = \frac{1}{\sqrt{1-t^2}}$ the hierarchical, $(\bullet, \bullet)_\omega$-orthogonal polynomials $q_s \in \Pi_s$ of Remark B.13 are called Chebyshev polynomials. There exists a closed form of all Gauss-Chebyshev quadratures

$$Q^s(g) = \frac{\pi}{s} \sum_{j=1}^{s} g\left(\cos\left(\frac{(2j-1)\pi}{2s}\right)\right) \approx \int_{-1}^{1} \frac{g(\tau)}{\sqrt{1-\tau^2}} \, \mathrm{d}\tau.$$

B.6. Radau and Lobatto quadrature

As we have seen, for Gauss formulas the quadrature nodes are in the interior of the interval (a, b). In some applications it can be useful, if one or both of the boundaries are quadrature nodes.

Definition B.18 (Radau quadrature). We call a quadrature formula Q^s_ω with pairwise different quadrature nodes $t_1, \dots, t_{s-1} \in [a, b)$ and $t_s = b$ a Radau quadrature if $Q_\omega(p) = Q^s_\omega(p)$ for all $p \in \Pi_{2s-2}$.

In contrast to Gauss quadratures we reduce the exactness for polynomials by one degree. Note, that Theorem B.10 and Lemma B.9 still apply. The quadrature weights are positive and convergence is guaranteed. For such formulas we need a variant of Lemma B.12.

Lemma B.19. *Let $w : (a, b) \to (0, \infty)$ be an admissible weight function and Q^s_ω an interpolation quadrature with $t_s = b$ and arbitrary quadrature nodes $t_1, \dots, t_{s-1}$. We define*

$$q_{s-1}(t) := \prod_{j=1}^{s-1} (t - t_j) \qquad \textit{and } r(t) := (b - t). \tag{B.20}$$

Then Q^s_ω is a Radau quadrature if and only if

$$\forall p \in \Pi_{s-2} : \quad \int_a^b \omega(\tau) r(\tau) q_{s-1}(\tau) p(\tau) \, \mathrm{d}\tau = 0. \tag{B.21}$$

In other words, $q_{s-1} \in \Pi_{s-1}$ is orthogonal to all polynomials of degree smaller than $s - 1$ with respect to the weighted L^2 scalar product with weight function $\tilde{\omega} := r\omega$.

Proof. For $p \in \Pi_{2s-2}$ there exists a solution $\tilde{p} \in \Pi_{s-1}$ to the interpolation problem (B.1) for $g = p$. Hence, due to the fundamental theorem of algebra there exists a $q \in \Pi_{s-2}$ such that

$$p - \tilde{p} = r q_{s-1} q.$$

since $p - \tilde{p}$ vanishes for all $t_1, \dots, t_s$. Finally, (B.21) together with

$$Q^s_\omega(r q_{s-1} q) = \sum_{j=1}^{s} \alpha_j (b - t_j) \prod_{k=1}^{s-1} (t_j - t_k) q(t_j) = 0 \tag{B.22}$$

leads to

$$Q_\omega(p) = Q_\omega(\tilde{p}) + Q_\omega(r q_{s-1} q) = Q_\omega(\tilde{p}) = Q^s_\omega(\tilde{p}) = Q^s_\omega(\tilde{p}) + Q^s_\omega(r q_{s-1} q) = Q^s_\omega(p),$$

i.e. Q^s_ω is exact for all polynomials of maximal degree $2s - 2$.

Vice versa, for $p \in \Pi_{s-2}$ there holds $p r q_{s-1} \in \Pi_{2s-2}$. Hence, (B.21) is a consequence of (B.22) and $Q_\omega(p r q_{s-1}) = Q^s_\omega(p r q_{s-1})$. □

Note, that $\tilde{\omega}$ is admissible, since r is positive on (a, b). Using the results of the preceding section, Radau formulas exist, are unique, have positive quadrature weights, and converge for all $g \in C^1([a, b])$. The quadrature nodes of a Radau quadrature Q^s_ω with s nodes are b and the $s - 1$ roots of the $(\bullet, \bullet)_{r\omega}$ orthogonal polynomials $p_{s-1} \in \Pi_{s-1}$.

Remark B.20. Radau quadratures can also be defined using the left boundary a instead of b as quadrature node. The only difference is, that r has to be replaced by $r(t) := t - a$.

Definition B.21 (Lobatto quadrature). We call an interpolation quadrature formula Q^s_ω with pairwise different quadrature nodes $t_2, \dots, t_{s-1} \in (a, b)$, $t_1 = a$, and $t_s = b$ a Lobatto quadrature if $Q_\omega(p) = Q^s_\omega(p)$ for all $p \in \Pi_{2s-3}$.

The polynomial r has to be replaced by $r(t) := (t - a)(b - t)$. In this case, Lemma B.9 and therewith Theorem B.10 cannot be used. But the rest of the theory still holds.

Example B.22 (Radau quadratures). The first Radau quadratures on the interval $[0, 1]$ with $\omega(t) \equiv 1$ are

$$\begin{aligned}
Q^1(g) &= g(1) \qquad \text{(right rectangular rule)},\\
Q^2(g) &= \frac{3}{4} g\left(\frac{1}{3}\right) + \frac{1}{4} g(1),\\
Q^3(g) &= \frac{16 - \sqrt{6}}{36} g\left(\frac{4 - \sqrt{6}}{10}\right) + \frac{16 + \sqrt{6}}{36} g\left(\frac{4 + \sqrt{6}}{10}\right) + \frac{1}{9} g(1).
\end{aligned}$$

Example B.23 (Lobatto quadratures). The first Lobatto quadratures on the interval $[0, 1]$ with $\omega(t) \equiv 1$ are

$$\begin{aligned}
Q^2(g) &= \frac{1}{2} g(0) + \frac{1}{2} g(1) \qquad \text{(trapezoidal rule)},\\
Q^3(g) &= \frac{1}{6} g(0) + \frac{2}{3} g\left(\frac{1}{2}\right) + \frac{1}{6} g(1),\\
Q^4(g) &= \frac{1}{12} g(0) + \frac{5}{12} g\left(\frac{1}{2} - \frac{1}{2\sqrt{5}}\right) + + \frac{5}{12} g\left(\frac{1}{2} + \frac{1}{2\sqrt{5}}\right) \frac{1}{12} g(1).
\end{aligned}$$

Bibliography

[But08] John C. Butcher: Numerical methods for ordinary differential equations, Wiley, Chichester, second edition, 2008.

[HNW93] Ernst Hairer, Syvert Paul Nørsett, Gerhard Wanner: Solving ordinary differential equations. I, Springer, Berlin, second edition, 1993.

[SWP12] Karl Strehmel, Rüdiger Weiner, Helmut Podhaisky: Numerik gewöhnlicher Differentialgleichungen, Springer, Berlin, second edition, 2012 [in German].

[Wal00] Wolfgang Walter: Gewöhnliche Differentialgleichungen, Springer, Berlin, second edition, 2000 [in German].

[DW11] Peter Deuflhard, Martin Weiser : Adaptive Lösung partieller Differentialgleichungen, De Gruyter , eISBN: 9783110218039 , 2011, [in German].