Georgios D. Akrivis
Computer Science Department, University of Ioannina, Greece,
e-mail: akrivis@cs.uoi.gr
and
BCAM – Basque Center for Applied Mathematics,
e-mail: gakrivis@bcamath.org

NUMERICAL METHODS FOR INITIAL VALUE PROBLEMS

Lecture Notes
from a course taught at
BCAM – Basque Center for Applied Mathematics,
Bilbao, Basque Country, Spain,
in November 2012

BILBAO, 2012

Contents

1. Initial value problems

We will discuss numerical methods for initial value problems for ordinary differential equations. Numerical methods have advantages and drawbacks. One of the most desirable properties of numerical schemes is that the numerical approximations "mimic" the behaviour of the exact solutions; whether this property is fulfilled or not depends on both the numerical scheme and the underlined equation. Since the behaviour of the exact solutions depends on properties of the underlined equation, several classes of equations are distinguished. In these notes we will mainly focus on two classes of equations: the first is the simplest class, for which all practical numerical methods for initial value problems should behave properly, while the second is motivated by the behaviour of evolution partial differential equations (p.d.e's). There are important classes of o.d.e's, such as Hamiltonian systems, stochastic o.d.e's etc., which we will *not* discuss here.

In the first chapter we will present some basic facts from the theory of o.d.e's such that also those who are not familiar with it will be able to understand our analysis of numerical methods. The second chapter is devoted to low order schemes with particular emphasis on the Euler method; we study it in detail and discuss some important properties of numerical methods. In the third chapter we study Runge–Kutta methods and in the fourth multistep schemes. Stable high order methods of these classes are widely used in practice.

We first mention some basic results from the theory of initial value problems (i.v.p.) for ordinary differential equations (o.d.e's).

Let $a, b \in \mathbb{R}, a < b$, $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ be a function, and $y_0 \in \mathbb{R}$. The typical *initial value problem* we will consider here is the following: Seek a

function $y : [a, b] \to \mathbb{R}$, such that

$$\begin{cases} y'(t) = f\big(t, y(t)\big), & a \le t \le b, \\ y(a) = y_0. \end{cases} \tag{1.1}$$

We will always assume that f is continuous for $(t, y) \in [a, b] \times \mathbb{R}$. (We will write $f \in C([a, b] \times \mathbb{R})$.) Every function $y \in C^1[a, b]$, satisfying both the differential equation in (1.1) and the initial condition $y(a) = y_0$, is called *solution* of the initial value problem (1.1). Obviously, if f is m times continuously differentiable in $[a, b] \times \mathbb{R}$, then the solution y is $m + 1$ times continuously differentiable in $[a, b]$.

In the theory of o.d.e's the initial value problem (1.1) is studied; in particular, conditions of f that ensure existence and/or uniqueness are given, and the continuous dependence of the solution on the initial data is investigated. Continuous dependence means here the following: If y is the solution of (1.1), and $\tilde{y}$ the solution of the corresponding problem with initial value $\tilde{y}_0$ instead of y_0, then, when the difference $|y_0 - \tilde{y}_0|$ is small, we want the corresponding difference of the solutions $\|y - \tilde{y}\|$ to be also small. (The choice of the norm of functions defined in $[a, b]$, in which we measure the difference, is, in general, part of the problem.)

This chapter consists of three short sections. The first concerns existence and uniqueness of solutions, and the second stability of solutions. In the third we generalize the results to systems of o.d.e's.

1.1 Existence and uniqueness

In the special case that f is a polynomial of degree at most one in y, the corresponding o.d.e. is called *linear* and problem (1.1) takes the form

$$\begin{cases} y'(t) = p(t)y(t) + q(t), & a \le t \le b, \\ y(a) = y_0. \end{cases} \tag{1.2}$$

If $p, q \in C[a, b]$, then problem (1.2) possesses a unique solution y, given by

$$y(t) = \mathrm{e}^{\int_a^t p(s)\,ds}\Big[y_0 + \int_a^t q(s)\mathrm{e}^{-\int_a^s p(\tau)\,d\tau}\,ds\Big], \quad a \le t \le b, \tag{1.3}$$

i.e.,

$$y(t) = \mathrm{e}^{\int_a^t p(s)\,ds} y_0 + \int_a^t q(s)\mathrm{e}^{\int_s^t p(\tau)\,d\tau}\,ds, \quad a \le t \le b.$$

We can derive (1.3) as follows: In the case $p = 0$, the equation can be solved by integration. Multiplying by an appropriate *integrating factor*, we can reduce the general case to the previous one. Indeed, we can write the differential equation in the form $y'(s) - p(s)y(s) = q(s)$ or, equvalently,

$$\left(\mathrm{e}^{-\int_a^s p(\tau)\,d\tau} y(s)\right)' = \mathrm{e}^{-\int_a^s p(\tau)\,d\tau} q(s).$$

Integrating this relation from a to t, we obtain (1.3).

For general f we can not give y in closed form. Actually, we can not even guarantee existence and/or uniqueness of solutions. It is indeed possible that there is no solution or that there are many solutions. For instance, consider the initial value problem

$$\begin{cases} y' = y^2, \quad 0 \le t \le 2, \\ y(0) = 1. \end{cases} \tag{1.4}$$

Assuming that a solution y exists, we easily see that $y(t) \ge 1$, since $y'(t) \ge 0$, whence y is increasing, and $y(0) = 1$. Therefore, we can write the differential equation in the form

$$\frac{y'(t)}{\big(y(t)\big)^2} = 1 \iff -\frac{d}{dt}\frac{1}{y(t)} = 1.$$

Integrating here from 0 to t, we get

$$-\frac{1}{y(t)} + \frac{1}{y(0)} = t \iff y(t) = \frac{1}{1-t}.$$

We infer that, for $0 \le t < 1$ the unique solution is

$$y(t) = \frac{1}{1-t};$$

notice that $y(t) \to \infty$ for $t \to 1^-$. Consequently, no solution of (1.4) exists (in the whole interval $[0, 2]$). Let us emphasize that f in (1.4), $f(t, y) = y^2$,

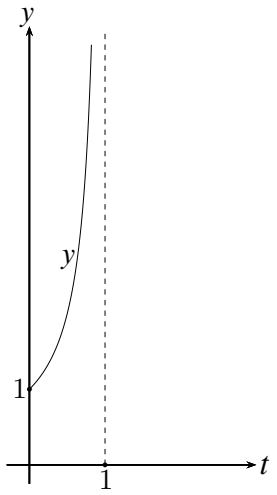

Figure 1.1: The 'solution' $y = \frac{1}{1-t}$ of problem (1.4) in the interval $[0, 1)$.

is as smooth as we like; so smoothness alone does not suffice for existence of solutions.

In contrast, the initial value problem

$$\begin{cases} y' = \sqrt{|y|}, \quad 0 \le t \le 1, \\ y(0) = 0, \end{cases} \tag{1.5}$$

has an infinite number of solutions, namely,

$$y(t) := 0, \ \ 0 \le t \le 1, \quad \text{and} \quad y(t) := \begin{cases} 0 \qquad , \ \ 0 \le t \le t^\star, \\ \dfrac{(t-t^\star)^2}{4}, \ \ t^\star < t \le 1, \end{cases}$$

for any $t^\star \in (0, 1)$. Indeed, first of all it is obvious from the o.d.e. that any solution y of initial value problem (1.5) is an increasing function. Now, if it vanishes in the interval $[0, t^\star]$ and is it does not vanish in a neighborhood at the right of $t^\star$, then it is positive in $(t^\star, 1]$. So, there we can write the o.d.e. in the form

$$y'(t) = \sqrt{y(t)} \iff \frac{y'(t)}{\sqrt{y(t)}} = 1 \iff 2\Big(\sqrt{y(t)}\Big)' = 1;$$

integrating the last relation in $[t^\star, t]$, and using our assumption $y(t^\star) = 0$, we obtain, for $t \geq t^\star$,

$$\sqrt{y(t)} - \sqrt{y(t^\star)} = \frac{t - t^\star}{2} \implies \sqrt{y(t)} = \frac{t - t^\star}{2} \implies y(t) = \frac{(t - t^\star)^2}{4}.$$

Notice that in this case the function f, $f(t, y) = \sqrt{|y|}$, is continuous but it is not differentiable (not even Lipschitz continuous) at $y = 0$.

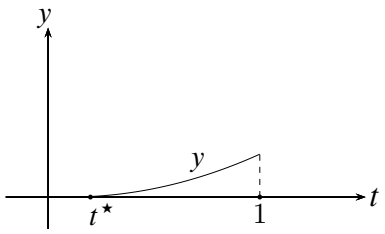

Figure 1.2: A solution of problem (1.5).

Theorem 1.1 (Existence and uniqueness.) *Assume that $f : [a, b] \times \mathbb{R} \to \mathbb{R}$ is continuous and satisfies the Lipschitz condition with respect to y, uniformly with respect to t, that is*

$$(1.6)\quad \exists L \geq 0 \ \forall t \in [a, b] \ \forall y_1, y_2 \in \mathbb{R} \quad \big|f(t, y_1) - f(t, y_2)\big| \leq L|y_1 - y_2|.$$

Then, for any initial value $y_0 \in \mathbb{R}$, initial value problem (1.1) *possesses a unique solution.* □

The proof of Theorem 1.1 is a straightforward application of the Banach fixed point theorem (i.e., the contraction theorem) applied to the integral form

$$(1.7)\qquad y = Ty$$

of (1.1), with the *integral operator* $T : C[a, b] \to C[a, b]$,

$$Tx(t) := (Tx)(t) := y_0 + \int_a^t f\big(s, x(s)\big)\, ds, \quad a \leq t \leq b.$$

A convenient choice of the Banach space is $(C[a, b], \|\cdot\|)$ with the norm $\|\cdot\|$,

$$\|x\| := \max_{a \leq t \leq b} \big(|x(t)|\mathrm{e}^{-2Lt}\big).$$

Notice the equivalence of the norms $\|\cdot\|$ and $\|\cdot\|_\infty$.

The Lipschitz condition (1.6) is very restrictive. Such a simple function as $f(t,y) := y^2$ does not satisfy (1.6), as we easily see. For $p, q \in C[a,b]$ functions $f(t,y) := p(t)y + q(t)$ (linear o.d.e.) and $f(t,y) := p(t)\sin y$ satisfy (1.6). If f is differentiable in its second variable, then it satisfies (1.6), if and only if f_y is bounded; for instance, if

$$\exists M \in \mathbb{R} \;\; \forall t \in [a,b] \;\; \forall y \in \mathbb{R} \quad \big|f_y(t,y)\big| \le M,$$

then f satisfies (1.6) (with $L := M$), as we easily see utilizing the mean value theorem. However, function f in problem (1.5), i.e., $f(y) := \sqrt{|y|}$, does not satisfy (1.6).

Condition (1.6) is referred to as "global" Lipschitz condition. Global because it is required for all reals y_1, y_2. This condition ensures existence and uniqueness in the whole interval $[a,b]$. Replacing (1.6) by a "local" Lipschitz condition, which is a realistic condition, satisfied by many functions, we can still show existence and uniqueness but in a subinterval $[a,b']$, with appropriate b', of the interval $[a,b]$. The exact result is:

Theorem 1.2 (Local existence and uniqueness of solutions.) *Let $c > 0$ and $f \in C\big([a,b] \times [y_0 - c, y_0 + c]\big)$. If f satisfies the Lipschitz condition, with respect to y, in $[a,b] \times [y_0 - c, y_0 + c]$, uniformly with respect to t, i.e.,*

$$\begin{aligned} &\exists L \ge 0 \;\; \forall t \in [a,b] \;\; \forall y_1, y_2 \in [y_0 - c, y_0 + c] \\ &\qquad \big|f(t,y_1) - f(t,y_2)\big| \le L|y_1 - y_2|, \end{aligned} \tag{1.8}$$

then problem (1.1) *is uniquely solvable, at least in an interval $[a,b']$, with*

$$b' := \min\big(b, a + \frac{c}{A}\big),$$

where

$$A := \max_{\substack{a \le t \le b \\ y_0 - c \le y \le y_0 + c}} \big|f(t,y)\big|. \qquad \square$$

Condition (1.8) is much milder than (1.6). Every function $f \in C\big([a,b] \times [y_0 - c, y_0 + c]\big)$, continuously differentiable in its second argument in the

interval $[y_0 - c, y_0 + c]$, satisfies (1.8). The proof of Theorem 1.2 is not difficult; is suffices to appropriately modify the proof of Theorem 1.1.

Let us also note that the continuity of f alone, $f \in C\big([a,b]\times\mathbb{R}\big)$, suffices for existence of a solution of (1.1) in some interval of the form $[a,c], c > a$. However, it does not ensure uniqueness, as we see from example (1.5). (In this example, as we easily see, the function $f(y) := \sqrt{|y|}$ does not satisfy the local Lipschitz condition, with respect to y, in any interval containing zero.)

1.2 Stability

In this section we will be concerned with stability properties of initial value problem (1.1). We assume that f is continuous. We will focus on two cases, which will play a central role in our study of numerical methods for initial value problems in the sequel: In the first case we assume that f satisfies the (global) Lipschitz condition (1.6), while in the second that it satisfies the so-called one-sided Lipschitz condition; cf. (1.13) in the sequel. For given initial values $y_0, z_0 \in \mathbb{R}$, we consider the initial value problems

$$(1.9)\quad \begin{cases} y' = f(t,y), \quad a \le t \le b, \\ y(a) = y_0, \end{cases} \quad \text{and} \quad \begin{cases} z' = f(t,z), \quad a \le t \le b, \\ z(a) = z_0. \end{cases}$$

First case. Assume that f satisfies the global Lipschitz condition (1.6). Then, according to Theorem 1.1, the initial value problems (1.9) possess unique solutions $y, z \in C^1[a,b]$, respectively. According to (1.9), letting $\varepsilon(t) := y(t) - z(t)$, we have $\varepsilon'(t) = f(t,y) - f(t,z)$, for $t \in [a,b]$. To obtain an estimate of $\big|\varepsilon(t)\big|$ or, equivalently, of $\big(\varepsilon(t)\big)^2$, we need information about its derivative. Multiplying the previous relation by $\varepsilon(t)$, we have

$$\varepsilon(t)\varepsilon'(t) = \big(f(t,y) - f(t,z)\big)\varepsilon(t),$$

from which, in view of the Lipschitz condition (1.6), we obtain

$$\varepsilon(t)\varepsilon'(t) = \frac{1}{2}\frac{d}{dt}\varepsilon^2(t) \le \big|f(t,y) - f(t,z)\big|\,\big|\varepsilon(t)\big| \le L\varepsilon^2(t),$$

for all $t \in [a,b]$. Consequently, setting $\varepsilon^2(t) =: \varphi(t)$ we have

$$(1.10)\qquad \varphi' - 2L\varphi \le 0, \quad t \in [a,b].$$

To solve this differential inequality, we multiply by the integrating factor e^{-2Lt}, and obtain

$$\mathrm{e}^{-2Lt}\varphi'(t) - 2L\mathrm{e}^{-2Lt}\varphi(t) = \frac{d}{dt}\left(\mathrm{e}^{-2Lt}\varphi(t)\right) \leq 0, \quad t \in [a,b].$$

We infer that $\mathrm{e}^{-2Lt}\varphi(t)$ is decreasing in $[a,b]$. In particular, we have

$$\mathrm{e}^{-2Lt}\varphi(t) \leq \mathrm{e}^{-2La}\varphi(a), \quad a \leq t \leq b,$$

from which we obtain

$$\left|\varepsilon(t)\right| \leq \mathrm{e}^{L(t-a)}\left|\varepsilon(a)\right|, \quad a \leq t \leq b, \tag{1.11}$$

and

$$\max_{a\leq t\leq b}\left|y(t) - z(t)\right| \leq \mathrm{e}^{L(b-a)}|y_0 - z_0|. \tag{1.12}$$

Estimate (1.12) expresses the *continuous dependence* of the solution y of (1.1) on the initial data $y_0 \in \mathbb{R}$, in the maximum norm $\|\cdot\|_\infty$,

$$\|y\|_\infty := \max_{a\leq t\leq b}\left|y(t)\right|.$$

Notice that the constant on the right-hand side of (1.12) increases exponentially with the Lipschitz constant L. When L is very large, the corresponding estimate is not particularly useful in practice. In the sequel we will derive another estimate, avoiding the Lipschitz condition (1.6), provided f satisfies (1.13).

Second case. In this case we assume that f satisfies the *one-sided Lipschitz condition*; cf. (1.13) in the sequel. The motivation for this condition is that it is often satisfied in applications, in particular in the case of parabolic p.d.e's. We write it in the form

$$\forall t \in [a,b] \ \ \forall y_1, y_2 \in \mathbb{R} \quad \big(f(t,y_1) - f(t,y_2)\big)(y_1 - y_2) \leq 0, \tag{1.13}$$

which can be easily generalized to systems of o.d.e's and also to more general equations; see Exercise 1.5. In the scalar case, condition (1.13) means simply that f is a decreasing function of its second argument, for each fixed value of

its first variable. We now proceed to the stability proof along the lines of the corresponding proof in the first case. Multiplying $\varepsilon'(t) = f(t,y) - f(t,z)$ by $\varepsilon(t)$, we obtain

$$\varepsilon(t)\varepsilon'(t) = \big(f(t,y) - f(t,z)\big)\varepsilon(t),$$

whence, in view of (1.13),

$$\varepsilon(t)\varepsilon'(t) = \frac{1}{2}\frac{d}{dt}\varepsilon^2(t) \leq 0,$$

for all $t \in [a,b]$. Therefore, ε^2 is a decreasing function of t. Hence, $|\varepsilon|$ is also decreasing and we infer that

$$\max_{a\leq t\leq b} \big|y(t) - z(t)\big| \leq |y_0 - z_0|. \tag{1.14}$$

Notice that the Lipschitz constant L does not enter in (1.14) (actually we have not assumed the Lipschitz condition (1.6)), in contrast to the estimate (1.12).

Uniqueness of solutions in the case f satisfies the one-sided Lipschitz condition is a trivial consequence of (1.14). Moreover, (1.13) guarantees also existence of solutions, provided f is continuous in $[a,b]\times\mathbb{R}$. Indeed, first, local existence in an interval of the form $[a,b')$ is ensured by the continuity of f. To show global existence, in the whole interval $[a,b]$, we will utilize (1.13) to prove that all possible solutions y are bounded in $[a,b]$. In fact, we have

$$y'(t) = \big[f\big(t,y(t)\big) - f(t,0)\big] + f(t,0),$$

whence

$$y'(t)y(t) = \big[f\big(t,y(t)\big) - f(t,0)\big]y(t) + f(t,0)y(t).$$

Now, obviously, the term on the left-hand side is twice the derivative of $\big(y(t)\big)^2$, the first term on the right-hand side is non-positive and the second can be easily estimated with the inequality $2xz \leq x^2 + z^2$. Then, we obtain

$$\big[\big(y(t)\big)^2\big]' \leq \big(f(t,0)\big)^2 + \big(y(t)\big)^2,$$

whence

$$\big[\mathrm{e}^{-t}\big(y(t)\big)^2\big]' \leq \mathrm{e}^{-t}\big(f(t,0)\big)^2.$$

Integrating this relation in the interval $[a,t]$, we get

$$\mathrm{e}^{-t}\big(y(t)\big)^2 - \mathrm{e}^{-a}\big(y(a)\big)^2 \le \int_a^t \mathrm{e}^{-s}\big(f(s,0)\big)^2\,ds,$$

and obtain easily the desired inequality

$$\big(y(t)\big)^2 \le \mathrm{e}^b\Big[(y_0)^2\mathrm{e}^{-a} + \int_a^b \mathrm{e}^{-s}\big(f(s,0)\big)^2\,ds\Big], \quad a \le t \le b.$$

We conclude that y is bounded in the interval $[a,b]$. This fact ensures global existence of the solution y in $[a,b]$; indeed, it is well known that, assuming continuity of f, a solution of the initial value problem in an interval of the form $[a,s)$ can not be extended to the interval $[a,s]$ in two cases only, namely when its limit from the left at s is either ∞ or $-\infty$, $\lim_{t\to s^-} y(t) = \infty$ or $\lim_{t\to s^-} y(t) = -\infty$. But, when y is bounded in $[a,b]$, both these possibilities are excluded, and we infer that y is indeed a global solution.

Now, we will consider some special cases; the motivation is two-fold, to get some insight of the one-sided Lipschitz condition and to explain why we will consider some special initial value problems in the sequel, often referred to as *test problems*. If f is linear in y, $f(t,y) = \lambda(t)y + \mu(t)$, cf. initial value problem (1.2), then, obviously, it satisfies condition (1.13), if and only if the function $\lambda(t)$ takes on only non-positive values. Let us also note that, although in the general case stability refers to the difference of two solutions, in the linear case it suffices to study the behaviour of one (nontrivial) solution of the corresponding homogeneous equation, since, obviously, the difference of two solutions is a solution of the corresponding homogeneous equation. In other words, the function μ is irrelevant for stability, and it suffices to study initial value problems of the form

$$\begin{cases} y' = \lambda(t)y, \quad a \le t \le b, \\ y(a) = y_0. \end{cases} \tag{1.15}$$

The stability estimate (1.14) takes now the form

$$\max_{a\le t\le b} \big|y(t)\big| \le |y_0|, \tag{1.16}$$

provided $\lambda(t)$ takes on only non-positive values. Also, without loss of generality we can choose an arbitrary non-vanishing initial value y_0 in the initial value problem (1.15), since, if y is the solution for $y_0 = 1$, then the solution in the case of an arbitrary initial value y_0 is the product $y_0 y$. Furthermore, the case $\lambda(t)$ is constant, i.e., independent of t, is of particular interest. So we will also consider the test problem

$$\begin{cases} y' = \lambda y, \quad t \geq 0, \\ y(0) = 1, \end{cases} \tag{1.17}$$

with a real constant λ. The case λ is non-positive is more interesting, since then the one-sided Lipschitz condition (1.13) is satisfied. Obviously, the solution of the test problem (1.17) is $y(t) = \mathrm{e}^{\lambda t}$; we are interested in the behaviour of the numerical approximations when numerical schemes are applied to this problem. Indeed, in the sequel we will consider a more general test problem, namely (1.17) with complex constant λ, with complex-valued solution y, that is

$$\begin{cases} y' = \lambda y, \quad t \geq 0, \\ y(0) = 1; \end{cases} \tag{1.18}$$

again, particularly interesting is the case when the real part of λ is non-positive, $\mathrm{Re}\,\lambda \leq 0$, since then $|y(t)| = \mathrm{e}^{(\mathrm{Re}\,\lambda)t}$ is decreasing.

1.3 Systems of o.d.e's

The generalization of the results of sections 1.1 and 1.2 to *systems of first order o.d.e's* is straightforward. Let $m \in \mathbb{N}$, $f : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$, and $y_0 \in \mathbb{R}^m$. We seek a vector-valued function $y : [a,b] \to \mathbb{R}^m$ such that

$$\begin{cases} y'(t) = f\big(t, y(t)\big), \quad a \leq t \leq b, \\ y(a) = y_0. \end{cases} \tag{1.19}$$

The results of section 1.1 are also valid for problem (1.19), if we replace the absolute value by any norm $\|\cdot\|$ of $\mathbb{R}^m$. The analogue of Theorem 1.1, for instance, reads in this case as follows:

Theorem 1.3 (Existence and uniqueness of solutions for systems of o.d.e's.) *Let $f : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function, satisfying the Lipschitz condition with respect to y, uniformly with respect to t, in a norm $\|\cdot\|$ of $\mathbb{R}^m$, that is*

$$(1.20)\quad \exists L \in \mathbb{R}\ \ \forall t \in [a,b]\ \ \forall y_1, y_2 \in \mathbb{R}^m\ \ \big\| f(t,y_1) - f(t,y_2) \big\| \le L \|y_1 - y_2\|.$$

Then, for any initial value $y_0 \in \mathbb{R}^m$, initial value problem (1.19) *is uniquely solvable.* □

Condition (1.20) is, again, very restrictive. In case the function f is continuously differentiable for $(t,y) \in [a,b] \times \mathbb{R}^m$, then it satisfies the Lipschitz condition, if and only if the first partial derivatives of f with respect to the variables y_i are bounded in $(t,y) \in [a,b] \times \mathbb{R}^m$. For example, if

$$M := \sup_{\substack{1 \le i \le m \\ (t,y) \in [a,b] \times \mathbb{R}^m}} \sum_{j=1}^{m} \Big| \frac{\partial f_i}{\partial y_j}(t,y) \Big| < \infty,$$

then f satisfies (1.20) with respect to the norm with $L = M$.

The analogue of problem (1.2) is now of the form

$$(1.21)\qquad \begin{cases} y'(t) = M(t)y(t) + g(t), & a \le t \le b, \\ y(a) = y_0, \end{cases}$$

with $g(t) \in \mathbb{R}^m$ and $M(t) \in \mathbb{R}^{m,m}$, for $t \in [a,b]$. This problem possesses a unique solution, if, e.g., g and M are continuous functions of t, for $t \in [a,b]$.

Problems of the form (1.19) arise often in applications. Moreover, initial value problems for *higher order differential equations* can be reduced to initial value problems for systems of the form (1.19). Consider, for example, the initial value problem

$$(1.22)\qquad \begin{cases} y^{(m)}(t) = f\big(t, y(t), y'(t), \dots, y^{(m-1)}(t)\big), & a \le t \le b, \\ y^{(i)}(a) = y_i, & i = 0, \dots, m-1. \end{cases}$$

Letting

$$z(t) := \big(y(t), y'(t), \dots, y^{(m-1)}(t)\big)^T, \quad z_0 := (y_0, y_1, \dots, y_{m-1})^T,$$

problem (1.22) can be written as

$$
\left\{
\begin{array}{l}
z'(t) = \begin{pmatrix} z_2(t) \\ z_3(t) \\ \vdots \\ z_m(t) \\ f\big(t, z_1(t), \ldots, z_m(t)\big) \end{pmatrix}, \quad a \le t \le b, \\
\\
z(a) = z_0.
\end{array}
\right. \tag{1.23}
$$

Analogously, initial value problems for systems of higher order differential equations can be reduced to systems of the form (1.19).

We close this chapter by a short discussion of stability properties of solutions of initial value problems for systems of o.d.e's. In case f satisfies the Lipschitz condition (1.20) with respect to the Euclidean norm, the analogue of estimate (1.12) is still valid, as we easily see; cf. Exercise 1.3 and the hint given there. Stability in an arbitrary norm follows easily from the estimate in the Euclidean norm by utilizing the equivalence of norms in $\mathbb{R}^m$. The right-hand side in the stability estimate is now multiplied by an appropriate constant.

Next let us look at the case that f satisfies the one-sided Lipschitz condition. Without loss of generality, we restrict ourselves to the case of the Euclidean norm. Then the generalization of (1.13) reads as follows: A function $f : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ satisfies the one-sided Lipschitz condition with respect to its second argument, if

$$
\forall t \in [a,b] \;\; \forall x, \tilde{x} \in \mathbb{R}^m \quad \big(f(t,x) - f(t,\tilde{x}), x - \tilde{x}\big) \le 0, \tag{1.24}
$$

with $(\cdot,\cdot)$ the Euclidean inner product in $\mathbb{R}^m$. The analogue of the estimate (1.14) is valid also in this case; cf. Exercise 1.5. The initial value problem with functions f satisfying (1.24) is a particularly interesting test problem that will attract our interest in our study of numerical methods in subsequent chapters. In the case of linear systems of o.d.e's, cf. problem (1.21), it is easily seen that function $f(t,y) = M(t)y + g(t)$ satisfies (1.24), if and only if the

matrices $M(t) \in \mathbb{R}^{m,m}, t \in [a,b]$, are negative semidefinite, that is

$$\forall t \in [a,b] \;\; \forall x \in \mathbb{R}^m \quad \big(M(t)x, x\big) \leq 0; \tag{1.25}$$

this condition is the counterpart of the non-positivity of the function $\lambda(t)$ in the scalar case; see problem (1.15).

A useful test problem is

$$\begin{cases} y'(t) = My(t), & t \geq 0, \\ y(0) = y_0, \end{cases} \tag{1.26}$$

with a 2×2 matrix of the form

$$M := \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}.$$

Obviously, M can be written in the form

$$M = \alpha \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \beta \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$

i.e., as the linear combination of two particularly simple matrices, one symmetric and one antisymmetric, respectively. Therefore, we have

$$(Mx, x) = \alpha \|x\|^2 \quad \forall x \in \mathbb{R}^2.$$

Notice that the linear equation $y' = \lambda y$ of test problem (1.18) with coefficient $\lambda = \alpha + \beta \mathrm{i}, \alpha, \beta \in \mathbb{R}$, can be written as a real system in the form

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}' = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = M \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

with y_1 and y_2 the real and imaginary parts of y, respectively. This is our motivation to consider test problem (1.26). The eigenvalues of M are obviously $\alpha \pm \beta \mathrm{i}$, i.e, λ and $\bar{\lambda}$. Also, the matrix M is negative semidefinite, compare to (1.25), if and only if α is non-positive.

Existence and uniqueness proofs for solutions of initial value problems can be found in all o.d.e. books, for instance in Coddington and Levinson [6], Birkhoff and Rota [3], and Walter [24].

Exercises

1.1 Consider the initial value problem

$$\begin{cases} y' = y^2, \quad 0 \le t < b, \\ y(0) = \alpha. \end{cases} \tag{1.27}$$

If α is positive, show that the largest possible value of b, for which the problem possesses a solution, is $b = 1/\alpha$. If α is negative, show that the problem possesses a solution even for $b = \infty$.

1.2 Consider the initial value problem

$$\begin{cases} y' = \sqrt{|y|}, \quad 0 \le t < b, \\ y(0) = \alpha. \end{cases} \tag{1.28}$$

If α is positive, determine the unique solution of the problem in $[0, \infty)$, i.e., with $b = \infty$. If α is negative, determine the unique solution for $b = 2\sqrt{-\alpha}$.
[*Hint*: Show that

$$y(t) = \big(\frac{t}{2} + \sqrt{\alpha}\big)^2, \text{ for } \alpha > 0, \quad \text{and} \quad y(t) = -\big(-\frac{t}{2} + \sqrt{-\alpha}\big)^2, \text{ for } \alpha < 0.\,]$$

1.3 The stability estimates (1.11) and (1.12) can be generalized to systems of o.d.e's: Let $f : [a, b] \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function satisfying the Lipschitz condition (1.20) with respect to the Euclidean norm $\|\cdot\|$ of $\mathbb{R}^m$. Let y and z be the solutions of the initial value problems

$$\begin{cases} y' = f(t, y), \quad t \in [a, b], \\ y(a) = y_0, \end{cases} \quad \text{and} \quad \begin{cases} z' = f(t, z), \quad t \in [a, b], \\ z(a) = z_0, \end{cases}$$

respectively. Show that

$$\|y(t) - z(t)\| \le \mathrm{e}^{L(t-a)} \|y_0 - z_0\|$$

for all $t \in [a, b]$,
[*Hint*: Let $(\cdot, \cdot)$ denote the Euclidean inner product in $\mathbb{R}^m$. Then, for a differentiable function $x : [a, b] \to \mathbb{R}^m$, we have

$$\begin{aligned} \frac{d}{dt}\|x(t)\|^2 &= \frac{d}{dt}\big[\big(x_1(t)\big)^2 + \cdots + \big(x_m(t)\big)^2\big] = 2\big[x_1(t)x_1'(t) + \cdots + x_m(t)x_m'(t)\big] \\ &= 2\big(x'(t), x(t)\big).\,] \end{aligned}$$

1.4 Consider the initial value problems (1.9), and assume this time that the continuous function $f : [a,b] \times \mathbb{R} \to \mathbb{R}$ satisfies the condition

$$\forall t \in [a,b] \;\; \forall y_1, y_2 \in \mathbb{R} \quad \big(f(t,y_1) - f(t,y_2)\big)(y_1 - y_2) \le \nu (y_1 - y_2)^2,$$

with a constant ν. (Notice that for $\nu = 0$ this condition coincides with (1.13).) Conditions of this form are also referred to as one-sided Lipschitz conditions. Show that

$$|y(t) - z(t)| \le \mathrm{e}^{\nu(t-a)} |y_0 - z_0|,$$

for all $t \in [a,b]$.

1.5 The stability estimate (1.14) as well as Exercise 1.4 can be generalized to systems of o.d.e's. The generalization of (1.14) is: Let $f : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function satisfying the one-sided Lipschitz condition (1.24), with respect to its second argument. Let y and z be the solutions of the initial value problems

$$\begin{cases} y' = f(t,y), & t \in [a,b], \\ y(a) = y_0, \end{cases} \quad \text{and} \quad \begin{cases} z' = f(t,z), & t \in [a,b], \\ z(a) = z_0, \end{cases}$$

respectively. Show that $\|y(\cdot) - z(\cdot)\|$ is a decreasing function, whence, in particular,

$$\|y(t) - z(t)\| \le \|y_0 - z_0\|,$$

for all $t \in [a,b]$. We used here the notation $(\cdot,\cdot)$ and $\|\cdot\|$ for the Euclidean inner product and the Euclidean norm, respectively, in $\mathbb{R}^m$.

1.6 Consider the initial value problem

$$\begin{cases} y' = f(t,y), & t \in [a,b], \\ y(a) = y_0, \end{cases}$$

and assume that the function f satisfies the one-sided Lipschitz condition given in Exercise 1.4. Check that with the change of variables

$$u(t) := \mathrm{e}^{-\nu(t-a)} y(t)$$

the problem takes the form

$$\begin{cases} u' = F(t,u), & t \in [a,b], \\ u(a) = y_0, \end{cases}$$

with

$$F(t,v) := \mathrm{e}^{-\nu(t-a)} f\left(t, \mathrm{e}^{\nu(t-a)} v\right) - \nu v$$

and show that F satisfies condition (1.13),

$$\forall t \in [a,b] \;\; \forall y_1, y_2 \in \mathbb{R} \quad \big(F(t,y_1) - F(t,y_2)\big)(y_1 - y_2) \le 0.$$

1.7 (The Gronwall inequality in integral form.) Let φ be a continuous function in the interval $[0,T]$, and $\alpha, \beta \in \mathbb{R}$ with $\beta \ge 0$. If

$$\varphi(t) \le \alpha + \beta \int_0^t \varphi(s)\, ds \quad \forall t \in [0,T],$$

show that

$$\varphi(t) \le \alpha\, \mathrm{e}^{\beta t} \quad \forall t \in [0,T].$$

[*Hint*: Let ε be a positive number and check that the function $\psi, \psi(t) := (\alpha + \varepsilon)\mathrm{e}^{\beta t}$, $t \in [0,T]$, satisfies the integral equation

$$\psi(t) = \alpha + \varepsilon + \beta \int_0^t \psi(s)\, ds \quad \forall t \in [0,T].$$

Obviously, $\varphi(0) < \psi(0)$. Assume t_0 is the smallest number in the interval $[0,T]$ such that $\varphi(t_0) = \psi(t_0)$. Show that this leads to the contradiction $\varphi(t_0) < \psi(t_0)$.]

1.8 (The Gronwall inequality in differential form.) Assume that the function φ of Exercise 1.7 is continuously differentiable in the interval $[0,T]$ and satisfies the inequality

$$\varphi'(t) \le \beta\varphi(t) \quad \forall t \in [0,T].$$

Show that

$$\varphi(t) \le \varphi(0)\mathrm{e}^{\beta t} \quad \forall t \in [0,T].$$

[*Hint*: Show that

$$\varphi(t) \le \varphi(0) + \beta \int_0^t \varphi(s)\, ds \quad \forall t \in [0,T]$$

and utilize Exercise 1.7. Alternatively, write the inequality in the form

$$\left(\mathrm{e}^{-\beta s}\varphi(s)\right)' \le 0$$

and integrate from 0 to t.]

1.9 Let $a \in \mathbb{R}$ and $f : [0, \infty) \to \mathbb{R}$ be a continuous function. Show that the solution y of the initial value problem

$$\begin{cases} y'(t) = ay(t) + f(t), & t \geq 0, \\ y(0) = y_0 \end{cases}$$

is given by

$$y(t) = \mathrm{e}^{at} y_0 + \int_0^t \mathrm{e}^{a(t-s)} f(s)\, ds, \quad t \geq 0;$$

cf. (1.3). Notice that the first term on the right-hand side represents the solution of the problem for the homogeneous equation

$$\begin{cases} x'(t) = ax(t), & t \geq 0, \\ x(0) = y_0, \end{cases}$$

while the integrand $\mathrm{e}^{a(t-s)} f(s)$ represents the value at t of the solution of problem

$$\begin{cases} x'(t) = ax(t), & t \geq s, \\ x(s) = f(s). \end{cases}$$

Interpret (1.3) in a similar way.

1.10 Let $M \in \mathbb{R}^{m,m}$ be a matrix.

a) In analogy to the definition of the exponential function e^x for real arguments x, we define the matrix e^M as

$$\mathrm{e}^M := \sum_{\ell=0}^{\infty} \frac{1}{\ell!} M^\ell.$$

Let $\|\cdot\|$ be a matrix norm. Prove that

$$\forall \varepsilon > 0 \ \ \exists n \in \mathbb{N} \ \ \forall k \in \mathbb{N} \quad \Big\| \sum_{\ell=n}^{n+k} \frac{1}{\ell!} M^\ell \Big\| \leq \varepsilon,$$

i.e., the series converges, whence the matrix e^M is well defined.

[*Hint*: Use the estimate

$$\Big\| \sum_{\ell=n}^{n+k} \frac{1}{\ell!} M^\ell \Big\| \leq \sum_{\ell=n}^{n+k} \frac{1}{\ell!} \|M\|^\ell$$

and the fact that the series $\sum_{\ell=0}^{\infty} \frac{1}{\ell!} x^\ell$ converges for all $x \in \mathbb{R}$.]

b) Consider the initial value problem

$$\begin{cases} y'(t) = My(t), & t \geq 0, \\ y(0) = y_0. \end{cases}$$

Show that its solution y is given by

$$y(t) = \mathrm{e}^{tM} y_0, \quad t \geq 0.$$

[*Hint*: Show that

$$\left(\mathrm{e}^{tM}\right)' = \Big(\sum_{\ell=0}^{\infty} \frac{1}{\ell!} t^\ell M^\ell\Big)' = \sum_{\ell=1}^{\infty} \frac{1}{(\ell-1)!} t^{\ell-1} M^\ell = M\mathrm{e}^{tM}.]$$

c) Let $E(t)$ denote the *solution operator* of the initial value problem of part b), i.e., $E(t) = \mathrm{e}^{tM}$, whence $y(t) = E(t)y_0$. In analogy to the relation $\mathrm{e}^{x+y} = \mathrm{e}^x \mathrm{e}^y$ for real numbers x and y, for matrices $A, B \in \mathbb{R}^{m,m}$ there holds $\mathrm{e}^{A+B} = \mathrm{e}^A \mathrm{e}^B$, provided A and B *commute*, i.e., $AB = BA$. Assuming this fact, show that the solution operator $E(t)$ satisfies the *semigroup property*, that is

$$E(\sigma + \tau) = E(\sigma)E(\tau) \quad \forall \sigma, \tau \geq 0.$$

This means that whether we consecutively solve the problems

$$\begin{cases} x'(t) = Mx(t), & 0 \leq t \leq \sigma, \\ x(0) = y_0 \end{cases}$$

and

$$\begin{cases} x'(t) = Mx(t), & \sigma \leq t \leq \sigma + \tau, \\ x(\sigma) = E(\sigma)y_0 \end{cases}$$

or directly the original problem in the interval $[0, \sigma+\tau]$, we obtain the same result, $y(\sigma+\tau) = x(\sigma+\tau)$. In other words, to the value of the solution at the time level $\sigma+\tau$ we can be led either by solving the original problem in the interval $[0, \sigma+\tau]$, or by first solving the problem in the interval $[0, \sigma]$ and subsequently in the interval $[\sigma, \sigma+\tau]$, of course with the right initial value $E(\sigma)y_0$.

1.11 Let $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ be a diagonal matrix and $M \in \mathbb{R}^{m,m}$ be a diagonalizable matrix, $M = U\Lambda U^{-1}$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. Prove that

a) $\mathrm{e}^{\Lambda} = \operatorname{diag}\left(\mathrm{e}^{\lambda_1}, \ldots, \mathrm{e}^{\lambda_m}\right)$

b) $\mathrm{e}^{M} = U\mathrm{e}^{\Lambda}U^{-1}$, whence $\mathrm{e}^{tM} = U\mathrm{e}^{t\Lambda}U^{-1}$, for $t \in \mathbb{R}$.

1.12 Let $f : [0, \infty) \to \mathbb{R}^m$ be a continuous function. With the notation of Exercise 1.10, consider the initial value problem

$$\begin{cases} y'(t) = My(t) + f(t), & t \geq 0, \\ y(0) = y_0. \end{cases}$$

Show that

$$\left(\mathrm{e}^{-tM} y(t)\right)' = \mathrm{e}^{-tM} f(t)$$

and infer that

$$y(t) = \mathrm{e}^{tM} y_0 + \int_0^t \mathrm{e}^{(t-s)M} f(s)\, ds, \quad t \geq 0.$$

(Compare with the corresponding result in Exercise 1.9.) This relation can be written in the form

$$y(t) = E(t)y_0 + \int_0^t E(t-s) f(s)\, ds,$$

where $E(t) = \mathrm{e}^{tM}$; this relation is referred to as *Duhamel's principle*. Notice that the first term on the right-hand side is the value at t of the solution of the corresponding homogeneous problem

$$\begin{cases} x'(t) = Mx(t), & t \geq 0, \\ x(0) = y_0, \end{cases}$$

while $E(t-s) f(s)$ is the value at t of the solution of problem

$$\begin{cases} x'(t) = Mx(t), & t \geq s, \\ x(s) = f(s). \end{cases}$$

1.13 Let $M \in \mathbb{C}^{m,m}$ be a matrix with eigenvalues $\lambda_1, \ldots, \lambda_m$ with non-positive real parts, $\operatorname{Re} \lambda_i \leq 0, i = 1, \ldots, m$. Consider the initial value problem

$$\begin{cases} y'(t) = My(t), & t \geq 0, \\ y(0) = y_0 \end{cases}$$

with non-vanishing y_0, $y_0 \neq 0$. Our aim in this Exercise, as well as in the following three Exercises, is to study the behaviour of the ratio

$$\frac{\|y(t)\|}{\|y_0\|},$$

with $\|\cdot\|$ a norm of $\mathbb{C}^m$.

a) If $m = 1$, show that

$$|y(t)| \leq |y_0|, \quad t \geq 0.$$

b) If $m = 2$, and $M = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$, in which case $\lambda_1 = \lambda_2 = 0$, show that the solution $y(t)$ is given by

$$y(t) = \begin{pmatrix} (y_0)_1 + (y_0)_2 t \\ (y_0)_2 \end{pmatrix}, \quad t \geq 0,$$

with $(y_0)_1$ and $(y_0)_2 \neq 0$ the components of y_0, and infer that

$$\frac{\|y(t)\|}{\|y_0\|} \to \infty, \quad t \to \infty,$$

in contrast to the case $m = 1$ that we examined in part a).

1.14 Let $\mu \in \mathbb{R}$. Using the notation of Exercise 1.13 we assume now that $M = \left(\begin{smallmatrix} -1 & \mu \\ 0 & 0 \end{smallmatrix}\right)$, in which case the eigenvalues are $\lambda_1 = -1$ και $\lambda_2 = 0$. The solution $y(t)$ is given by

$$y(t) = \begin{pmatrix} (y_0)_1 \mathrm{e}^{-t} + (y_0)_2 \mu (1 - \mathrm{e}^{-t}) \\ (y_0)_2 \end{pmatrix}, \quad t \geq 0.$$

In particular, for $(y_0)_1 = 0$ and a norm $\|\cdot\|_p (p \geq 1)$ in $\mathbb{R}^2$, show that

$$\frac{\|y(t)\|_p}{\|y_0\|_p} \leq \left(1 + |\mu|^p\right)^{1/p}, \quad t \geq 0.$$

Compare with the result in the case $m = 1$, examined in Exercise 1.13a, where the corresponding ratio does not exceed one.

1.15 Let $\lambda \in \mathbb{C}$ be a number with *negative* real part, $\mathrm{Re}\,\lambda < 0$. Use the notation of Exercise 1.13 and assume that the matrix M has entries λ in its diagonal, one in its upper diagonal, and zero elsewhere,

$$M = \begin{pmatrix} \lambda & 1 & & & 0 \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 0 & & & & \lambda \end{pmatrix};$$

obviously, the eigenvalues are $\lambda_1 = \cdots = \lambda_m = \lambda$. Check that the solution $y(t) = \big(y_1(t), \ldots, y_m(t)\big)^T$ is given by

$$\begin{cases} y_m(t) = y_m(0)\mathrm{e}^{\lambda t}, \\ y_i(t) = y_i(0)\mathrm{e}^{\lambda t} + \displaystyle\int_0^t y_{i+1}(s)\mathrm{e}^{\lambda(t-s)}\,ds, \quad i = m-1, \ldots, 1, \end{cases}$$

and utilize the fact that the function φ, $\varphi(t) := \int_0^t |\mathrm{e}^{\lambda(t-s)}|\,ds$, is uniformly bounded to infer that

$$\|y(t)\|_\infty \le C\,\|y_0\|_\infty, \quad t \ge 0,$$

for some constant C. Compare with the case examined in Exercise 1.13b.

1.16 We use the notation of Exercise 1.13 and assume this time, besides $\operatorname{Re}\lambda_i \le 0, i = 1, \ldots, m$, that $\operatorname{Re}\lambda_i < 0$, if λ_i is a *multiple* eigenvalue. Show that

$$\|y(t)\| \le C\,\|y_0\|, \quad t \ge 0,$$

for some constant C, depending also on the norm $\|\cdot\|$.
[*Hint*: For $m = 1$ the claim is obviously valid; cf. Exercise 1.13a. For $m > 1$, let $T \in \mathbb{C}^{m,m}$ be a matrix such that $T^{-1}MT = J$ is the Jordan normal form of M. With $x(t) := T^{-1}y(t)$ write the system of o.d.e's in the form

$$x'(t) = Jx(t).$$

This problem can be decomposed into problems of the form studied in Exercise 1.15 and problems with $m = 1$. Consequently,

$$\|x(t)\|_\infty \le \widetilde{C}\,\|x(0)\|_\infty, \quad t \ge 0,$$

and the equivalence of norms in $\mathbb{C}^m$ leads to the desired result.]

1.17 With the notation of Exercise 1.13, we assume that the matrix $M \in \mathbb{R}^{m,m}$ is *symmetric*, and that its (real) eigenvalues are non-positive, $\lambda_i \le 0, 1 \le i \le m$, (i.e., the matrix is negative semidefinite).

a) Let $\varphi : (-\infty, 0] \to \mathbb{R}$ be a bounded, continuous function. The operator (matrix) $\varphi(M) \in \mathbb{R}^{m,m}$ is defined as follows: Consider the orthonormal, with respect to the Euclidean inner product $(\cdot,\cdot)$ in $\mathbb{R}^m$, eigenvectors $v^{(i)}, i = 1, \ldots, m$, of M,

such that $Mv^{(i)} = \lambda_i v^{(i)}, 1 \le i \le m$. For every $v \in \mathbb{R}^m$ we define the action of $\varphi(M)$ by the relation ('spectral' representation of M)

$$\varphi(M)v = \sum_{i=1}^{m} \varphi(\lambda_i)(v, v^{(i)})v^{(i)}\,.$$

Show that

$$\|\varphi(M)\|_2 = \max_{1\le i\le m} |\varphi(\lambda_i)|,$$

where $\|\cdot\|_2$ is the matrix norm induced by the Euclidean norm in $\mathbb{R}^m$.

b) Utilizing the representation

$$y(t) = \mathrm{e}^{tM} y(0), \quad t \ge 0,$$

of the solution y of our problem, show that

$$\|y(t)\|_2 \le \mathrm{e}^{t(\max_i \lambda_i)}\|y(0)\|_2, \quad t \ge 0.$$

c) Show that the definition e^{tM} we used here is compatible with the one of Exercise 1.10a.

1.18 (Square root of a matrix.) To get some insight in the definition of $\varphi(M)$ of the previous Exercise, we focus here on the square root of a matrix. Let $M \in \mathbb{R}^{m,m}$ be a symmetric and positive semidefinite matrix, that is, such that $(Mx, x) \ge 0$, for all $x \in \mathbb{R}^m$. Then the (real) eigenvalues of M are non-negative, $\lambda_i \ge 0, 1 \le i \le m$. We consider the orthonormal, with respect to the Euclidean inner product $(\cdot,\cdot)$ in $\mathbb{R}^m$, eigenvectors $v^{(i)}, i = 1, \ldots, m$, of M, such that $Mv^{(i)} = \lambda_i v^{(i)}, 1 \le i \le m$, and, in analogy to the definition $\varphi(M)$ in the previous Exercise, we define the square root $M^{1/2} \in \mathbb{R}^{m,m}$ of M through the relation

$$M^{1/2}v = \sum_{i=1}^{m} \sqrt{\lambda_i}(v, v^{(i)})v^{(i)} \quad \forall v \in \mathbb{R}^m.$$

Check that

$$M^{1/2}M^{1/2}v = \sum_{i=1}^{m} \lambda_i (v, v^{(i)})v^{(i)} = Mv \quad \forall v \in \mathbb{R}^m,$$

and infer that $M^{1/2}M^{1/2} = M$, a fact that explains why we called the matrix $M^{1/2}$ square root of M.

1.19 Let $M \in \mathbb{R}^{m,m}$ be a *negative semidefinite* matrix, $(Mx, x) \leq 0$ for $x \in \mathbb{R}^m$. Consider the initial value problem

$$\begin{cases} y'(t) = My(t), & t \geq 0, \\ y(0) = y_0. \end{cases}$$

Show that the Euclidean norm $\|y(\cdot)\|$ is a decreasing function; see Exercise 1.5.

1.20 It is well known that an $m \times m$ symmetric real matrix is negative semidefinite, if and only if all its eigenvalues are non-positive. Let $M \in \mathbb{R}^{m,m}$ be an arbitrary matrix. With M^T the transpose of M, the matrix can be decomposed in a symmetric and an antisymmetric part, M_s and M_a, respectively,

$$M = M_s + M_a \quad \text{with} \quad M_s := \frac{1}{2}(M + M^T),\ M_a := \frac{1}{2}(M - M^T).$$

Since $(M_a x, x) = 0$, for $x \in \mathbb{R}^m$, we have $(Mx, x) = (M_s x, x)$, for $x \in \mathbb{R}^m$; in other words, M is negative semidefinite, if and only if M_s is negative semidefinite, i.e., if and only if all eigenvalues of M_s are non-positive.

Let $M = \left(\begin{smallmatrix} \lambda_1 & 1 \\ 0 & \lambda_2 \end{smallmatrix}\right)$. Show that M is negative semidefinite, if and only if $\lambda_1, \lambda_2 \leq 0$ and $\lambda_1\lambda_2 \geq 1/4$. Notice that λ_1 and λ_2 are the eigenvalues of M.

Notice, in particular, that the matrix $M = \left(\begin{smallmatrix} 0 & 1 \\ 0 & 0 \end{smallmatrix}\right)$ of Exercise 1.13b is *not* negative semidefinite.

[*Hint*: The eigenvalues μ_1, μ_2 of M_s are

$$\mu_{1,2} = \frac{1}{2}\big[(\lambda_1 + \lambda_2) \pm \sqrt{(\lambda_1 - \lambda_2)^2 + 1}\big]. \,]$$

1.21 Consider the initial value problem

$$\begin{cases} x'(t) = -2x(t) + y(t), & t \geq 0, \\ y'(t) = 2x(t) - 2y(t), & t \geq 0, \\ x(0) = x_0, \\ y(0) = y_0. \end{cases}$$

Show that the function $[x(\cdot)]^2 + [y(\cdot)]^2$ is decreasing.

[*Hint*: The matrix $\left(\begin{smallmatrix} -2 & 1 \\ -2 & -2 \end{smallmatrix}\right)$ is negative definite. See Exercise 1.19.]

1.22 Let $M \in \mathbb{R}^{m,m}$ be an *antisymmetric* matrix, i.e., such that $M^T = -M$, whence for its entries we have $M_{ij} = -M_{ji}$, $i, j = 1, \ldots, m$. Consider the initial value problem

$$\begin{cases} y'(t) = My(t), \quad t \geq 0, \\ y(0) = y_0. \end{cases}$$

Show that the Euclidean norm $\|y(\cdot)\|$ is a constant function, $\|y(t)\| = \|y(0)\|$, for all $t \geq 0$.

[*Hint*: Since $M^T = -M$, for $x, y \in \mathbb{R}^m$ we have $(Mx, y) = -(x, My)$. In particular, $(Mx, x) = 0$ for $x \in \mathbb{R}^m$. Take in the system of o.d.e's the inner product with $y(t)$ and utilize this property. Indeed, the property $(Mx, x) = 0$, for all $x \in \mathbb{R}^m$, characterizes the antisymmetric matrices, as we easily see using the relation $(M(x + y), x + y) = (Mx, x) + (My, y) + (Mx, y) + (x, My)$, for $x, y \in \mathbb{R}^m$.]

1.23 Let $M \in \mathbb{R}^{m,m}$ be a symmetric matrix. Consider the initial value problem

$$\begin{cases} y'(t) = \mathrm{i}My(t), \quad t \geq 0, \\ y(0) = y_0 \end{cases}$$

for a function $y : [0, \infty) \to \mathbb{C}^m$, where i is the imaginary unit. Show that the Euclidean norm $\|y(\cdot)\|$ is a constant function (the "energy" is conserved), $\|y(t)\| = \|y(0)\|$, for all $t \geq 0$.

[*Hint*: Take in the system of o.d.e's the inner product with $y(t)$ and utilize the fact that in the complex case we have

$$\frac{d}{dt}\|y(t)\|^2 = \frac{d}{dt}\big(y(t), y(t)\big) = \big(y'(t), y(t)\big) + \big(y(t), y'(t)\big) = 2\,\mathrm{Re}\,\big(y'(t), y(t)\big),$$

i.e., $\mathrm{Re}\,\big(y'(t), y(t)\big) = \frac{1}{2}\frac{d}{dt}\|y(t)\|^2$. Furthermore, $(Mz, z) \in \mathbb{R}$, for all $z \in \mathbb{C}^m$.]

2. The Euler method

In this chapter we analyze the Euler methods, both the explicit and the implicit, and some other low order schemes. These methods are simple and, from a pedagogical point of view, appropriate to introduce various important concepts, like consistency, stability properties and convergence. Later on, in chapters 3 and 4, we will generalize these concepts to more advanced methods, like Runge–Kutta and multistep schemes.

2.1 Explicit Euler method

This is the simplest numerical method for initial value problems. We consider the initial value problem (1.1) and assume that it possesses a unique solution.

Let $a = t^0 < t^1 < \cdots < t^N = b$ be a partition of the interval $[a, b]$. The numerical methods for (1.1) give usually approximations y^i to the nodal values $y(t^i), i = 0, \ldots, N$, of the solution y at the nodes of the partition t^n (time levels). Often, we use a uniform partition, that is, for $N \in \mathbb{N}$, we let $h := (b-a)/N$ be the time step and let $t^i := a + ih, i = 0, \ldots, N$. The explicit Euler method yields the approximations $y^1, \ldots, y^N$,

$$y^{n+1} = y^n + hf(t^n, y^n), \quad n = 0, \ldots, N-1, \tag{2.1}$$

with $y^0 := y_0$ given as in (1.1), in the case of a uniform partition with 'time step' h. The method is called explicit since y^{n+1} can be explicitly computed in terms of y^n; in other words, y^{n+1} is not implicitly defined as the solution of some equation.

In this case the uniformity of the partition is not essential; in the case of

non-uniform partitions (2.1) has to be modified in the form

$$y^{n+1} = y^n + (t^{n+1} - t^n) f(t^n, y^n).$$

Notice, however, that for some other methods (like multistep methods) the uniformity of the partition is essential (at least for the analysis) and the generalization to non-uniform partitions nontrivial.

2.1.1 Derivation of the method

There are various ways to construct numerical methods for initial value problems; the most important are probably numerical integration and numerical differentiation. Here we will derive the explicit Euler method in three different ways.

Numerical differentiation. We consider the differential equation at the point t^n,

$$y'(t^n) = f\big(t^n, y(t^n)\big),$$

and approximate $y'(t^n)$ by the difference quotient

$$\frac{1}{h}\big[y(t^{n+1}) - y(t^n)\big],$$

to obtain

$$\frac{1}{h}\big[y(t^{n+1}) - y(t^n)\big] \approx f\big(t^n, y(t^n)\big).$$

Replacing here the nodal values $y(t^m)$ by y^m and $\approx$ by $=$, we get

$$\frac{y^{n+1} - y^n}{h} = f(t^n, y^n),$$

i.e., (2.1).

Numerical integration. Integrating the differential equation from t^n to t^{n+1}, we obtain

$$\int_{t^n}^{t^{n+1}} y'(t)\,dt = \int_{t^n}^{t^{n+1}} f\big(t, y(t)\big)\,dt,$$

i.e.,

$$y(t^{n+1}) - y(t^n) = \int_{t^n}^{t^{n+1}} f\big(t, y(t)\big)\,dt. \tag{2.2}$$

We now approximate the integral on the right-hand side by the left rectangular rule and get

$$y(t^{n+1}) - y(t^n) \approx hf\big(t^n, y(t^n)\big).$$

As before, replacing here the nodal values $y(t^m)$ by y^m and $\approx$ by $=$, we get

$$y^{n+1} - y^n = hf(t^n, y^n),$$

i.e., (2.1).

Taylor expansion. By Taylor expanding the solution y around the point t^n, we have

$$y(t^{n+1}) = y(t^n) + hy'(t^n) + O(h^2).$$

Using here the differential equation, we obtain

$$y(t^{n+1}) = y(t^n) + hf\big(t^n, y(t^n)\big) + O(h^2),$$

whence, neglecting second order terms,

$$y(t^{n+1}) \approx y(t^n) + hf\big(t^n, y(t^n)\big).$$

As before, replacing here the nodal values $y(t^m)$ by y^m and $\approx$ by $=$, we get

$$y^{n+1} = y^n + hf(t^n, y^n),$$

i.e., (2.1).

2.1.2 Consistency

We obtain the *consistency error* E^n of the method by simply replacing the approximations in (2.1) by the corresponding values of the exact solution,

$$E^n := y(t^{n+1}) - y(t^n) - hf\big(t^n, y(t^n)\big), \tag{2.3}$$

$n = 0, \ldots, N-1$. In other words, the consistency error is the amount by which the exact solution misses satisfying the numerical scheme, i.e., misses being approximate solution. Notice that the consistency error is a quantity that is useful in the analysis of the method; it can not be really computed in practice, since the exact solution y is not known.

Remark 2.1 (Alternative definition.) The consistency error is not always defined as in (2.3) in the literature. Often, particularly in the numerical p.d.e. literature, consistency error is called the quantity $\widetilde{E}^n$,

$$h\widetilde{E}^n = y(t^{n+1}) - y(t^n) - hf\big(t^n, y(t^n)\big). \tag{2.4}$$

This definition is motivated by the fact that the Euler method can be equivalently written in the form

$$\frac{y^{n+1} - y^n}{h} = f(t^n, y^n), \quad n = 0, \ldots, N-1, \tag{2.5}$$

Of course, $h\widetilde{E}^n = E^n$. Which definition one prefers to use, is just a matter of taste. □

Remark 2.2 (Local error.) An alternative way to view the consistency error E^n is the following: Assume we start with the exact solution $y(t^n)$ at t^n (rather than with the approximation y^n, which we actually do in practice) and perform just one step with the Euler method. Denote $\tilde{y}^{n+1}$ the resulting approximation,

$$\tilde{y}^{n+1} = y(t^n) + hf\big(t^n, y(t^n)\big).$$

Then, (2.3) can be written in the form

$$E^n = y(t^{n+1}) - \tilde{y}^{n+1}, \tag{2.6}$$

i.e., the consistency error is the error after one time step, when starting with the exact value $y(t^n)$. This is the reason why the consistency error is often also referred to as *local error*. □

Although we can not really compute the consistency error, it is useful in the analysis of the method. We now proceed to investigate the asymptotic behaviour (as h decreases to zero) of the consistency error. Using the differential equation, we first write it in the form

$$E^n = y(t^{n+1}) - y(t^n) - hy'(t^n). \tag{2.7}$$

Remark 2.3 (Local error expressed in terms of y.) Before we proceed, let us note that the fact that we expressed the consistency error in (2.7) in terms of a singe function, namely y, in the sense that f does not enter in this form, is of importance. It does not only make the study of its asymptotic behaviour almost trivial, but it also implies that its behaviour is independent of the underlined equation (in the sense that it does not depend on f), provided y is smooth enough. This is possible for all multistep methods but not for high order Runge–Kutta methods; see Exercise 2.26. The *order reduction phenomenon* for high order Runge–Kutta methods is due exactly to this fact, that is that the consistency error can *not* be expressed in terms of a single function. □

Now to determine the asymptotic behaviour, it suffices to Taylor expand y around any point in $[0, T]$. For convenience, we expand around t^n, and obtain

$$E^n = \big[y(t^n) + hy'(t^n) + \frac{h^2}{2}y''(\xi^n)\big] - y(t^n) - hy'(t^n),$$

i.e.,

$$E^n = \frac{h^2}{2}y''(\xi^n) \tag{2.8}$$

with ξ^n an appropriate point in (t^n, t^{n+1}).

We thus see that E^n is of second order, if y is twice continuously differentiable and y'' does not vanish,

$$\max_{0\le n\le N-1} |E^n| \le \frac{1}{2}\|y''\|_\infty h^2, \tag{2.9}$$

with $\|\cdot\|_\infty$ the maximum norm in $[0, T]$, $\|\varphi\|_\infty := \max_{0\le t\le T} |\varphi(t)|$.

Notice also that when the solution y is a polynomial of degree at most one, then the consistency error vanishes and the initial value problem is integrated exactly by the method, in the sense that the approximations coincide with the corresponding exact nodal values, $y^n = y(t^n), n = 0, \dots, N$.

Since an estimate of the form (2.9) holds for the consistency error, and 2 is the largest exponent of h for which such an estimate is valid, we say that the consistency error is of second order. On the other hand, the (consistency)

order of the method is *one*, that is the order of the consistency error reduced by one. (If you want the consistency error to be of the order of the method, rather than the order of the method plus one, simply replace E^n by $\widetilde{E}^n$.)

A numerical method with consistency order at least one is called *consistent*.

Remark 2.4 (Consistency for systems of o.d.e's.) Let us now briefly consider the consistency of the Euler method for systems of o.d.e's. Of course, the consistency error E^n is defined as in (2.3), but it is now a vector in $\mathbb{R}^m$, for each n. Also, (2.7) is valid in this case as well. However, the analogue of (2.8) is not valid in this form, since Taylor's formula with remainder evaluated at a point is not valid for vector-valued functions; indeed for each component y_i of y we have, in general, different points $\xi^{n,i}$. Therefore, the correct analogue of (2.8) is in this case

$$E^n = \frac{h^2}{2}\begin{pmatrix} y''(\xi^{n,1}) \\ \vdots \\ y''(\xi^{n,m}) \end{pmatrix}. \tag{2.10}$$

Thus, we can easily estimate the maximum norm of the consistency error in the form

$$\max_{0\le n\le N-1} \|E^n\|_\infty \le \frac{1}{2} \max_{a\le t\le b} \|y''(t)\|_\infty h^2; \tag{2.11}$$

compare to (2.8). If we want to estimate the consistency error in another norm $\|\cdot\|$ of $\mathbb{R}^m$, we can just use the equivalence of norms in $\mathbb{R}^n$.

Alternatively, we can use the integral form of the remainder in the Taylor expansion,

$$\begin{aligned} \varphi(s) = \varphi(t^\star) + \varphi'(t^\star)(s-t^\star) + \cdots + \frac{1}{\ell!}\varphi^{(\ell)}(t^\star)(s-t^\star)^\ell \\ + \frac{1}{\ell!}\int_{t^\star}^{s} (s-t)^\ell \varphi^{(\ell+1)}(t)\,dt, \end{aligned} \tag{2.12}$$

which is valid also for vector-valued functions. Starting from (2.7) and using this expansion with $t^\star = t^n$, $s = t^{n+1}$, and $\ell = 1$, we obtain

$$E^n = \int_{t^n}^{t^{n+1}} (t^{n+1} - t) y''(t)\,dt, \tag{2.13}$$

which is an alternative form of (2.8). Therefore, for any norm $\|\cdot\|$ of $\mathbb{R}^m$, we infer that

$$\|E^n\| \le \max_{t^n \le t \le t^{n+1}} \|y''(t)\| \int_{t^n}^{t^{n+1}} (t^{n+1} - t)\, dt = \frac{h^2}{2} \max_{t^n \le t \le t^{n+1}} \|y''(t)\|,$$

whence

$$\max_{0 \le n \le N-1} \|E^n\| \le \frac{1}{2} \max_{a \le t \le b} \|y''(t)\| h^2. \tag{2.14}$$

This estimate is actually valid for more general differential equations, with smooth solution $y : [a, b] \to X$, with values in a normed linear space $\big(X, \|\cdot\|\big)$.

Let us also note that both consistency representations (2.8) and (2.10) follow also from (2.13). □

2.1.3 Stability

We first state and prove an auxiliary result, that will be often used throughout the notes in stability proofs.

Lemma 2.1 (Auxiliary result.) *Let δ be a positive number and $K, d_0, d_1, \ldots$ be non-negative numbers such that*

$$d_{i+1} \le (1 + \delta) d_i + K, \quad i = 0, 1, 2, \ldots . \tag{2.15}$$

Then, we can estimate d_n in terms of d_0, K, δ and n as follows

$$d_n \le d_0 \mathrm{e}^{n\delta} + K \frac{\mathrm{e}^{n\delta} - 1}{\delta}, \quad n = 0, 1, 2, \ldots . \tag{2.16}$$

Proof. For $n = 0$ estimate (2.16) is obviously valid. For $n \ge 1$, we have, due to (2.15),

$$d_n \le (1 + \delta)^n d_0 + K\big[1 + (1 + \delta) + (1 + \delta)^2 + \cdots + (1 + \delta)^{n-1}\big],$$

as is easily seen inductively, and we infer that

$$d_n \le (1 + \delta)^n d_0 + K \frac{(1 + \delta)^n - 1}{\delta},$$

which leads to the desired estimate (2.16), since obviously $e^{\delta} \geq 1 + \delta$. □

Now, let $y^n, z^n, 0 \leq n \leq N$, be given, respectively, by

$$\begin{cases} y^{n+1} = y^n + hf(t^n, y^n), \quad 0 \leq n \leq N-1, \\ y^0 = y_0, \end{cases} \tag{2.17}$$

and

$$\begin{cases} z^{n+1} = z^n + hf(t^n, z^n), \quad 0 \leq n \leq N-1, \\ z^0 = z_0, \end{cases} \tag{2.18}$$

for given starting approximations $y_0, z_0 \in \mathbb{R}$. In other words, y^n and z^n are (explicit) Euler approximations to the solution of the equation $y' = f(t, y)$, corresponding to different initial values y_0 and z_0. This method (and, more generally, all single step methods, i.e., methods in which y^{n+1} is given in terms of the previous approximation y^n and data, only) is called *stable* (for a certain class of equations), if there is a constant C depending on f but independent of h and of $y^n, z^n, 0 \leq n \leq N$, such that

$$\max_{1 \leq n \leq N} |y^n - z^n| \leq C|y_0 - z_0|. \tag{2.19}$$

For simplicity, we considered here a uniform partition $t^n = a + nh, 0 \leq n \leq N, Nh = b - a$, of the interval $[a, b]$; the generalization to non-uniform partitions is obvious.

The stability of a method is a property of fundamental importance; in practice, due to rounding errors, for instance, or to the fact that nonlinear equations can not be solved exactly, in general, we do not compute the exact approximations y^n but rather some perturbations $\tilde{y}^n$. Stable methods are not too sensitive to errors of this kind.

We assume now that f satisfies the global Lipschitz condition (1.6) and will prove stability of the Euler method. Subtracting (2.18) from (2.17), we get

$$y^{n+1} - z^{n+1} = y^n - z^n + h\big[f(t^n, y^n) - f(t^n, z^n)\big].$$

Using now the triangle inequality and the Lipschitz condition (1.6), we obtain

$$|y^{n+1} - z^{n+1}| \leq |y^n - z^n| + hL|y^n - z^n|, \quad n = 0, \ldots, N-1,$$

whence, inductively,

$$|y^n - z^n| \le (1+hL)^n |y^0 - z^0| \le \mathrm{e}^{hLn}|y^0 - z^0|, \quad n = 1, \ldots, N. \tag{2.20}$$

Thus, we infer that

$$|y^n - z^n| \le \mathrm{e}^{L(t^n - a)}|y_0 - z_0|, \quad n = 0, \ldots, N, \tag{2.21}$$

and

$$\max_{0 \le n \le N} |y^n - z^n| \le \mathrm{e}^{L(b-a)}|y_0 - z_0|, \tag{2.22}$$

which are the discrete analogues of the stability estimates (1.11) and (1.12), respectively, for the continuous problem. Notice that in this proof we essentially used Lemma 2.1, with $K = 0$.

Notice also that it is straightforward to generalize the stability proof of the Euler method for systems of o.d.e's, in a norm $\|\cdot\|$, assuming f satisfies Lipschitz condition (1.20).

B–stability. The Euler method can be easily applied to initial value problems for systems of o.d.e's; the formula (2.1) remains the same, the only difference being that now the approximation y^n are vectors, $y^n \in \mathbb{R}^m$.

A natural question now is whether it mimics the behaviour of the exact solutions, in case f satisfies the one-sided Lipschitz condition (1.24); cf. Exercise 1.5. Since this is a very important property, it deserves a definition:

Definition 2.1 (B–stability.) A single-step method is called *B–stable*, if, when applied to an appropriate class of test problems, namely initial value problems (1.19) with right-hand side f satisfying the one-sided Lipschitz condition (1.24), for any given starting approximations y^0 and z^0, it yields approximations y^n and z^n, respectively, such that $\|y^n - z^n\|, n = 0, \ldots, N$, is a decreasing sequence,

$$\|y^{n+1} - z^{n+1}\| \le \|y^n - z^n\|, \quad n = 0, \ldots, N, \tag{2.23}$$

with $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^m$. □

As we will see, the explicit Euler method is *not* B–stable. Before that, we slightly relax the stability requirement and introduce the so-called *A–stability* property. Since the A–stability property concerns the behaviour of numerical approximations for *linear* differential equations with constant coefficients, we do not need to consider two sequences of approximation; one sequence of approximations for the corresponding homogeneous equation suffices.

Definition 2.2 (A–stability.) A single-step method is called *A–stable*, if, when applied to the scalar test problem (1.18), with a complex coefficient λ with non-positive real part, $\mathrm{Re}\,\lambda \leq 0$, it yields approximations y^n such that $|y^n|$, $n = 0, \ldots,$ is a decreasing sequence,

$$|y^{n+1}| \leq |y^n|, \quad n = 0, \ldots . \tag{2.24}$$

□

Notice that the solution of the test problem (1.18) is $y(t) = \mathrm{e}^{\lambda t}$ and that its absolute value $|y(t)| = \mathrm{e}^{(\mathrm{Re}\,\lambda)t}$ is a decreasing function. In other words, the A–stability property requires that the numerical approximations mimic the behaviour of the exact solution of test problem (1.18). Notice also that the initial value y_0 (in this case $y_0 = 1$) is irrelevant for the A–stability (provided $y_0 \neq 0$).

All single step methods used in practice yield the same approximations when applied to test problems (1.18) or (1.26). Therefore, it is obvious that the B–stability implies A–stability,

$$\text{B–stability} \implies \text{A–stability}.$$

Actually, although the A–stability property is defined to mimic the behaviour of the solutions of the scalar test problem (1.18), it does more: A–stable schemes mimic the behaviour of the solutions of systems of linear o.d.e's with constant coefficients, $y' = My + g(t)$, with a negative semidefinite matrix M, that is, such that the one-sided Lipschitz condition is satisfied; see Remark 2.7 in the sequel.

A natural question is whether these two stability properties are equivalent. The answer is *negative*, as we will see in Remark 2.6.

Let us now return to the explicit Euler method. We have already mentioned that it is not B–stable. Actually we will show that it is not even A–stable. (Indeed, all *explicit* methods used in practice share this property with the explicit Euler method: they are not A–stable.)

Indeed, applying the explicit Euler method to (1.18), we obtain $y^{n+1} = y^n + h\lambda y^n$, i.e.,

$$y^{n+1} = r(h\lambda)y^n, \quad \text{with} \quad r(z) := 1 + z, \tag{2.25}$$

whence

$$|y^{n+1}| = |1 + h\lambda|\,|y^n|. \tag{2.26}$$

Obviously, if the complex number $h\lambda$ in the non-positive complex half-plane (since we assumed that $\mathrm{Re}\,\lambda \leq 0$) is outside the disk of radius one centered at the point -1, then $|1 + h\lambda| > 1$, whence, for non-vanishing y^n,

$$|y^{n+1}| > |y^n|.$$

In particular, the explicit Euler method is not A–stable; thus it in not B–stable, either.

As we just saw, the explicit Euler method is not A–stable. However, the numerical approximations y^n do mimic the behaviour of the exact solution of (1.18), at least in some cases; more precisely, when $|1 + h\lambda| \leq 1$, in other words, when $h\lambda$ is in the disk of radius one centered at the point -1. This information is useful and leads us to the following definition of the *stability region*:

Definition 2.3 (Stability region.) Consider the scalar test problem (1.18), with an arbitrary but fixed complex coefficient λ, and discretize it by a single-step method, with a fixed time step h. The *stability region* S of the method consists of all points $z = h\lambda$ in the complex plane $\mathbb{C}$, with the property that the method yields approximations y^n such that $|y^n|, n = 0, \ldots,$ is a decreasing sequence,

$$|y^{n+1}| \leq |y^n|, \quad n = 0, \ldots . \tag{2.27}$$

The intersection of S with the real axis is called *stability "interval"* of the method.

Let $\vartheta \in (0, \pi/2)$. The method is called $A(\vartheta)$*–stable*, if the sector S_ϑ,

$$S_\vartheta := \{z \in \mathbb{C} : z = -\rho e^{i\varphi},\ \rho \geq 0,\ |\varphi| \leq \vartheta\},$$

is contained in its stability region S. □

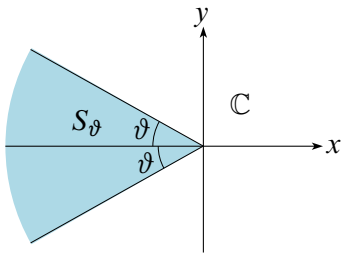

Figure 2.1: The sector S_ϑ in the complex plane.

It is obvious from Definitions 2.2 and 2.3 that a method is A–stable, if and only if the non-positive complex half-plane $\mathbb{C}^-$,

$$\mathbb{C}^- := \{z \in \mathbb{C} : \mathrm{Re}\, z \leq 0\},$$

is contained in its stability region S.

Notice that the stability region S of the explicit Euler method is the unit disc centered at -1,

$$S := \{z \in \mathbb{C} : |1 + z| \leq 1\}; \tag{2.28}$$

cf. (2.26).

Remark 2.5 (A comment on (2.25).) Let us look at the consistency error E^n of the explicit Euler method, in the special case that it is applied to (1.18), with a fixed λ. According to (2.25), the consistency error can be written in the form

$$E^n = y(t^{n+1}) - r(\lambda h)y(t^n).$$

Now, the solution is $y(t) = e^{\lambda t}$, whence $y(t^{n+1}) = e^{\lambda h} y(t^n)$, and thus

$$E^n = \left[e^{\lambda h} - r(\lambda h)\right] y(t^n).$$

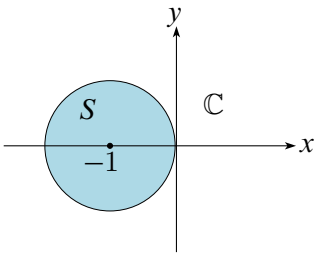

Figure 2.2: The stability region S of the explicit Euler method.

Since $\mathrm{e}^{\lambda h} = 1 + \lambda h + O(h^2)$, we infer that

$$\mathrm{e}^{\lambda h} - r(\lambda h) = O(h^2),$$

which is natural because the consistency error is of order two. In other words, since the consistency error is of order two, the first two terms of $r(z)$ (which are actually the only terms in this case) have to be 1 and z. The converse is, in general, not true; from a relation of the form

$$\mathrm{e}^z - r(z) = O(z^{p+1}), \quad \text{as } z \to 0,$$

we can not conclude that the order of the method is p. We will discuss this point in detail in chapter 3. □

Remark 2.6 (The trapezoidal method is A–stable but not B–stable.) With the standard notation, the time step $y^n \mapsto y^{n+1}$ of the *trapezoidal method* for the initial value problem (1.1) is

$$y^{n+1} = y^n + \frac{h}{2}\big[f(t^n, y^n) + f(t^{n+1}, y^{n+1})\big], \quad n = 0, \ldots, N-1. \tag{2.29}$$

This is an implicit second order method, often used in practice. Before we proceed, let us comment on the derivation of the method and explain why it is called trapezoidal method. Our starting point is (2.2). We approximate the integral on the right-hand side by the trapezoidal rule and get

$$y(t^{n+1}) - y(t^n) \approx \frac{h}{2}\big[f\big(t^n, y(t^n)\big) + f\big(t^{n+1}, y(t^{n+1})\big)\big].$$

Replacing here the nodal values $y(t^m)$ by y^m and $\approx$ by $=$, we are led to (2.29).

First claim: The trapezoidal method is A–stable.

Indeed, applying the method to the test problem (1.18), we obtain

$$y^{n+1} = y^n + \frac{h}{2}(\lambda y^n + \lambda y^{n+1}),$$

whence

$$y^{n+1} = r(h\lambda)y^n, \quad \text{with} \quad r(z) := \frac{1+\frac{z}{2}}{1-\frac{z}{2}}. \tag{2.30}$$

Thus we infer that the stability region S of the trapezoidal method is

$$S = \{z \in \mathbb{C} : |z+2| \le |z-2|\} = \mathbb{C}^-.$$

In particular, the trapezoidal method is indeed A–stable.

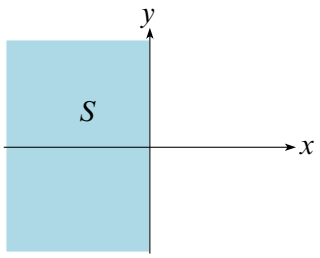

Figure 2.3: The stability region $S = \mathbb{C}^-$ of the trapezoidal method.

Second claim: The trapezoidal method is *not* B–stable.

Indeed, applying the trapezoidal method to the test problem (1.15), with a continuous function λ, we have

$$y^{n+1} = y^n + \frac{h}{2}\big[\lambda(t^n)y^n + \lambda(t^{n+1})y^{n+1}\big].$$

Notice that we again applied the method to a linear scalar o.d.e.; the only difference to (1.18) is that in this case we allow a variable coefficient. Assume now that λ takes on only non-positive values (in which case the one-sided Lipschitz condition (1.13) is satisfied). Fix a time step h and let, for instance, λ be such that $h\lambda(t^n) = -8$ and $h\lambda(t^{n+1}) = -1$. Then, we have

$$y^{n+1} = y^n + \frac{1}{2}\big(-8y^n - y^{n+1}\big),$$

whence $y^{n+1} = -2y^n$, an thus $|y^{n+1}| = 2|y^n|$, which proves our claim.

See also Exercise 2.22 for a more general result of this form for a one parameter family of schemes. □

Remark 2.7 (A von Neumann theorem and an application.) Let $\big(H, (\cdot,\cdot)\big)$ be a complex Hilbert space. We denote by $\|\cdot\|$ both the norm associated to the inner product, $\|v\| := (v,v)^{1/2}$, and the induced operator norm, for bounded linear operators $A : H \to H$,

$$\|A\| := \sup_{\substack{v\in H\\ v\neq 0}} \frac{\|Av\|}{\|v\|}.$$

Then, according to a von Neumann theorem, for any linear operator $A : H \to H$ and any rational function r, we have

$$\big[\,\mathrm{Re}(Av,v) \le 0 \;\forall v \in H\,\big] \implies \|r(A)\| \le \sup_{\mathrm{Re}\, z\le 0} |r(z)|; \tag{2.31}$$

in other words, if the linear operator $A : H \to H$ satisfies the condition $\mathrm{Re}(Av,v) \le 0$, for all $v \in H$, then, for any rational function r, the norm $\|r(A)\|$ does not exceed $\sup_{\mathrm{Re}\, z\le 0} |r(z)|$. Of course, the result is particularly useful in case the rational function r is uniformly bounded in the negative complex half-plane.

The result is also valid in the case of real Hilbert spaces. Now the assumption is $(Av,v) \le 0$, for all $v \in H$, i.e., that the linear operator A is negative semidefinite, and the estimate reads as before,

$$\big[(Av,v) \le 0 \;\forall v \in H\big] \implies \|r(A)\| \le \sup_{\mathrm{Re}\, z\le 0} |r(z)|. \tag{2.32}$$

Let us now apply this Theorem to A–stable single-step schemes. Consider a linear system of o.d.e's with constant coefficients, $y' = My$, and assume that M is negative semidefinite; see (1.25). Then, the time step $y^n \mapsto y^{n+1}$ of the single-step schemes studied in this chapter, can be expressed in the form

$$y^{n+1} = r(hM)y^n, \tag{2.33}$$

with r a rational function; see (2.25), (2.30), (2.56), as well as ($\star$) in Exercise 2.22, which includes all these methods as special cases. Indeed, (2.33) is

valid for all Runge–Kutta methods, which will be the subject of chapter 3; see (3.57) and (3.58). From relation (2.33), we immediately obtain

$$\|y^{n+1}\| \leq \|r(hM)\|\,\|y^n\|. \tag{2.34}$$

Now, since M is negative semidefinite and h is positive, the matrix hM is also negative semidefinite. Combining (2.32) with the A–stability of the method, we infer

$$\|r(hM)\| \leq \sup_{\mathrm{Re}\, z \leq 0} |r(z)| \leq 1,$$

and (2.34) yields

$$\|y^{n+1}\| \leq \|y^n\|. \tag{2.35}$$

The desired property (2.23) follows now from the fact that the difference $y^m - z^m$ of two sequences of approximations for the solution of the inhomogeneous system of o.d.e.'s $y' = My + g(t)$ satisfies the relation

$$y^{n+1} - z^{n+1} = r(hM)(y^n - z^n)$$

(see, for instance, Exercise 2.24) and from the results stated above.

For an elementary proof of (2.35), avoiding the use of the von Neumann theorem, in the case of a symmetric matrix M we refer to Exercises 3.20 and 3.19. □

Remark 2.8 (Equivalent requirement for A–stability.) As we already mentioned, when single-step methods are applied to the test problem (1.18), then they can be written in the form

$$y^{n+1} = r(h\lambda)y^n;$$

see (2.33). Therefore, we obviously have, $y^n = \big(r(h\lambda)\big)^n y^0$, whence

$$|y^n| = |r(h\lambda)|^n\,|y^0|, \quad n = 0, \ldots, N.$$

We can thus easily see that condition (2.24) in the Definition 2.2 of the A–stability, is actually equivalent to the seemingly less stringest requirement that the approximations $(y^n)_{n\in\mathbb{N}}$ are *bounded*, when the method is applied to test problem (1.18). □

2.1.4 Convergence, error estimates

In this paragraph we will estimate the (global) error

$$\varepsilon^n := y(t^n) - y^n, \quad n = 0, \ldots, N, \tag{2.36}$$

of the explicit Euler approximations. In particular, we will establish convergence of the approximations. This will be done by combining the stability and the consistency of the method.

Since the method is not B–stable, it is not suitable for problems with right-hand side satisfying the one-sided Lipschitz condition. Therefore, we will here only consider the case that f satisfies the global Lipschitz condition (1.6).

Theorem 2.1 (Error estimate.) *Let $f \in C\big([a,b]\times\mathbb{R}\big)$ be a function satisfying the Lipschitz condition* (1.6)*, and assume that the solution y of* (1.1) *is twice continuously differentiable, $y \in C^2[a,b]$. If $y^0, \ldots, y^N$ are the explicit Euler approximations for a uniform partition of the interval $[a,b]$ with time step $h = (b-a)/N$, given by* (2.1)*, then the estimate*

$$\max_{0\le n\le N} \big|y(t^n) - y^n\big| \le \frac{M}{2L}\big(\mathrm{e}^{L(b-a)} - 1\big)h \tag{2.37}$$

is valid, with

$$M := \|y''\|_\infty = \max_{a\le t\le b} \big|y''(t)\big|$$

and L the Lipschitz constant in (1.6)*.*

Proof. Subtracting equation (2.1) from the consistency error equation (2.3), we arrive at the *error equation*

$$\varepsilon^{n+1} = \varepsilon^n + h\big[f\big(t^n, y(t^n)\big) - f\,(t^n, y^n)\big] + E^n. \tag{2.38}$$

Using now the Lipschitz condition, we obtain

$$|\varepsilon^{n+1}| \le |\varepsilon^n| + hL|\varepsilon^n| + |E^n|,$$

whence

$$|\varepsilon^{n+1}| \le (1+hL)|\varepsilon^n| + \max_{0\le m\le N-1} |E^m|, \quad n = 0, \ldots, N-1. \tag{2.39}$$

Utilizing Lemma 2.1, with obvious notation, we infer from (2.39)

$$|\varepsilon^n| \le \mathrm{e}^{nhL}|\varepsilon^0| + \frac{\mathrm{e}^{nhL}-1}{hL} \max_{0\le m\le N-1} |E^m|, \quad n = 0, \dots, N,$$

whence, since $\varepsilon^0 = 0$ and $nh \le b - a$,

$$|\varepsilon^n| \le \frac{\mathrm{e}^{L(b-a)}-1}{hL} \max_{0\le m\le N-1} |E^m|, \quad n = 0, \dots, N.$$

Finally, using in this relation the consistency estimate (2.9), we obtain the desired error estimate (2.37). □

In Theorem 2.1 we established and estimate of the error of the explicit Euler method of the form

$$\max_{0\le n\le N} \left|y(t^n) - y^n\right| \le Ch, \quad h = \frac{b-a}{N}, \tag{2.40}$$

with a constant C, independent of h, depending on the data of our problem, namely a, b, y_0 and f (and of the solution y, which is uniquely determined by the data). We notice that the bound in (2.40) is the product of a constant times the time step h to the power one. Therefore, we say that the *order* of the method is (at least) one.

It is easily seen that the order is exactly one, that is, in general, we can not replace Ch by an expression that decays faster as h decreases to zero, even if we assume that y is as smooth as we like. (Of course, there are problems for which this is possible, for instance, the problem

$$\begin{cases} y' = 1, \quad 0 \le t \le 1, \\ y(0) = 0, \end{cases}$$

is integrated exactly by the method, since the solution is $y(t) = t$, whence $y'' = 0$, and thus $M = 0$ in (2.37); therefore, $\varepsilon^n = 0, n = 0, \dots, N$.)

To prove our claim, we consider the initial value problem

$$\begin{cases} y' = 2t, \quad 0 \le t \le 1, \\ y(0) = 0. \end{cases} \tag{2.41}$$

The unique solution of (2.41) is $y(t) = t^2, 0 \leq t \leq 1$. This choice is motivated by the fact that y'' is a constant function; in particular, the consistency error E^n does not vary with n (see (2.8)), whence no cancellation of the consistency errors is possible, and this allows us to easily observe the behaviour of the (global) error. Let $N \in \mathbb{N}, h := 1/N, t^n := nh, n = 0, \ldots, N$, and $y^0, \ldots, y^N$ be the explicit Euler approximations for problem (2.41). Applying (2.1) to (2.41), we obtain

$$y^{n+1} = y^n + 2ht^n,$$

i.e.,

$$y^{n+1} = y^n + 2nh^2, \quad n = 0, \ldots, N-1.$$

We thus inductively see that

$$y^n = y^0 + 2\big[1 + 2 + \cdots + (n-1)\big]h^2 = y^0 + n(n-1)h^2, \quad n = 0, \ldots, N,$$

whence, since $y^0 := 0$,

$$y^n = n(n-1)h^2, \quad n = 0, \ldots, N.$$

For $n = N$ we get $y^N = N(N-1)h^2$, whence, since $Nh = 1$, $y^N = 1 - h$. Therefore, the error at the final point $t = t^N = 1$ is

$$y(1) - y^N = h,$$

i.e., of order exactly one, and this proves our claim.

We infer that the order of the explicit Euler method is exactly one. This is true, provided the solution y is sufficiently smooth; cf. Theorem 2.1. If the solution is only once continuously differentiable, then we can still prove convergence, namely an estimate with right-hand side converging to zero as h decreases to zero, but the bound is not of first order any more. For this result and for an estimate in case y' is Hölder continuous, we refer to Exercise 2.2.

The estimate (2.37) is a typical *a priori* error estimate for a method for initial value problems. It is useful because it gives us information about the required stability properties of the method, the decay rate of the error, the

regularity needed, and the highest possible order of the method. On the other hand, if we really want to compute the error bound for a certain value of the step-size h, we can not do it, because the constant on the right-hand side of (2.37) depends on the unknown exact solution; *a posteriori* error estimates yield computable error bounds (but we will not discuss such estimates in these notes).

Remark 2.9 (Approximate solution in $[a,b]$.) By the Euler method, as well as by many methods we will discuss in the sequel, we only determine approximations y^n of the nodal values $y(t^n)$ of the solution y at the nodes t^n. Some people do not find this natural, namely that we approximate a function by a sequence $y^0,\dots,y^N$. If we wish, however, we can define approximate solutions as functions in $[a,b]$, for instance by interpolating the nodal values by appropriate splines. What appropriate is, depends on our purposes. In the a priori error analysis, for instance, it completely suffices to define an approximate solution $\tilde{y}$, such that the error in the supremum norm,

$$\sup_{a\le t\le b}|y(t)-\tilde{y}(t)|,$$

is of the same order as in the discrete maximum norm,

$$\max_{0\le n\le N}|y(t^n)-y^n|.$$

For instance, in the case of the Euler method, we can choose either $\tilde{y}$ piecewise constant, $\tilde{y}(t)=y^n, t\in[t^n,t^{n+1})$, or piecewise linear, by linearly interpolating between the nodal approximations y^n and y^{n+1},

$$\tilde{y}(t)=y^n+\frac{y^{n+1}-y^n}{t^{n+1}-t^n}(t-t^n),\quad t\in[t^n,t^{n+1}).$$

Without getting into the details, let us note that in the a posteriori theory, only the second choice is appropriate. □

The error estimate in Theorem 2.1 can be easily generalized to the case of initial value problems for systems of first order o.d.e's. Let $\|\cdot\|$ be a norm in $\mathbb{R}^m$. We established two consistency estimates for the case of systems; see

(2.11) and (2.14). Depending on which of the two we will use, we will derive two error estimates; the first is expressed in terms of a constant C_1 such that

$$\forall x \in \mathbb{R}^m \quad \|x\| \leq C_1 \|x\|_\infty. \tag{2.42}$$

The analogue to Theorem 2.1 in the case of systems is:

Theorem 2.2 (Error estimates for systems of o.d.e's.) *Let $f : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$ be a function satisfying the Lipschitz condition* (1.10)*. Assume $y = (y_1, \ldots, y_m)^T, y_i \in C^2[a,b], i = 1, \ldots, m$, is the solution of initial value problem* (1.9)*. If $y^0, \ldots, y^N$ are the approximations to $y(t^n)$, given by the explicit Euler method using a uniform partition of $[a,b]$ with step-size $h = (b-a)/N$, then there holds*

$$\max_{0 \leq n \leq N} \|y(t^n) - y^n\| \leq \frac{M}{2L} C_1 \left[\mathrm{e}^{L(b-a)} - 1 \right] h, \tag{2.43}$$

with $M := \max_{a \leq t \leq b} \|y''(t)\|_\infty$. Furthermore,

$$\max_{0 \leq n \leq N} \|y(t^n) - y^n\| \leq \frac{\tilde{M}}{2L} \left[\mathrm{e}^{L(b-a)} - 1 \right] h, \tag{2.44}$$

with $\tilde{M} := \max_{a \leq t \leq b} \|y''(t)\|$. □

The proof of Theorem 2.2 is a straightforward generalization of the proof of Theorem 2.1, and is left as an Exercise (Exercise 2.5).

Remark 2.10 (Advantages and drawbacks.) All numerical methods for initial value problems used in practice have both advantages and drawbacks. This is the reason why there are so many of them. They may be suitable for a certain class of equations but not suitable for another class of equations. Here we summarize advantages and drawbacks of the explicit Euler method. The method is explicit, very easy to program, and it requires only one evaluation of the right-hand side f per time level. On the other hand, the order of the method is very low, just one, and if we want to compute accurate approximations we have to use very small time steps; this increases the total computational cost.

Also in practice, the numerical approximations y^n suffer from roundoff errors, particularly if we compute with very small time step. Furthermore, the method is not B–stable, whence it is not appropriate for equations satisfying the one-sided Lipschitz condition; actually, its stability region is very small. □

2.2 Implicit Euler method

In this section we will analyze the *implicit Euler method*, also referred to as *backward Euler method*.

With standard notation, the method yields approximations y^m to $y(t^m)$ defined by

$$y^{n+1} = y^n + hf(t^{n+1}, y^{n+1}), \quad n = 0, \ldots, N-1, \tag{2.45}$$

with starting value $y^0 := y_0$.

In (2.45) the unknown y^{n+1} is implicitly defined as the solution of an equation, and this is why the method is called implicit. This is a computational drawback, compared to the explicit Euler method; however, as we will see later on, the implicit Euler method has more advantageous stability properties, more precisely it is B–stable.

The implicit Euler method can be derived in a completely analogue manner as the explicit one. For instance, using numerical differentiation, we consider the differential equation at the point t^{n+1},

$$y'(t^{n+1}) = f\big(t^{n+1}, y(t^{n+1})\big),$$

and approximate $y'(t^{n+1})$ by the backward difference quotient

$$\frac{1}{h}\big[y(t^{n+1}) - y(t^n)\big],$$

to obtain

$$\frac{1}{h}\big[y(t^{n+1}) - y(t^n)\big] \approx f\big(t^{n+1}, y(t^{n+1})\big).$$

Replacing here the nodal values $y(t^m)$ by y^m and $\approx$ by $=$, we get

$$\frac{y^{n+1} - y^n}{h} = f(t^{n+1}, y^{n+1}),$$

i.e., (2.45).

2.2.1 Existence and uniqueness of the approximations

Since the method is implicit, existence and uniqueness of the approximations defined by (2.45) is not obvious. Actually, without any conditions on f and/or h, existence and uniqueness is not guaranteed. For instance, in the linear case $f(t,y) = \lambda y$, with a real constant λ, (2.45) reads $y^{n+1} = y^n + h\lambda y^{n+1}$, whence

$$(1-\lambda h)y^{n+1} = y^n. \tag{2.46}$$

Assume now that λ is positive and $h = 1/\lambda$. Then, in case $y^n \neq 0$ equation (2.46) does not have a solution, while in case $y^n = 0$ any real number y^{n+1} is a solution of (2.46).

We will next consider two cases, when f satisfies the global Lipschitz condition and the one-sided Lipschitz condition, respectively.

Fitst case: f satisfies the global Lipschitz condition. We consider the function g,

$$g(x) := y^n + hf(t^{n+1}, x), \quad x \in \mathbb{R}.$$

Obviously, any y^{n+1} is a solution of equation (2.45), if and only if it is a fixed point of g. Thus, it suffices to find conditions under which g has exactly one fixed point. In view of the Lipschitz condition, we have

$$\big|g(x_1) - g(x_2)\big| \leq hL|x_1 - x_2|, \quad x_1, x_2 \in \mathbb{R},$$

and see easily that, for h sufficiently small such that $hL < 1$, the function g is a contraction in $\mathbb{R}$, and we infer that it possesses exactly one fixed point, namely the solution y^{n+1} of (2.45).

This proof can be immediately extended to the case of systems of first order o.d.e's.

Second case: f satisfies the one-sided Lipschitz condition. We consider the function g,

$$g(x) := x - y^n - hf(t^{n+1}, x), \quad x \in \mathbb{R}.$$

Obviously, any y^{n+1} is a solution of equation (2.45), if and only if it is a root of g. First, since $f(t^{n+1}, \cdot)$ is decreasing, the function g is obviously strictly

increasing; thus, it can have at most one solution. This shows uniqueness of the approximations. To show existence, we note that, again since $f(t^{n+1},\cdot)$ is decreasing, for $x \le 0$, obviously $g(x) \le x - y^n - hf(t^{n+1},0)$, and infer that $g(x)$ tends to $-\infty$ as x tends to $-\infty$. Furthermore, for $x \ge 0$, we have $g(x) \ge x - y^n - hf(t^{n+1},0)$, whence $g(x)$ tends to ∞ as x tends to ∞. In particular, g takes on both negative and positive values. Utilizing the continuity of g and the intermediate value theorem, we infer that g possesses at least one root, y^{n+1}. We have thus shown existence and uniqueness of the approximate solutions, without any restriction on the step-size h.

The proof we just presented can not be generalized to the case of systems of o.d.e's. We refer to Exercise 2.14 for an alternative proof that can be easily extended to the case of systems of first order o.d.e's; see Exercises 2.15 and 2.16.

2.2.2 Consistency

The *consistency error* E^n of the implicit Euler method is given by

$$E^n := y(t^{n+1}) - y(t^n) - hf\big(t^{n+1}, y(t^{n+1})\big), \tag{2.47}$$

$n = 0, \ldots, N-1$.

Using in (2.47) the o.d.e. (1.1), we can express E^n in terms of y only,

$$E^n = y(t^{n+1}) - y(t^n) - hy'(t^{n+1}). \tag{2.48}$$

Thus, the consistency error can be estimated in a completely analogue way to the explicit Euler case; for instance, it is easily seen that the analogue of (2.13) reads as follows

$$E^n = -\int_{t^n}^{t^{n+1}} (t - t^n) y''(t)\, dt. \tag{2.49}$$

We infer that the implicit Euler method is a first order method, i.e., its order is one.

2.2.3 Stability

The implicit Euler method has very good stability properties; it is B–stable and thus also A–stable.

Now, let $y^n, z^n, 0 \le n \le N$, be given, respectively, by

$$\begin{cases} y^{n+1} = y^n + hf(t^{n+1}, y^{n+1}), \quad 0 \le n \le N-1, \\ y^0 = y_0, \end{cases} \tag{2.50}$$

and

$$\begin{cases} z^{n+1} = z^n + hf(t^{n+1}, z^{n+1}), \quad 0 \le n \le N-1, \\ z^0 = z_0, \end{cases} \tag{2.51}$$

for given starting approximations $y_0, z_0 \in \mathbb{R}$. In other words, y^n and z^n are implicit Euler approximations to the solution of the equation $y' = f(t, y)$, corresponding to different initial values y_0 and z_0. For simplicity, we considered here a uniform partition $t^n = a + nh, 0 \le n \le N, Nh = b - a$, of the interval $[a, b]$; the generalization to non-uniform partitions is obvious.

We assume now that f satisfies the global Lipschitz condition (1.6) and will prove stability of the Euler method. Subtracting (2.51) from (2.50), we get

$$y^{n+1} - z^{n+1} = y^n - z^n + h\big[f(t^{n+1}, y^{n+1}) - f(t^{n+1}, z^{n+1})\big]. \tag{2.52}$$

Using now the triangle inequality and the Lipschitz condition (1.6), we obtain

$$|y^{n+1} - z^{n+1}| \le |y^n - z^n| + hL|y^{n+1} - z^{n+1}|, \quad n = 0, \ldots, N-1,$$

whence, for h sufficiently small, such that $Lh \le 1/2$, say,

$$|y^{n+1} - z^{n+1}| \le \frac{1}{1 - Lh}|y^n - z^n|, \quad n = 0, \ldots, N-1.$$

Now, for $Lh \le 1/2$, we have

$$\frac{1}{1 - Lh} \le 1 + 2Lh,$$

and infer that

$$|y^{n+1} - z^{n+1}| \le (1 + 2Lh)|y^n - z^n|, \quad n = 0, \ldots, N-1,$$

whence, inductively,

$$|y^n - z^n| \le (1 + 2hL)^n |y^0 - z^0| \le \mathrm{e}^{2hLn}|y^0 - z^0|, \quad n = 1, \ldots, N. \tag{2.53}$$

Thus, we get

$$|y^n - z^n| \le \mathrm{e}^{2L(t^n - a)}|y_0 - z_0|, \quad n = 0, \ldots, N, \tag{2.54}$$

and

$$\max_{0 \le n \le N} |y^n - z^n| \le \mathrm{e}^{2L(b-a)}|y_0 - z_0|, \tag{2.55}$$

i.e., stability of the implicit Euler method.

Notice also that it is straightforward to generalize the stability proof of the implicit Euler method for systems of o.d.e's, in a norm $\|\cdot\|$, assuming f satisfies Lipschitz condition (1.20).

Let us now show that the method is B–stable; in particular this shows stability of the method for initial value problem (1.19) with right-hand side f satisfying the one-sided Lipschitz condition (1.24). We start from (2.52). Taking there the Euclidean inner product with $y^{n+1} - z^{n+1}$ and using the one-sided Lipschitz condition (1.24), we obtain

$$\|y^{n+1} - z^{n+1}\|^2 \le (y^n - z^n, y^{n+1} - z^{n+1}),$$

whence, in view of the Cauchy–Schwarz inequality,

$$\|y^{n+1} - z^{n+1}\|^2 \le \|y^n - z^n\| \, \|y^{n+1} - z^{n+1}\|,$$

and the desired estimate $\|y^{n+1} - z^{n+1}\| \le \|y^n - z^n\|, n = 0, \ldots, N$, follows.

Since the B–stability property is stronger than the A–stability property, we infer that the implicit Euler method is also A–stable. Let us, however, show directly this property and at the same time determine the stability region of

the method. Applying the method to the test problem (1.18), we have $y^{n+1} = y^n + h\lambda y^{n+1}$, i.e.,

$$y^{n+1} = r(h\lambda)y^n, \quad \text{with} \quad r(z) := \frac{1}{1-z}. \tag{2.56}$$

We infer that the stability region S of the method is

$$S = \{z \in \mathbb{C} : |z-1| \geq 1\},$$

that is the exterior of the open unit disc in the complex plane centered at 1. In particular, the method is indeed A–stable.

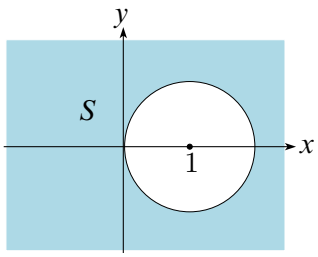

Figure 2.4: The stability region S of the implicit Euler method.

2.2.4 Convergence, error estimates

In this section we will estimate the error

$$\varepsilon^n := y(t^n) - y^n, \quad n = 0, \ldots, N, \tag{2.57}$$

of the implicit Euler approximations. In particular, we will establish convergence of the approximations. This will be done by combining the stability and the consistency of the method.

Since the estimate in case f satisfies the Lipschitz condition is easily derived by combining stability and consistency, an in the case of the explicit Euler method, we shall consider here only the case that f satisfies the one-sided Lipschitz condition (1.13).

Subtracting (2.45) from (2.47), we obtain the error equation

$$\varepsilon^{n+1} = \varepsilon^n + h\Big[f\big(t^{n+1}, y(t^{n+1})\big) - f\big(t^{n+1}, y^{n+1}\big)\Big] + E^n.$$

Multiplying here by ε^{n+1} and using (1.13), we get,

$$(\varepsilon^{n+1})^2 \le \varepsilon^n \varepsilon^{n+1} + E^n \varepsilon^{n+1} \quad \text{or} \quad |\varepsilon^{n+1}| \le |\varepsilon^n| + \max_{0\le m\le N} |E^m|,$$

and infer easily that

$$|\varepsilon^n| \le n \max_{0\le m\le N} |E^m|.$$

Utilizing here the consistency estimate (2.49), we obtain the desired error estimate

$$\max_{0\le n\le N} |\varepsilon^n| \le \frac{b-a}{2} \max_{a\le t\le b} |y''(t)|h. \tag{2.58}$$

We emphasize that the important fact in this estimate is that the Lipschitz constant L of f does not enter; we have not even assumed that f satisfies the Lipschitz condition. Our assumption is that f satisfies the one-sided Lipschitz condition (1.13).

Exercises

2.1 Let c, C be positive constants, $N \in \mathbb{N}$, $h := c/N$ and numbers $\varepsilon^0, \dots, \varepsilon^N$, with $\varepsilon^0 \ne 0$.

a) If

$$(\star) \qquad |\varepsilon^{n+1}| \le (1 + Ch)|\varepsilon^n|, \quad n = 0, \dots, N-1,$$

show that there exists a constant $\widetilde{C}$, *independent of h and N*, such that

$$\max_{0\le n\le N} |\varepsilon^n| \le \widetilde{C}|\varepsilon^0|.$$

b) If instead of $(\star)$ we assume that

$$|\varepsilon^{n+1}| \le (1 + Ch^p)|\varepsilon^n|, \quad n = 0, \dots, N-1,$$

with $p > 1$, show that the conclusion in a) is again valid.

c) If instead of $(\star)$ we assume that

$$|\varepsilon^{n+1}| \leq (1 + Ch^q)|\varepsilon^n|, \quad n = 0, \ldots, N-1,$$

with $q < 1$, show that the conclusion in a) is, in general, *not* valid.

[*Hint*: Choose $\varepsilon^n := (1 + Ch^q)^n \varepsilon^0, n = 1, \ldots, N$, and show that

$$(1 + Ch^q)^N \geq 1 + NCh^q = 1 + Cc\frac{1}{h^{1-q}} \to \infty, \quad h \to 0.\,]$$

d) If instead of $(\star)$ we assume that

$$|\varepsilon^{n+1}| \leq (\lambda + Ch)|\varepsilon^n|, \quad n = 0, \ldots, N-1,$$

with $\lambda > 1$, show that the conclusion in a) is, in general, *not* valid.

[*Hint*: Choose $\varepsilon^n := \lambda^n \varepsilon^0, n = 1, \ldots, N$.]

Comment. inequalities of the form $(\star)$ play a key role in proving stability of numerical methods, as we already saw in section 2.1.3 and as we will see several times in the sequel. The aim of this Exercise it to give some insight into this kind of estimates.

2.2 Let $f \in C\big([a,b] \times \mathbb{R}\big)$ be a function satisfying the global Lipschitz condition (1.6). We assume that the solution y of the initial value problem (1.1) is continuously differentiable. If $y^0, \ldots, y^N$ are the explicit Euler approximations, for a uniform partition of the interval $[a,b]$, with time step $h := (b-a)/N$, show that

$$\max_{0 \leq n \leq N} \big|y(t^n) - y^n\big| \to 0, \quad N \to \infty,$$

i.e., that the method is still convergent. Assuming, in addition, that y' satisfies in $[a,b]$ the *Hölder condition* with exponent $\alpha \in (0,1]$ (why do we restrict ourselves to the interval $(0,1]$? which functions satisfy this condition with exponent $\alpha > 1$?), i.e.,

$$(\star) \qquad \forall x, \tilde{x} \in [a,b] \quad \big|y'(x) - y'(\tilde{x})\big| \leq \lambda |x - \tilde{x}|^\alpha,$$

with some constant $\lambda \in \mathbb{R}$, show that

$$\max_{0 \leq n \leq N} \big|y(t^n) - y^n\big| \leq \frac{\lambda}{(1+\alpha)L}\big(\mathrm{e}^{L(b-a)} - 1\big)h^\alpha.$$

In case y' satisfies the Lipschitz condition, that is $(\star)$ with $\alpha = 1$, compare the result to the one of Theorem 2.1.

[*Hint*: Let E^n,

$$E^n := y(t^{n+1}) - y(t^n) - hf\big(t^n, y(t^n)\big) = y(t^{n+1}) - y(t^n) - hy'(t^n),$$

be the consistency error of the method, and

$$E_h := \frac{1}{h} \max_{0 \le n \le N-1} |E^n|.$$

Show that the estimate corresponding to (2.39) is now

$$|\varepsilon^{n+1}| \le (1 + hL)|\varepsilon^n| + hE_h,$$

and infer that

$$\big|y(t^n) - y^n\big| \le \frac{\mathrm{e}^{nhL} - 1}{L} E_h, \quad n = 0, \ldots, N.$$

Finally,

$$E^n = \int_{t^n}^{t^{n+1}} \big[y'(t) - y'(t^n)\big]\, dt,$$

whence

$$E_h \le \max_{\substack{x, \tilde{x} \in [a,b] \\ |x - \tilde{x}| \le h}} \big|y'(x) - y'(\tilde{x})\big|.$$

If y' is continuous, then it is also uniformly continuous, whence $E_h \to 0$, as $h \to 0$. If y' satisfies ($\star$), then, obviously, we have $E_h \le \lambda h^\alpha$. Notice also that $|E^n| \le \lambda h^{1+\alpha}/(1+\alpha)$, whence $E_h \le \lambda h^\alpha/(1+\alpha)$, and infer that the desired error estimate is indeed valid.]

2.3 Let $a = t^0 < t^1 < \cdots < t^N = b$ be a partition of $[a, b]$. Assume that the partition is *quasi-uniform*, in the sense that there exists a constant μ, independent of N, such that

$$\min_{0 \le n \le N-1} h_n \ge \mu \max_{0 \le n \le N-1} h_n,$$

with $h_n := t^{n+1} - t^n$, $n = 0, \ldots, N-1$. Let

$$h := \max_{0 \le n \le N-1} h_n,$$

and formulate and prove a result analogue to Theorem 2.1 for this case.

2.4 Let $a = t^0 < t^1 < \cdots < t^N = b$ be an arbitrary partition of $[a, b]$, $h_n := t^{n+1} - t^n, n = 0, \ldots, N-1$, and

$$h := \max_{0 \le n \le N-1} h_n.$$

Formulate and prove a result analogue to Theorem 2.1 for this case.
[*Hint*: First show, as in (2.39), that

$$|\varepsilon^{n+1}| \le (1 + h_n L)|\varepsilon^n| + \frac{Mh}{2} h_n, \quad n = 0, \ldots, N-1.$$

Utilizing now the fact that $\varepsilon^0 = 0$, we have

$$\begin{aligned}
|\varepsilon^{n+1}| &\le \frac{Mh}{2}\big[h_n + (1 + h_n L)h_{n-1} + (1 + h_n L)(1 + h_{n-1}L)h_{n-2} \\
&\qquad + \cdots + (1 + h_n L)(1 + h_{n-1}L)\cdots(1 + h_1 L)h_0\big] \\
&\le \frac{Mh}{2}\Big[h_n + \mathrm{e}^{L(t^{n+1}-t^n)}(t^n - t^{n-1}) + \mathrm{e}^{L(t^{n+1}-t^{n-1})}(t^{n-1} - t^{n-2}) \\
&\qquad + \cdots + \mathrm{e}^{L(t^{n+1}-t^1)}(t^1 - t^0)\Big] \\
&\le \frac{Mh}{2}\mathrm{e}^{L(b-a)}\big(h_n + h_{n-1} + \cdots + h_0\big) = \frac{Mh}{2}\mathrm{e}^{L(b-a)}(b-a).
\end{aligned}$$

Show the inequality

$$\mathrm{e}^{L(t^{n+1}-t^k)}(t^k - t^{k-1}) \le \int_{t^{k-1}}^{t^k} \mathrm{e}^{L(t^{n+1}-s)}\, ds$$

and use it to improve the constant in the previous estimate.]

2.5 Prove Theorem 2.2.

2.6 Assume problem (1.1) possesses a unique solution $y \in C^2[a, b]$, and let $m, \ell \in \mathbb{R}$ be such that

$$\forall t \in [a, b] \quad m \le y(t) \le \ell,$$

and $\delta > 0$. If instead of (1.6) we assume the *local* Lipschitz condition, that is

$$\exists L \ge 0 \ \ \forall t \in [a, b] \ \ \forall y_1, y_2 \in [m - \delta, \ell + \delta] \quad \big|f(t, y_1) - f(t, y_2)\big| \le L|y_1 - y_2|$$

(a very useful and realistic condition), show that there exists a positive h_0 such that, for $h \in (0, h_0]$, the explicit Euler method for (1.1), with time-step h (uniform partition), yields approximations for which the estimate (2.37) holds true.

[*Hint*: Assume that

$$\frac{M}{2L}[\mathrm{e}^{L(b-a)}-1]h_0<\delta,$$

with

$$M:=\max_{a\le t\le b}|y''(t)|,$$

and show inductively that $y^k\in[m-\delta,\ell+\delta], k=0,\dots,N.$]

2.7 (The midpoint method.) With standard notation the step $y^n\mapsto y^{n+1}$ of the midpoint method is

$$y^{n+1}=y^n+hf(t^{n+\frac12},y^{n+\frac12}),$$

with

$$t^{n+\frac12}:=\frac12(t^n+t^{n+1})=t^n+\frac12 h\quad\text{and}\quad y^{n+\frac12}:=\frac12(y^n+y^{n+1}).$$

First, let us comment on the derivation of the method. Our starting point is (2.2). We approximate the integral on the right-hand side by the midpoint rule and get

$$y(t^{n+1})-y(t^n)\approx hf\big(t^{n+\frac12},y(t^{n+\frac12})\big).$$

Show that

$$y(t^{n+\frac12})=\frac12\big(y(t^n)+y(t^{n+1})\big)+O(h^2).$$

Thus, we have

$$y(t^{n+1})-y(t^n)\approx hf\big(t^{n+\frac12},\frac12\big(y(t^n)+y(t^{n+1})\big)\big).$$

Replacing here the nodal values $y(t^m)$ by y^m and $\approx$ by $=$, we are led to the midpoint method.

Show that the midpoint and trapezoidal methods coincide, when applied to homogeneous systems of linear o.d.e's with constant coefficients, $y'=My$, with a square matrix M.

Prove that the midpoint method is B–stable (while the trapezoidal method is not).

2.8 Let $M\in\mathbb{R}^{m,m}$ be an antisymmetric, invertible matrix, $y_0\in\mathbb{R}^m$, $y_0\neq 0$, and consider the initial value problem

$$\begin{cases}y'=My, & t\ge 0,\\ y(0)=y_0.\end{cases}$$

Then, the Euclidean norm $\|y(\cdot)\|$ is conserved, i.e., it is a constant function; see Exercise 1.22.

a) Consider the approximations $y^n, n \geq 0$, produced by the explicit Euler method, with a fixed time step $h > 0$. Show that $\|y^n\| \to \infty$, as $n \to \infty$.

[*Hint*: We have $y^{n+1} = y^n + hMy^n$, whence

$$\|y^{n+1}\|^2 = (y^n + hMy^n, y^n + hMy^n) = \|y^n\|^2 + h^2\|My^n\|^2,$$

and $\|Mx\| \geq c\|x\|$, for all $x \in \mathbb{R}^m$, with a positive constant c. Therefore $\|y^{n+1}\|^2 \geq (1 + c^2h^2)\|y^n\|^2$.]

b) Consider the approximations $y^n, n \geq 0$, produced by the implicit Euler method, with a fixed time step $h > 0$. Show that $\|y^n\| \to 0$, as $n \to \infty$.

[*Hint*: We have $y^{n+1} = y^n + hMy^{n+1}$, whence

$$\|y^n\|^2 = (y^{n+1} - hMy^{n+1}, y^{n+1} - hMy^{n+1}) = \|y^{n+1}\|^2 + h^2\|My^{n+1}\|^2,$$

whence $\|y^n\|^2 \geq (1 + c^2h^2)\|y^{n+1}\|^2$.]

c) Consider the approximations $y^n, n \geq 0$, produced by midpoint method (which coincides with the trapezoidal method in this case), with a fixed time step $h > 0$. Show that the method mimics the continuous problem, in the sense that $\|y^n\|$ is conserved, $\|y^n\| = \|y_0\|, n \in \mathbb{N}$.

[*Hint*: We have $y^{n+1} = y^n + \frac{h}{2}M(y^n + y^{n+1})$, whence

$$(y^{n+1} - y^n, y^{n+1} + y^n) = \frac{h}{2}\big(M(y^n + y^{n+1}), y^{n+1} + y^n\big) = 0,$$

i.e., $\|y^{n+1}\|^2 = \|y^n\|^2$.]

2.9 We discretize the initial value problem of Exercise 1.23 by the midpoint method (or equivalently by the trapezoidal method) with step size h. Show that the approximations conserve the Euclidean norm, $\|y^n\| = \|y_0\|, n \geq 0$.

2.10 Consider the initial value problem (1.1) and assume that the continuous function f satisfies (1.13). We discretize the problem by the midpoint rule, using, without loss of generality, a uniform partition of the interval $[a, b]$. Show that the approximations are well defined.

[*Hint*: For a given y^n, show that $x^\star = \frac{1}{2}(y^n + y^{n+1})$ is well defined.]

2.11 (Stability on the imaginary axis.) Consider the initial value problem

$$\begin{cases} y' = \mathrm{i}\mu y, \quad t \geq 0, \\ y(0) = 1, \end{cases}$$

with μ a non-vanishing real constant and $y : [0, \infty) \to \mathbb{C}$. Obviously, $y(t) = \mathrm{e}^{\mathrm{i}\mu t}$, and $\big|y(t)\big| = 1$, for all $t \geq 0$. For this problem consider:

a) The explicit Euler method.

b) The implicit Euler method.

c) The midpoint method.

Show that only the midpoint method mimics the behaviour of the continuous problem, in the sense that it yields approximations y^n such that $|y^n| = 1, n \in \mathbb{N}$.

2.12 Consider the initial value problem

$$\begin{cases} y' = \mathrm{i}\mu(t)y, \quad t \geq 0, \\ y(0) = 1, \end{cases}$$

with μ a real-valued function and $y : [0, \infty) \to \mathbb{C}$. Show that $\big|y(t)\big| = 1$, for all $t \geq 0$. For this problem consider:

a) The midpoint method.

b) The trapezoidal method.

Show that the midpoint method mimics the behaviour of the continuous problem, in the sense that it yields approximations y^n such that $|y^n| = 1, n \in \mathbb{N}$, while the trapezoidal method does not, in general. For instance, for a fixed h and μ such that $h\mu(t^n) = 4$ and $h\mu(t^{n+1}) = 2$, we have for the trapezoidal method

$$|y^{n+1}| = \sqrt{\frac{5}{2}}\,|y^n|.$$

Notice, however, that these methods coincide when μ is constant; cf. Exercise 2.11.

2.13 According to the Brouwer *fixed point theorem*, if K is a non-empty, *convex*, *closed* and *bounded* subset of $\mathbb{R}^m$ and $f : K \to K$ a continuous function, then f has at least one fixed point in K, that is there exists $x^\star \in K$ such that $f(x^\star) = x^\star$. (The simplest such case is when $m = 1$ and $K = [a, b]$, in which case the theorem reduces to the intermediate value theorem.) This result is useful in existence proofs of numerical approximations.

Use Brouwer's fixed point theorem to show the following particularly useful version of it: Let $g : \mathbb{R}^m \to \mathbb{R}^m$ be a continuous function such that $(g(x), x) \geq 0$, for all $x \in \mathbb{R}^m$ with $\|x\| = \alpha$, for a positive number α. Then, there exists $x^\star \in \mathbb{R}^m$ such

that $g(x^\star) = 0$ and $\|x^\star\| \le \alpha$. Here $(\cdot,\cdot)$ and $\|\cdot\|$ denote the Euclidean inner product and the Euclidean norm in $\mathbb{R}^m$, respectively.
[*Hint*: The set $K := \{x \in \mathbb{R}^m : \|x\| \le \alpha\}$ is obviously non-empty, convex, closed and bounded. If $g(x) \neq 0$ for all $x \in K$, then the function $f : K \to K$,

$$f(x) = -\alpha \frac{g(x)}{\|g(x)\|},$$

would be continuous and, according to Brouwer's fixed point theorem, would have a fixed point $x^\star$ στο K. This leads to the contradiction that $\|x^\star\| = \alpha$ and $\|x^\star\|^2 = (f(x^\star), x^\star) \le 0$.]

2.14 Give an existence and uniqueness proof of the approximations defined by the implicit Euler method for initial value problems with f satisfying the one-sided Lipschitz condition (1.13), for all h, which can be generalized to systems of first order o.d.e's; cf. Exercises 2.15 and 2.16.
[*Hint*: Consider the continuous function $g, g(x) := x - y^n - hf(t^{n+1}, x), x \in \mathbb{R}$. For $x \in \mathbb{R}$, we have

$$\begin{aligned} g(x) \cdot x &= x^2 - y^n \cdot x - hf(t^{n+1}, x) \cdot x \\ &= x^2 - y^n \cdot x - h\big[f(t^{n+1}, x) - f(t^{n+1}, 0)\big] \cdot x - hf(t^{n+1}, 0) \cdot x, \end{aligned}$$

whence, in view of (1.13), we get

$$g(x) \cdot x \ge x^2 - \big[y^n + hf(t^{n+1}, 0)\big] \cdot x.$$

Since $\big[y^n + hf(t^{n+1}, 0)\big] \cdot x \le \frac{1}{2}\big(x^2 + \big[y^n + hf(t^{n+1}, 0)\big]^2\big)$, we infer that

$$g(x) \cdot x \ge \frac{1}{2}\big(|x|^2 - |y^n + hf(t^{n+1}, 0)|^2\big) \quad \forall x \in \mathbb{R}.$$

For $x \in \mathbb{R}$, $|x| = |y^n + hf(t^{n+1}, 0)| + 1$, we thus have $g(x) \cdot x > 0$. We conclude that g has different signs at the boundary points of the interval

$$\big[-|y^n + hf(t^{n+1}, 0)| - 1, |y^n + hf(t^{n+1}, 0)| + 1\big],$$

and, therefore, it possesses at least one solution in this interval. As far as uniqueness is concerned, assuming that $x, y \in \mathbb{R}$ are roots of g, we have

$$g(x) - g(y) = 0$$

or

$$x - y = h\big[f(t^{n+1}, x) - f(t^{n+1}, y)\big].$$

Multiplying here by $x - y$ and using (1.13) we obtain $(x - y)^2 \le 0$, whence $x = y$.]

2.15 Consider an initial value problem

$$\begin{cases} y'(t) = f(t, y), \quad t \in [a, b], \\ y(a) = y_0, \end{cases}$$

with $f : [a, b] \times \mathbb{R}^m \to \mathbb{R}^m$ a continuous function satisfying the one-sided Lipschitz condition (1.24). Prove that the approximation of the implicit Euler method, in the case of a uniform partition of $[a, b]$ with time step h, are well defined.
[*Hint*: Consider the function $g : \mathbb{R}^m \to \mathbb{R}^m$, $g(x) := x - y^n - hf(t^{n+1}, x)$. Then,

$$\begin{aligned} \big(g(x), x\big) &= \|x\|^2 - (y^n + hf(t^{n+1}, 0), x) - h(f(t^{n+1}, x) - f(t^{n+1}, 0), x) \\ &\geq \|x\|^2 - (y^n + hf(t^{n+1}, 0), x) \\ &\geq \frac{1}{2}\big[\|x\|^2 - \|y^n + hf(t^{n+1}, 0)\|^2\big]. \end{aligned}$$

For $x \in \mathbb{R}^m$ such that $\|x\| = \|y^n + hf(t^{n+1}, 0)\| + 1$ the term on the right-hand side is obviously positive. Use now Exercise 2.13 to show existence of y^{n+1}. The uniqueness proof is very easy.]

2.16 With the notation of Exercise 2.15, we now assume that the function $f : [a, b] \times \mathbb{R}^m \to \mathbb{R}^m$ is continuous and satisfies the slightly more condition

$$\forall t \in [a, b] \quad \forall y_1, y_2 \in \mathbb{R}^m \quad \big(f(t, y_1) - f(t, y_2), y_1 - y_2\big) \leq \nu \|y_1 - y_2\|^2,$$

with a positive constant ν. Prove that the approximation of the implicit Euler method, in the case of a uniform partition of $[a, b]$ with time step h, are well defined, provided the product νh is sufficiently small.

2.17 Besides Brouwer's fixed point theorem, the following result, often referred to as Zarantonello's fixed point theorem, plays an important role in existence and uniqueness proofs of numerical approximations: Let $G : \mathbb{R}^m \to \mathbb{R}^m$ be a *strongly monotone* mapping, i.e., such that

$$\forall v, w \in \mathbb{R}^m \quad \big(G(v) - G(w), v - w\big) \geq c\|v - w\|^2,$$

with a positive constant c, which satisfies also the Lipschitz condition,

$$\forall v, w \in \mathbb{R}^m \quad \|G(v) - G(w)\| \leq L\|v - w\|.$$

Show that G vanishes at exactly one point.

[*Hint*: It suffices to show that the mapping

$$F : \mathbb{R}^m \to \mathbb{R}^m, \quad F(v) := v - \frac{c}{L^2} G(v),$$

has exactly one fixed point. First notice that

$$\|F(v) - F(w)\|^2 = \|v - w\|^2 - 2\frac{c}{L^2}\big(G(v) - G(w), v - w\big) + \frac{c^2}{L^4}\|G(v) - G(w)\|^2$$

and infer that

$$\|F(v) - F(w)\| \leq \sqrt{1 - \frac{c^2}{L^2}}\|v - w\|,$$

i.e., that F is a contraction.]

2.18 With the notation of Exercise 2.15, we now assume that the function $f : [a, b] \times \mathbb{R}^m \to \mathbb{R}^m$ satisfies the condition

$$\forall t \in [a, b] \;\; \forall y_1, y_2 \in \mathbb{R}^m \quad \big(f(t, y_1) - f(t, y_2), y_1 - y_2\big) \leq \nu \|y_1 - y_2\|^2,$$

with a positive constant ν, as well as the Lipschitz condition

$$\forall t \in [a, b] \;\; \forall y_1, y_2 \in \mathbb{R}^m \quad \|f(t, y_1) - f(t, y_2)\| \leq L\|y_1 - y_2\|.$$

Use Exercise 2.17 to prove that the approximation of the implicit Euler method, in the case of a uniform partition of $[a, b]$ with time step h, are well defined, provided the product νh is sufficiently small.

[Notice that we have assumed the Lipschitz condition here, but the Lipschitz constant does not enter in the restriction on the time step h. At a first glance to Exercises 2.13 and 2.17 we could say that it is advantageous to use the Brouwer's rather than Zarantonello's fixed point theorem, since in the second case the Lipschitz condition is required. However, the advantage of Zarantonello's fixed point theorem consists in the fact that it can be also used to establish existence of approximate solutions in infinite dimensional Banach spaces as well, as is the case for the contraction mapping theorem, in contrast to Brouwer's fixed point theorem, in which the space has to be finite dimensional, like the spaces $\mathbb{R}^m$.]

2.19 (Discrete Gronwall inequality, corresponding to the differential form: uniform partition.) Let $T > 0, N \in \mathbb{N}$ and $h := \frac{T}{N}$. Av $\alpha_0, \ldots, \alpha_N$ and $\varepsilon_0, \ldots, \varepsilon_{N-1}$ be non-negative numbers such that, with $\gamma > 0$,

$$\alpha_{n+1} \leq (1 + \gamma h)\alpha_n + h\varepsilon_n, \quad n = 0, \ldots, N - 1.$$

Show that

$$\max_{1 \leq n \leq N} \alpha_n \leq \mathrm{e}^{\gamma T}\alpha_0 + \frac{1}{\gamma}\mathrm{e}^{\gamma T} \max_{0 \leq n \leq N-1} \varepsilon_n.$$

2.20 (Discrete Gronwall inequality, corresponding to the differential form: non-uniform partition.) Let $T > 0, 0 = t^0 < t^1 < \cdots < t^N = T$ be a partition of $[0,T], h_n := t^{n+1} - t^n, n = 0,\dots,N-1$. If $\alpha_0,\dots,\alpha_N$ and $\varepsilon_0,\dots,\varepsilon_{N-1}$ are non-negative numbers such that, with $\gamma > 0$.

$$\alpha_{n+1} \le (1+\gamma h_n)\alpha_n + h_n\varepsilon_n, \quad n = 0,\dots,N-1,$$

prove that

$$\max_{1\le n\le N} \alpha_n \le \mathrm{e}^{\gamma T}\alpha_0 + \int_0^T \mathrm{e}^{\gamma(T-s)}\,ds \max_{0\le n\le N-1} \varepsilon_n;$$

cf. Exercise 2.4.

2.21 (Discrete Gronwall inequality, corresponding to the integral form.) Let $T > 0, N \in \mathbb{N}$ and $h := \frac{T}{N}$. If $\alpha_0,\dots,\alpha_N$ and E are non-negative numbers such that, with $\gamma > 0$,

$$\alpha_{n+1} \le E + \gamma h \sum_{\ell=0}^{n} \alpha_\ell, \quad n = 0,\dots,N-1,$$

prove that

$$\max_{1\le n\le N} \alpha_n \le C(h\alpha_0 + E)$$

with a constant C, independent of h.
[*Hint*: Let $\varphi_n := \gamma h \sum_{\ell=0}^{n} \alpha_\ell$. Then $\varphi_{n+1} - \varphi_n = \gamma h \alpha_{n+1}$, whence

$$\varphi_{n+1} \le \gamma h E + (1+\gamma h)\varphi_n, \quad n = 0,\dots,N-1.$$

Use now Exercise 2.19 to estimate the φ_n's, and, subsequently, use

$$\alpha_{n+1} \le E + \varphi_n$$

to estimate the α_n's.]

2.22 (The "theta-method": stability properties.) Let $\vartheta \in [0,1]$. With the usual notation, consider the so-called *theta-method* for problem (1.1),

$$y^{n+1} = y^n + (1-\vartheta)hf(t^n,y^n) + \vartheta hf(t^{n+1},y^{n+1}), \quad n = 0,\dots,N-1,$$

with $y^0 = y_0$. Which quadrature rule leads to this method? Which methods do we get for $\vartheta = 0$, $\vartheta = 1/2$, and $\vartheta = 1$, respectively? Show that applying the method to the test problem (1.18), we obtain

$$(\star) \qquad y^{n+1} = r(h\lambda)y^n, \quad \text{with} \quad r(z) := \frac{1+(1-\vartheta)z}{1-\vartheta z},$$

whence the stability region S of the method is

$$S = \{z \in \mathbb{C} : |(1-\vartheta)z + 1| \le |\vartheta z - 1|\}.$$

Let now $\vartheta \in (0,1)$. If $z = x + \mathrm{i}y$, with $x, y \in \mathbb{R}$, it is easily seen that z belongs to S, if and only if

$$(\star\star) \qquad (1-2\vartheta)(x^2+y^2) + 2x \le 0.$$

We already know that $S = \mathbb{C}^-$ for $\vartheta = 1/2$. For $\vartheta < 1/2$ and $\vartheta > 1/2$, relation $(\star\star)$ takes the form

$$\big(x + \frac{1}{1-2\vartheta}\big)^2 + y^2 \le \frac{1}{(1-2\vartheta)^2} \quad \text{and} \quad \big(x + \frac{1}{1-2\vartheta}\big)^2 + y^2 \ge \frac{1}{(1-2\vartheta)^2},$$

respectively. Therefore, the stability region S is the disc in the first case and the exterior of the open disc in the second case, respectively, centered at $\big(-\frac{1}{1-2\vartheta}, 0\big)$ with radius $1/|1-2\vartheta|$. In particular, we infer that the method is A–stable, if and only if $\vartheta \ge 1/2$. Notice that the origin is a boundary point of the disc mentioned above. Also, the radius of the disc is one for $\vartheta = 0$, it is an increasing function of ϑ in the interval $(0, 1/2)$, and tends to infinity as ϑ approaches $1/2$. The center of the disc moves on the negative real axis from -1 to $-\infty$, as ϑ increases from 0 to $1/2$. Then, for $\vartheta \in (1/2, 1]$, the radius decreases from ∞ to 1. The center of the disc moves on the positive real axis from ∞ to 1 as ϑ increases from $1/2$ to 1. See Figures 2.2–2.4 for the stability regions in the cases $\vartheta = 0$, $\vartheta = 1/2$, and $\vartheta = 1$, respectively.

Prove also that the method is B–stable, if and only if $\vartheta = 1$.

[*Hint*: The method is obviously B–stable for $\vartheta = 1$, since it reduces to the implicit Euler method. Also, it is not B–stable for $\vartheta = 0$, since it it is not even A–stable for $\vartheta < 1/2$. For $\vartheta \in (0,1)$, we follow the idea of Remark 2.6. With the notation used there, and for a fixed h, we let the function λ be such that $(1-\vartheta)h\lambda(t^n) = -4$ and $\vartheta h\lambda(t^{n+1}) = -\frac{1}{2}$, and infer that $y^{n+1} = -2y^n$, whence $|y^{n+1}| = 2|y^n|$, which proves that the method is not B–stable.]

2.23 (The "theta-method": consistency properties.) Let $\vartheta \in [0,1]$, and, with the usual notation, E^n denote the consistency error of the theta-method,

$$E^n = y(t^{n+1}) - y(t^n) - (1-\vartheta)hf\big(t^n, y(t^n)\big) - \vartheta hf\big(t^{n+1}, y(t^{n+1})\big).$$

Then,

$$E^n = y(t^{n+1}) - y(t^n) - (1-\vartheta)hy'(t^n) - \vartheta hy'(t^{n+1}),$$

whence

$$E^n = (1-\vartheta)\big[y(t^{n+1}) - y(t^n) - hy'(t^n)\big] + \vartheta\big[y(t^{n+1}) - y(t^n) - hy'(t^{n+1})\big].$$

Notice that the first and the second term, respectively, in brackets on the right-hand side, is the consistency error of the explicit and the implicit Euler method, respectively. In other words, the consistency error of the theta-method is a convex combination of the consistency errors of the explicit and implicit Euler methods. Thus, according to (2.13) and (2.49), we have

$$E^n = -(1-\vartheta)\int_{t^n}^{t^{n+1}} (t-t^{n+1})y''(t)\,dt - \vartheta\int_{t^n}^{t^{n+1}} (t-t^n)y''(t)\,dt,$$

whence

$$E^n = -\int_{t^n}^{t^{n+1}} \big(t-(1-\vartheta)t^{n+1} - \vartheta t^n\big)y''(t)\,dt,$$

i.e., with $t^{n+\frac{1}{2}} := \frac{1}{2}(t^n + t^{n+1})$ the midpoint of the interval $[t^n, t^{n+1}]$,

$$(\star) \qquad E^n = -\int_{t^n}^{t^{n+1}} \Big[t - t^{n+\frac{1}{2}} + \big(\vartheta - \frac{1}{2}\big)h\Big]y''(t)\,dt.$$

Show now that the method is of first order for $\vartheta \neq 1/2$ and of second order for $\vartheta = 1/2$.

[*Hint*: Obviously, the order of the consistency error E^n is at least two. In the case $\vartheta \neq 1/2$, show that the order of E^n is exactly two, if y'' is a non-vanishing constant function.

In the case $\vartheta = 1/2$, i.e., in the case of the trapezoidal method, from $(\star)$, we obtain

$$\begin{aligned} E^n &= -\frac{1}{2}\int_{t^n}^{t^{n+1}} \big[\big(t-t^{n+\frac{1}{2}}\big)^2\big]' y''(t)\,dt \\ &= -\frac{1}{8}h^2\big[y''(t^{n+1}) - y''(t^n)\big] + \frac{1}{2}\int_{t^n}^{t^{n+1}} \big(t-t^{n+\frac{1}{2}}\big)^2 y'''(t)\,dt \\ &= \frac{1}{2}\int_{t^n}^{t^{n+1}} \Big[\big(t-t^{n+\frac{1}{2}}\big)^2 - \frac{1}{8}h^2\Big]y'''(t)\,dt \end{aligned}$$

and infer that

$$(\star\star) \qquad E^n = \frac{1}{2}\int_{t^n}^{t^{n+1}} \big(t-t^{n+1}\big)\big(t-t^n\big)y'''(t)\,dt.$$

Alternatively, we can obtain ($\star\star$) by subtracting the Taylor expansion around t^n,

$$y(t^{n+\frac{1}{2}}) = y(t^n) + \frac{1}{2}hy'(t^n) + \frac{1}{8}h^2y''(t^n) + \frac{1}{2}\int_{t^n}^{t^{n+\frac{1}{2}}} \big(t^{n+\frac{1}{2}} - t\big)^2 y'''(t)\,dt,$$

from the corresponding expansion around t^{n+1}.]

2.24 (The "theta-method" for inhomogeneous linear equations.) With the standard notation, show that the time step $y^n \mapsto y^{n+1}$ of the theta-method applied to the inhomogeneous linear equation $y' = \lambda y + f(t)$ is

$$y^{n+1} = \frac{1 + (1-\vartheta)\lambda h}{1 - \vartheta\lambda h} y^n + \frac{1-\vartheta}{1-\vartheta\lambda h} hf(t^n) + \frac{\vartheta}{1-\vartheta\lambda h} hf(t^{n+1}).$$

2.25 (The midpoint method: consistency.) The consistency error E^n of the midpoint method (see Exercise 2.7) is

$$E^n := y(t^{n+1}) - y(t^n) - hf\big(t^{n+\frac{1}{2}}, \frac{1}{2}\big(y(t^n) + y(t^{n+1})\big)\big).$$

For reasons that will become clear in the sequence, we split E^n in the form $E^n = E_1^n + E_2^n$, with

$$\begin{cases} E_1^n := y(t^{n+1}) - y(t^n) - hy'(t^{n+\frac{1}{2}}), \\ E_2^n := h\Big[f\big(t^{n+\frac{1}{2}}, y(t^{n+\frac{1}{2}})\big) - f\big(t^{n+\frac{1}{2}}, \frac{1}{2}\big(y(t^n) + y(t^{n+1})\big)\big)\Big]. \end{cases}$$

Notice that E_1^n is expressed in terms of y only. By Taylor expanding around $t^{n+\frac{1}{2}}$, we easily see that

$$E_1^n = \frac{1}{2}\int_{t^n}^{t^{n+\frac{1}{2}}} (t^n - t)^2 y'''(t)\,dt + \frac{1}{2}\int_{t^{n+\frac{1}{2}}}^{t^{n+1}} (t^{n+1} - t)^2 y'''(t)\,dt.$$

We thus see that E_1^n is of order 3.

Furthermore, again by Taylor expanding around $t^{n+\frac{1}{2}}$, we can write

$$\widetilde{E}_2^n := y(t^{n+\frac{1}{2}}) - \frac{1}{2}\big(y(t^n) + y(t^{n+1})\big)$$

in the form

$$(\star) \qquad \widetilde{E}_2^n = -\frac{1}{2}\int_{t^n}^{t^{n+\frac{1}{2}}} (t - t^n) y''(t)\,dt - \frac{1}{2}\int_{t^{n+\frac{1}{2}}}^{t^{n+1}} (t^{n+1} - t) y''(t)\,dt.$$

Therefore, if f is locally Lipschitz in y, we have

$$\|E_2^n\| \leq Lh\|\widetilde{E}_2^n\|,$$

and infer easily that E_2^n is also of order 3.

Notice for later use that if $y''(t)$ is constant, then $(\star)$ yields

$$(\star\star) \qquad \widetilde{E}_2^n = -\frac{h^2}{8}y''(t);$$

in particular, in contrast to E_1^n, the quantity $\widetilde{E}_2^n$ does not vanish, if y is a polynomial of degree two.

2.26 (Polynomial order.) The *polynomial order* $\tilde{p}$ of a numerical method for initial value problems is the largest integer such that the method integrates initial value problems with solution $y \in \mathbb{P}_{\tilde{p}}$, i.e., polynomial of degree at most $\tilde{p}$, exactly; that is, the approximations y^n coincide with the exact nodal values $y(t^n)$. This is the case, if and only if the consistency error E^n of the method vanishes, whenever $y \in \mathbb{P}_{\tilde{p}}$.

Show that $\tilde{p} = 1$ for both Euler methods, the explicit and the implicit, $\tilde{p} = 2$ for the trapezoidal method, and $\tilde{p} = 1$ for the midpoint method. More generally, $\tilde{p} = 1$ for the theta-method, when $\vartheta \neq 1/2$, and $\tilde{p} = 2$, when $\vartheta = 1/2$.

[*Hint*: For the theta-method the result follows from the representation $(\star)$ of the consistency error in Exercise 2.23, in case $\vartheta \neq 1/2$, and from the Hint in the same Exercise, in case $\vartheta = 1/2$.

For the midpoint method the result follows from Exercise 2.25; see, in particular, the representation $(\star\star)$ in Exercise 2.25, that leads easily to $\tilde{p} < 2$.]

Comment. The polynomial order plays a key role in the analysis of the methods, when they are applied to evolution p.d.e's; it is sometimes referred to as *order of strict accuracy*. Notice that the polynomial order $\tilde{p} = 1$ of the midpoint method (which is a Runge–Kutta method but not a multistep method) is strictly less than its order $p = 2$. In contrast, for all other methods considered here (which are also multistep methods) the polynomial order $\tilde{p}$ coincides with the order of the method. This it typical for multistep methods and for Runge–Kutta methods: the polynomial order of every multistep method coincides with its order, while the polynomial order of a Runge–Kutta method is, in general, less than its order; more about this in chapters 3 and 4. This fact is related to the order reduction phenomenon for Runge–Kutta methods, first observed and analyzed by Michel Crouzeix; see also Remark 2.3.

2.27 (Implicit–explicit Euler method.) We write an initial value problem in the form

$$(\star) \qquad \begin{cases} y' = f(t,y) + g(t,y), \quad a \le t \le b, \\ y(a) = y_0, \end{cases}$$

i.e., we decompose the right-hand side of the o.d.e. into two parts. With the usual notation, consider the following method for problem $(\star)$

$$(\star\star) \qquad y^{n+1} = y^n + hf(t^{n+1}, y^{n+1}) + hg(t^n, y^n), \quad n = 0, \dots, N-1,$$

with $y^0 := y_0$. Obviously, method $(\star\star)$ is a combination of the implicit and the explicit Euler methods, and it reduces to them, when $g = 0$ and $f = 0$, respectively. Prove that the order of accuracy of the new method is one, equal to the order of the methods we combined to construct it. Assume now that f satisfies the one-sided Lipschitz condition (1.13) and g satisfies the Lipschitz condition (1.6) with constant L. Prove stability, with constant independent of f, and convergence (error estimate) of the method.
[*Hint*: Use the o.d.e. to check that

$$\begin{aligned} y(t^{n+1}) - y(t^n) - hf\big(t^{n+1}, y(t^{n+1})\big) - hg\big(t^n, y(t^n)\big) \\ = y(t^{n+1}) - y(t^n) - hy'(t^{n+1}) + h\big[G(t^{n+1}) - G(t^n)\big] \end{aligned}$$

with $G(t) := g\big(t, y(t)\big)$.]
[*Comment*: In some cases, when the functions f and g exhibit different behaviour, method $(\star\star)$ combines the advantages of both methods, from which it was constructed, without inheriting their drawbacks. For instance, if we use only the explicit Euler method, the constant in the error estimate necessarily depends also on f. On the other hand, if f is, e.g., linear, the computation of y^{n+1} in $(\star\star)$ is very easy, while if we use only the implicit Euler method and g is nonlinear, then to advance in time we need to solve a nonlinear equation at every time level.]

3. Runge–Kutta methods

In this chapter we will study a very important class of numerical methods for initial value problems, the class of Runge–Kutta methods. Particular methods of this class were introduced about 120 years ago[1].

In section 3.1 we introduce the class of Runge–Kutta methods and give some examples. The *explicit* Runge–Kutta methods are well defined, in the sense that the approximations they yield are well defined; existence and uniqueness of approximations for *implicit* Runge–Kutta methods, for sufficiently small step-size h, will be established in section 3.2, where we also prove stability of Runge–Kutta methods. The implementation of implicit Runge–Kutta methods is generally expensive; however, these methods are very interesting, since many of them combine excellent stability properties with high order accuracy. Sections 3.3 and 3.4 are devoted to the study of the order of accuracy and of the convergence of the methods. In section 3.5 we investigate a special subclass of these methods, the *collocation methods*, while in section 3.6 we study the absolute stability (and some of its extensions) of Runge–Kutta methods.

3.1 Preliminaries: Notation and examples

The RK (Runge–Kutta) methods are *single-step methods*, i.e., methods that for the computation of the approximation y^{n+1} use only the approximation at the previous time level, namely y^n.

[1] C. Runge: *Über die numerische Auflösung von Differentialgleichungen*. Math. Annal. **46** (1895) 167–178. W. Kutta: *Beitrag zur näherungsweisen Integration totaler Differentialgleichungen*. ZAMP **46** (1901) 435–453.

We first consider the initial value problem (1.1): Seek a function $y : [a,b] \to \mathbb{R}$ such that

$$\begin{cases} y' = f(t,y), \quad a \le t \le b, \\ y(a) = y_0. \end{cases} \tag{3.1}$$

Let $q \in \mathbb{N}, \tau_i \in \mathbb{R}, i = 1,\dots,q$ (usually $0 \le \tau_i \le 1$), $a_{ij} \in \mathbb{R}, i,j = 1,\dots,q$, and $b_i \in \mathbb{R}, i = 1,\dots,q$. For $\psi : \mathbb{R} \to \mathbb{R}$ we approximate the integrals

$$\int_0^{\tau_i} \psi(s)\,ds$$

by sums (quadrature rules, or numerical integration rules)

$$\sum_{j=1}^{q} a_{ij}\psi(\tau_j), \quad i = 1,\dots,q,$$

and the integral

$$\int_0^1 \psi(s)\,ds$$

by the sum

$$\sum_{j=1}^{q} b_j \psi(\tau_j).$$

In other words, the constants a_{ij}, τ_i, b_i describe $q+1$ quadrature rules. In these rules τ_i are the *nodes*, b_i are the *weights* of the rule for the approximation of the integral in the interval $[0,1]$, and $a_{ij}, j = 1,\dots,q$, are the weights of the rule for the approximation of the integral in the interval $[0,\tau_i]$. Each set of such constants describes a RK method. As usually, we write these constants in the form of a *Runge–Kutta tableau* (notation of J. Butcher)

$$\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1q} & \tau_1 \\ a_{21} & a_{22} & \dots & a_{2q} & \tau_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{q1} & a_{q2} & \dots & a_{qq} & \tau_q \\ \hline b_1 & b_2 & \dots & b_q & \end{array} = \begin{array}{c|c} A & \tau \\ \hline b^T & \end{array}, \tag{3.2}$$

with $A = (a_{ij}) \in \mathbb{R}^{q,q}, b = (b_1, \ldots, b_q)^T$ and $\tau = (\tau_1, \ldots, \tau_q)^T$.

Let, exactly as in the previous chapter, $N \in \mathbb{N}, h := \frac{b-a}{N}, t^n := a + nh, n = 0, \ldots, N$. Denote by $y^n, n = 0, \ldots, N$, the approximation of the nodal values $y(t^n)$ of the exact solution of (3.1) at t^n, produced by a RK method. We furthermore introduce the intermediate points

$$t^{n,i} := t^n + \tau_i h, \quad i = 1, \ldots, q. \tag{3.3}$$

Integrating both sides of the o.d.e. $y' = f(t, y)$ from t^n to $t^{n,i}, i = 1, \ldots, q$, with respect to t, we have

$$y(t^{n,i}) - y(t^n) = \int_{t^n}^{t^n + \tau_i h} f\big(t, y(t)\big)\, dt.$$

Now

$$\int_{t^n}^{t^n + \tau_i h} f\big(t, y(t)\big)\, dt = h \int_0^{\tau_i} f\big(t^n + sh, y(t^n + sh)\big)\, ds,$$

whence

$$y(t^{n,i}) = y(t^n) + h \int_0^{\tau_i} f\big(t^n + sh, y(t^n + sh)\big)\, ds, \quad i = 1, \ldots, q.$$

Approximating the integrals on the right-hand sides of this relation by the quadrature rules with nodes τ_j and weights a_{ij}, we are led to the approximations $y^{n,i}$ to the values $y(t^{n,i})$ that satisfy the relations

$$y^{n,i} = y^n + h \sum_{j=1}^{q} a_{ij} f(t^{n,j}, y^{n,j}), \quad i = 1, \ldots, q. \tag{3.4}$$

Now, integrating the o.d.e. $y' = f(t, y)$ from t^n to t^{n+1}, using the change of variables $t = t^n + hs$, and approximating the resulting integral in the interval $[0, 1]$ by the quadrature rule with nodes τ_i and weights $b_i, i = 1, \ldots, q$, we are led to the definition of the approximation y^{n+1} to the nodal value $y(t^{n+1})$,

$$y^{n+1} := y^n + h \sum_{i=1}^{q} b_i f(t^{n,i}, y^{n,i}). \tag{3.5}$$

Thus, the RK method described by the tableau (3.2) produces the approximations $y^0, \ldots, y^N$, given by the relations

$$(3.6)\qquad \left.\begin{cases} y^0 := y_0 \\ y^{n,i} = y^n + h\sum_{j=1}^{q} a_{ij} f(t^{n,j}, y^{n,j}),\ 1 \le i \le q, \\ y^{n+1} := y^n + h\sum_{i=1}^{q} b_i f(t^{n,i}, y^{n,i}) \end{cases}\right\},\ n = 0, \ldots, N-1.$$

Relations (3.6) describe the general RK methods with q *intermediate stages*. The intermediate stages $y^{n,i}$ are the solutions of system (3.4). This is a nonlinear system, in general, and in the case of problem (3.1) it consists of q equations for the q unknowns $y^{n,i}$. We will study the solvability of problem (3.4) in the next section, but note already here that, for sufficiently small h, the intermediate stages $y^{n,i}, i = 1, \ldots, q$, are uniquely defined. The stages $y^{n,i}, i = 1, \ldots, q$, are approximations to $y(t^{n,i})$, but are only used for the computation of the nodal approximations y^{n+1} through (3.5). In the error estimates for RK methods (sections 3.3 and 3.4) we will not study the approximation properties of $y^{n,i}$ to $y(t^{n,i})$; we are mainly interested in the approximation properties of y^{n+1} to the nodal values $y(t^{n+1})$.

An alternative way to describe the time stepping of the general RK method with q intermediate stages is the following: We set $k^{n,i} = f(t^{n,i}, y^{n,i})$, and write relations (3.4) and (3.5) in the form

$$(3.7)\qquad \begin{cases} k^{n,i} = f\big(t^{n,i}, y^n + h\sum_{j=1}^{q} a_{ij} k^{n,j}\big), \quad i = 1, \ldots, q, \\ y^{n+1} := y^n + h\sum_{i=1}^{q} b_i k^{n,i}. \end{cases}$$

The application of a RK method corresponding to the tableau (3.2) to initial value problems for *systems* of m first order o.d.e's,

$$(3.8)\qquad \begin{cases} y' = f(t, y), \quad a \le t \le b, \\ y(a) = y_0, \end{cases}$$

$y : [a, b] \to \mathbb{R}^m$, is obvious: All formulas, e.g., (3.6), (3.7), remain the same, with the only difference that in this case y^0, y^n, $y^{n,i}$, $k^{n,i} \in \mathbb{R}^m$. The nonlinear system (3.4) consists now of mq equations with mq unknowns, the components of the q vectors $y^{n,i} \in \mathbb{R}^m$, which are uniquely defined for sufficiently small h; see section 3.2.

Comment. Let us consider two RK tableaus of the form

$$\begin{array}{c|c} A & \tau \\ \hline b^T & \end{array} = \begin{array}{cc|c} a_{11} & a_{12} & \tau_1 \\ a_{21} & a_{22} & \tau_2 \\ \hline b_1 & b_2 & \end{array} \quad \text{and} \quad \begin{array}{c|c} \widetilde{A} & \tilde{\tau} \\ \hline \tilde{b}^T & \end{array} = \begin{array}{cc|c} a_{22} & a_{21} & \tau_2 \\ a_{12} & a_{11} & \tau_1 \\ \hline b_2 & b_1 & \end{array},$$

i.e., such that the second is obtained from the first by an appropriate renumbering, more precisely by a permutation of the lines of the matrix A (which drifts also the corresponding components of the vector τ) and of the columns of A (which drifts also the corresponding components of the vector b). As we will see, these two tableaus describe the same RK method.

For the first RK tableau we will use the usual notation, see (3.4) and (3.5), while for the second we denote by $\tilde{t}^{n,i}$ the intermediate nodes, $\tilde{t}^{n,1} = t^n + \tau_2 h = t^{n,2}$ and $\tilde{t}^{n,2} = t^n + \tau_1 h = t^{n,1}$, by $\tilde{y}^{n,i}$ the intermediate stages and by $\tilde{y}^{n+1}$ the approximation of the nodal value $y(t^{n+1})$, assuming that the approximation of the nodal value $y(t^n)$ is y^n. Then, we have

$$(3.4') \qquad \begin{cases} \tilde{y}^{n,1} = y^n + h\big[a_{22} f(\tilde{t}^{n,1}, \tilde{y}^{n,1}) + a_{21} f(\tilde{t}^{n,2}, \tilde{y}^{n,2})\big] \\ \tilde{y}^{n,2} = y^n + h\big[a_{12} f(\tilde{t}^{n,1}, \tilde{y}^{n,1}) + a_{11} f(\tilde{t}^{n,2}, \tilde{y}^{n,2})\big] \end{cases}$$

και

$$(3.5') \qquad \tilde{y}^{n+1} = y^n + h\big[b_2 f(\tilde{t}^{n,1}, \tilde{y}^{n,1}) + b_1 f(\tilde{t}^{n,2}, \tilde{y}^{n,2})\big].$$

Comparing (3.4) and (3.4′) we immediately infer that $\tilde{y}^{n,1} = y^{n,2}$ and $\tilde{y}^{n,2} = y^{n,1}$, whence (3.5) and (3.5′) yield $\tilde{y}^{n+1} = y^{n+1}$. Consequently, both tableaus describe indeed the same method.

This is valid in general: If a tableau results from another by a permutation of two lines i και j or of the columns i and j (procedure that can be repeated

several times), i.e., as we say, by appropriate renumbering, then both tableaus describe the same RK method. □

In case the matrix A of a RK method, possibly after appropriate renumbering, is *strictly lower triangular*, that is, in case $a_{ij} = 0$ for $j \geq i$, the intermediate stages $y^{n,i}$ (or $k^{n,i}$) can be recursively calculated by simple substitutions,

$$\begin{cases} y^{n,1} = y^n \\ y^{n,2} = y^n + ha_{21} f(t^{n,1}, y^{n,1}) \\ \quad \vdots \\ y^{n,q} = y^n + h \sum_{j=1}^{q-1} a_{qj} f(t^{n,j}, y^{n,j}) \end{cases}$$

or

$$\begin{cases} k^{n,1} = f(t^{n,1}, y^n) \\ k^{n,2} = f(t^{n,2}, y^n + ha_{21} k^{n,1}) \\ \quad \vdots \\ k^{n,q} = f\left(t^{n,q}, y^n + h \sum_{j=1}^{q-1} a_{qj} k^{n,j}\right). \end{cases}$$

Such RK methods are called *explicit*. All other RK methods are *implicit*. An important class of implicit RK methods are the so-called *semiimplicit*, in which the matrix A, possibly after appropriate renumbering, is *lower triangular*, i.e., such that $a_{ij} = 0$ for $j > i$. In this case the system (3.4) is of the form

$$\begin{cases} y^{n,1} = y^n + ha_{11} f(t^{n,1}, y^{n,1}) \\ y^{n,2} = y^n + ha_{21} f(t^{n,1}, y^{n,1}) + ha_{22} f(t^{n,2}, y^{n,2}) \\ \quad \vdots \\ y^{n,q} = y^n + h \sum_{j=1}^{q} a_{qj} f(t^{n,j}, y^{n,j}). \end{cases}$$

The q systems (in case $m = 1$ the q equations) are now decomposed. The first relation is an $m \times m$ nonlinear system, with solution $y^{n,1}$. Substituting $y^{n,1}$

in the second relation and solving an $m \times m$ system we determine $y^{n,2}$ etc. Notice that solving q $m \times m$ systems is by far less expensive that solving a $qm \times qm$ system.

Examples of Runge–Kutta methods

As we will see later on, a RK method is consistent, if and only if $b_1 + \cdots + b_q = 1$, i.e., if and only if the quadrature formula with nodes $\tau_1, \ldots, \tau_q$ and weights $b_1, \ldots, b_q$ integrates constant functions exactly in the interval $[0, 1]$. Also, usually, RK methods satisfy the property $a_{i1} + \cdots + a_{iq} = \tau_i$, i.e., the quadrature formula with nodes $\tau_1, \ldots, \tau_q$ and weights $a_{i1}, \ldots, a_{iq}$ integrates constant functions exactly in the interval $[0, \tau_i], i = 1, \ldots, q$. All examples we will see here fulfill these properties.

1. The RK tableau

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

describes the explicit Euler method (2.1). Indeed, we have

$$\begin{cases} y^{n,1} = y^n \\ y^{n+1} = y^n + hf(t^n, y^{n,1}), \end{cases}$$

whence

$$y^{n+1} = y^n + hf(t^n, y^n).$$

2. The RK tableau

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

describes the implicit Euler method (2.45). Indeed, we have

$$\begin{cases} y^{n,1} = y^n + hf(t^{n+1}, y^{n,1}) \\ y^{n+1} = y^n + hf(t^{n+1}, y^{n,1}). \end{cases}$$

For sufficiently small h, the function g, $g(x) := y^n + hf(t^{n+1}, x)$, is a contraction in $\mathbb{R}$ and the first equation has a unique solution $y^{n,1}$. Since the right-hand sides of these relations coincide, we have $y^{n,1} = y^{n+1}$, whence, replacing $y^{n,1}$ in the second relation by y^{n+1}, we write the implicit Euler method in

its usual form as

$$y^{n+1} = y^n + hf(t^{n+1}, y^{n+1}).$$

3. The RK tableau

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

describes the midpoint method,

$$y^{n+1} = y^n + hf\Big(t^n + \frac{h}{2}, \frac{1}{2}(y^n + y^{n+1})\Big), \quad n = 0, \dots, N-1; \tag{3.9}$$

see Exercises 2.7 and 2.25. Indeed, we have

$$\begin{cases} y^{n,1} = y^n + \dfrac{h}{2} f(t^n + \dfrac{h}{2}, y^{n,1}) \\ y^{n+1} = y^n + hf(t^n + \dfrac{h}{2}, y^{n,1}). \end{cases}$$

For sufficiently small h, the function $g, g(x) := y^n + \frac{h}{2} f(t^n + \frac{h}{2}, x)$, is a contraction, and the above relations immediately yield $y^{n,1} = \frac{1}{2}(y^n + y^{n+1})$. Replacing $y^{n,1}$ in the second relation by $\frac{1}{2}(y^n + y^{n+1})$, we are led to the midpoint method (3.9).

4. The RK tableau

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

describes the trapezoidal method (2.29). Indeed, we have

$$\begin{cases} y^{n,1} = y^n \\ y^{n,2} = y^n + \dfrac{h}{2}\big[f(t^n, y^{n,1}) + f(t^{n+1}, y^{n,2})\big] \\ y^{n+1} = y^n + \dfrac{h}{2}\big[f(t^n, y^{n,1}) + f(t^{n+1}, y^{n,2})\big]. \end{cases}$$

We immediately infer that $y^{n,1} = y^n$, and, for sufficiently small h, $y^{n,2} = y^{n+1}$. Substituting $y^{n,1}$, $y^{n,2}$ by y^n, y^{n+1}, respectively, in the third relation, we obtain (2.29).

5. More generally, the RK tableau

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

describes the theta-method, discussed in Exercises 2.22 and 2.23. This can be easily seen as in the previous example.

6. It is easily seen that the RK tableau

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

describes the method

$$\begin{cases} y^{n,2} := y^n + \dfrac{h}{2} f(t^n, y^n) \\ y^{n+1} := y^n + hf\left(t^n + \dfrac{h}{2}, y^{n,2}\right), \end{cases} \tag{3.10}$$

which is referred to as *improved Euler method* or *explicit midpoint method*.

7. An interesting family of semiimplicit, two-stage RK methods is described by the one-parameter tableau

$$\begin{array}{cc|c} \mu & 0 & \mu \\ 1-2\mu & \mu & 1-\mu \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array} \tag{3.11}$$

with $\mu \in \mathbb{R}$. In general, the order of accuracy of these methods is $p = 2$. For $\mu = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$ we obtain the very interesting (2,3) (two stages and order of

accuracy $p = 3$) *diagonally implicit RK methods RK* ((2,3) DIRK), that is, semiimplicit methods with equal diagonal entries.

8. The two-stage method described by the tableau

$$\begin{array}{cc|c} \frac{1}{4} & \frac{1}{4}-\mu & \frac{1}{2}-\mu \\ \frac{1}{4}+\mu & \frac{1}{4} & \frac{1}{2}+\mu \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array} \quad , \quad \mu = \frac{\sqrt{3}}{6}\,, \tag{3.12}$$

has very interesting properties. The corresponding RK method is the only two-stage RK method of order of accuracy 4. All other two-stage RK methods have order of accuracy at most three. This method is known as *two-stage Gauss–Legendre method*. The reason is that the entries $\tau_i = \frac{1}{2} \pm \mu$ and $b_i = 1/2$ of (3.12) are the nodes and the weights, respectively, of the Gauss quadrature formula with two nodes in the interval $[0, 1]$ and with weight function $w(x) = 1$.

9. The RK tableaus

$$\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ -1 & 2 & 0 & 1 \\ \hline \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \end{array} \tag{3.13}$$

$$\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & \frac{2}{3} & 0 & \frac{2}{3} \\ \hline \frac{1}{4} & 0 & \frac{3}{4} & \end{array} \tag{3.14}$$

$$\begin{array}{ccc|c} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{3}{4} & 0 & \frac{3}{4} \\ \hline \frac{2}{9} & \frac{1}{3} & \frac{4}{9} & \end{array} \tag{3.15}$$

describe three explicit, three-stage RK methods of order three, referred to as *third order Kutta method*, *third order Heun method*, and *Ralston method*, respectively.

10. Finally, the so-called *classical Runge–Kutta method* is explicit, has four stages and order of accuracy also four. It is described by the tableau

$$\begin{array}{cccc|c} 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 1 \\ \hline \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \end{array} \quad . \tag{3.16}$$

Two widely used families of RK methods are given in Remarks 3.3 and 3.4. More examples can be found in the literature.

3.2 Solvability and stability of RK methods

In this section we will show that the Runge–Kutta approximations are uniquely defined, at least for sufficiently small h. Then, we will prove stability of Runge– Kutta methods.

3.2.1 Solvability

We will first show that the RK approximations (3.6) are, for h sufficiently small, well defined. For the explicit methods, there is nothing to prove, since the intermediate stages $y^{n,i}$ are, in that case, obviously, uniquely defined. It remains to show that also in the case of implicit RK methods the system (3.4) is uniquely solvable, at least for sufficiently small h. For simplicity, we focus on the case of a scalar o.d.e., i.e., on the initial value problem (3.1). We assume that f satisfies the global Lipschitz condition

$$\exists L \geq 0 \ \ \forall t \in [a,b] \ \ \forall y_1, y_2 \in \mathbb{R} \quad \left| f(t,y_1) - f(t,y_2) \right| \leq L|y_1 - y_2|. \tag{3.17}$$

Proposition 3.1 (Existence and uniqueness of the approximations.) *Assume that* (3.17) *is satisfied, and let* $h < \frac{1}{\gamma}$, *with*

$$\gamma := L \max_{1\le i\le q} \sum_{j=1}^{q} |a_{ij}|.$$

Then, the system (3.4) *possesses a unique solution* $y^{n,1}, \dots, y^{n,q}$.

Proof. Consider the mapping $F : \mathbb{R}^q \to \mathbb{R}^q$,

$$F_i(x) := y^n + h \sum_{j=1}^{q} a_{ij} f(t^{n,j}, x_j), \quad 1 \le i \le q,$$

with $x = (x_1, \dots, x_q)^T$ and $F(x) = \big(F_1(x), \dots, F_q(x)\big)^T$.

Obviously, every fixed-point of F is a solution of system (3.4) and, conversely, every solution of the system (3.4) is a fixed-point of F. Therefore, it suffices to show that, for $h < \frac{1}{\gamma}$, F possesses exactly one fixed-point.

Using (3.17) we have, for $x, \tilde{x} \in \mathbb{R}^q$ and $i = 1, \dots, q$,

$$\begin{aligned} \big|F_i(x) - F_i(\tilde{x})\big| &= \Big|h \sum_{j=1}^{q} a_{ij}\big[f(t^{n,j}, x_j) - f(t^{n,j}, \tilde{x}_j)\big]\Big| \\ &\le hL \sum_{j=1}^{q} |a_{ij}||x_j - \tilde{x}_j|, \end{aligned}$$

whence

$$(3.18)\quad \forall x, \tilde{x} \in \mathbb{R}^q \ \ \forall i \in \{1, \dots, q\} \quad \big|F_i(x) - F_i(\tilde{x})\big| \le \gamma h \max_j |x_j - \tilde{x}_j|.$$

Denoting by $\|\cdot\|_\infty$ the maximum norm in $\mathbb{R}^q$, from (3.18) we obviously get

$$(3.19)\qquad \forall x, \tilde{x} \in \mathbb{R}^q \quad \|F(x) - F(\tilde{x})\|_\infty \le \gamma h \|x - \tilde{x}\|_\infty.$$

Due to the hypothesis $\gamma h < 1$, the mapping F is a *contraction* in $(\mathbb{R}^q, \|\cdot\|_\infty)$, and has, therefore, exactly one fixed-point, and the proof is complete. □

Consequently, the nonlinear in general system (3.4) possesses a unique solution, if $h < 1/\gamma$. (A completely analogous condition is needed in the case

of *systems* of o.d.e's of the form (3.8); see Exercise 3.1.) When the Lipschitz constant is large (stiff systems), this condition on the step-size h is very restrictive. Under some conditions on the function f (for instance, the one-sided Lipschitz condition (1.13) or appropriate modifications of it in the case of systems of o.d.e's) and for suitable implicit RK methods (including, e.g., the implicit Euler method, the midpoint method, the semiimplicit methods (3.11) for $\mu > 0$, the Gauss–Legendre method (3.12) with $q = 2$ intermediate stages and its generalizations for arbitrary q etc.), it is possible to show existence and uniqueness of the quantities $y^{n,i}$ for all h. (See Exercises 3.30–3.33 at the end of this chapter.)

3.2.2 Stability

Let us now proceed to a first study of the *stability* of RK methods; we will show that they satisfy, like the explicit Euler method, discrete analogues of the estimate (1.12).

Definition 3.1 (Stability of RK methods.) We say that a Runge–Kutta method is *stable*, if, for initial value problems of the form (3.1) and under the conditions of Proposition 3.1, there exists a constant C, independent of h, such that for sequences $y^0, \dots, y^N$, $z^0, \dots, z^N$, satisfying, respectively, relations (3.6) and

$$\text{(3.20)} \qquad \begin{cases} z^{n,i} = z^n + h\sum_{j=1}^{q} a_{ij} f(t^{n,j}, z^{n,j}), \ \ 1 \le i \le q, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 0, \dots, N-1, \\ z^{n+1} := z^n + h\sum_{i=1}^{q} b_i f(t^{n,i}, z^{n,i}), \end{cases}$$

there holds

$$\text{(3.21)} \qquad \max_{1 \le n \le N} |y^n - z^n| \le C\,|y^0 - z^0|.$$

We will now prove that *all* RK methods are stable in this (weak) sense. We will formulate the result a little bit more generally, in the way we will need it later on in the derivation of error estimates.

Proposition 3.2 (Stability of RK methods.) *Consider a RK method, assume that the conditions of Proposition* 3.1 *are fulfilled, and let* $y^0, \dots, y^N$ *be the approximations defined by* (3.6). *Consider also the quantities* $z^{n,i}, 0 \le n \le N-1, 1 \le i \le q$, *and* $z^n, 0 \le n \le N$, *of* $\mathbb{R}$, *for which we assume that the following relations hold true, for* $0 \le n \le N-1$,

$$\text{(3.22)} \qquad \begin{cases} z^0 \in \mathbb{R} \quad \textit{given} \\ z^{n,i} = z^n + h \displaystyle\sum_{j=1}^{q} a_{ij} f(t^{n,j}, z^{n,j}), \qquad i = 1, \dots, q, \\ z^{n+1} = z^n + h \displaystyle\sum_{i=1}^{q} b_i f(t^{n,i}, z^{n,i}) + \rho^n, \end{cases}$$

with $\rho^0, \dots, \rho^{N-1}$ *given numbers. Then, there exist constants* C_1 *and* C_2, *independent of* h, *such that*

$$\text{(3.23)} \qquad \max_{1 \le n \le N} |y^n - z^n| \le C_1 |y^0 - z^0| + \frac{C_2}{h} \max_{0 \le n \le N-1} |\rho^n|.$$

Proof. Subtracting the second relations of (3.6) and (3.22) and using (3.17) we obtain, for every $i = 1, \dots, q$,

$$|y^{n,i} - z^{n,i}| \le |y^n - z^n| + hL \sum_{j=1}^{q} |a_{ij}| |y^{n,j} - z^{n,j}|,$$

whence

$$|y^{n,i} - z^{n,i}| \le |y^n - z^n| + h\gamma \max_j |y^{n,j} - z^{n,j}|.$$

Consequently,

$$\max_i |y^{n,i} - z^{n,i}| \le |y^n - z^n| + h\gamma \max_j |y^{n,j} - z^{n,j}|.$$

In view of the assumption $\gamma h < 1$, this relation yields

$$|y^{n,i} - z^{n,i}| \le \frac{1}{1 - \gamma h} |y^n - z^n|, \quad i = 1, \dots, q, \ \ n = 0, \dots, N-1.$$

Therefore, if $h \le h_0 < 1/\gamma$, there exists a constant C, independent of h, such that

$$|y^{n,i} - z^{n,i}| \le C\,|y^n - z^n|, \quad i = 1, \dots, q, \;\; n = 0, \dots, N-1. \tag{3.24}$$

Subtracting now the last relations in (3.6) and (3.22), and using again (3.17), we get

$$|y^{n+1} - z^{n+1}| \le |y^n - z^n| + hL \sum_{i=1}^{q} |b_i||y^{n,i} - z^{n,i}| + |\rho^n|,$$

which combined with (3.24) yields, for $n = 0, \dots, N-1$,

$$|y^{n+1} - z^{n+1}| \le \Big(1 + hLC \sum_{i=1}^{q} |b_i|\Big)|y^n - z^n| + |\rho^n|. \tag{3.25}$$

Utilizing here Lemma 2.1 and setting $C' := LC \sum_{i=1}^{q} |b_i|$ we immediately infer that

$$\max_{1 \le n \le N} |y^n - z^n| \le \mathrm{e}^{C'(b-a)}|y^0 - z^0| + \frac{\mathrm{e}^{C'(b-a)} - 1}{C'h} \max_{0 \le n \le N-1} |\rho^n|, \tag{3.26}$$

which is the desired estimate (3.23). □

Notice that, for $\rho^n = 0$, the stability estimate (3.21) is a trivial consequence of (3.23). Completely analogous results hold also for the case of *systems* of o.d.e's of the form (3.8); see Exercise 3.1.

3.3 Order of accuracy and convergence of RK methods

In this section we will introduce the consistency error and the order of accuracy of a RK method, we will see how we can use the consistency error to derive bounds for the error of the method, and will see how we can determine the order of the method in some examples.

We assume that the right-hand side f of the o.d.e. $y' = f(t, y)$, and consequently also the solution y of the initial value problem (3.1), are sufficiently

regular functions. We consider a RK method (3.6), and assume that the step-size h is sufficiently small, such that the assumption in Proposition 3.1 is fulfilled, and thus the method is well defined. To define the order of accuracy of a RK method we first introduce, for $n = 0, \ldots, N-1$, the quantities $E^n \in \mathbb{R}$ as follows:

$$\text{(3.27)} \qquad \begin{cases} \zeta^{n,i} = y(t^n) + h \sum_{j=1}^{q} a_{ij} f(t^{n,j}, \zeta^{n,j}), & 1 \le i \le q, \\ E^n := \left[y(t^n) + h \sum_{i=1}^{q} b_i f(t^{n,i}, \zeta^{n,i}) \right] - y(t^{n+1}) \end{cases}$$

(see the corresponding definition and the related remarks in the case of the explicit Euler method in subsection 2.1.2). Starting at the exact nodal value $y(t^n)$, rather than at the approximation y^n, and performing one step with the method, we obtain as approximation at the time level t^{n+1} the expression in brackets in the second relation of (3.27); E^n is the difference between this approximation and the exact nodal value y^{n+1}. The quantities $\zeta^{n,i}, i = 1, \ldots, q$, and consequently also E^n are well defined according to Proposition 3.1. The quantity E^n is called *consistency error* or *local discretization error* or simply *local error* of the method. We say that the *order of accuracy* (or simply that the *order*) of the Runge–Kutta method is p, if p is the largest integer, for which there exist a constant $\widetilde{C}$, independent of h and N, (depending on the data of the problem and on the solution, which is assumed sufficiently smooth, as well as on the concrete RK method) such that

$$\text{(3.28)} \qquad \max_{0 \le n \le N-1} |E^n| \le \widetilde{C} h^{p+1}.$$

Let us emphasize that we require an estimate of the form (3.28) to be valid for *all* initial value problems of the form (3.1) with sufficiently regular functions f and y. The consistency error can not be computed but plays a key role in the analysis of the method.

The following interpretation of the consistency error is evident from its definition, see relations (3.27): If we replace the approximation by the exact solution, i.e., y^n by $y(t^n)$ and y^{n+1} by $y(t^{n+1})$, in (3.4) and (3.5), then (3.5) is

not satisfied exactly any more; the difference of both sides in the new relation is exactly the quantity E^n, which measures the error of the RK method after one step, with initial value $y(t^n)$.

The RK method (3.6) is called *consistent*, if its order p is at least one; a sufficient and necessary condition for the consistency of a RK method is $b_1 + \cdots + b_q = 1$; see Exercise 3.9.

In the case of *$m \times m$ systems of o.d.e's* of the form (3.8) the quantities E^n, that are well defined through relations (3.27), provided h is sufficiently small (see Exercise 3.1), are vectors in $\mathbb{R}^m$. We say that the order of the RK is p, if instead of (3.28) the analogue estimate

$$\max_{0\le n\le N-1} \|E^n\|_\infty \le \widetilde{C} h^{p+1}, \tag{3.29}$$

holds true. The order of RK methods for systems of o.d.e's coincides with their order for scalar o.d.e's.

The determination of the order of accuracy of RK methods is a key point in the analysis of RK methods for o.e.d's. From the consistency estimate (3.28) (or from (3.29)) we immediately obtain (utilizing also the stability of the method; see Proposition 3.2) that the (global) discretization *error* of the method is $O(h^p)$, i.e., of order p:

Theorem 3.1 (Error estimates.) *Assume that the Lipschitz condition* (3.17) *is satisfied and that problem* (3.1) *possesses a sufficiently regular solution. With the constant γ introduced in Proposition* 3.1*, let $h_0 > 0$ be such that $\gamma h_0 < 1$, και $0 < h \le h_0$. We consider the RK method* (3.6) *and assume that the consistency error estimate* (3.28) *holds true. Then we have*

$$\max_{0\le n\le N} \left|y(t^n) - y^n\right| \le \frac{\widetilde{C}}{C'}\left[\mathrm{e}^{C'(b-a)} - 1\right] h^p, \tag{3.30}$$

with the constants $\widetilde{C}$ and C, C', independent of h, defined in (3.28) *and in the proof of Proposition* 3.2*, respectively.*

Proof. We write (3.27) in the form

$$\begin{cases} \zeta^{n,i} = y(t^n) + h\sum_{j=1}^{q} a_{ij} f(t^{n,j}, \zeta^{n,j}), & 1 \le i \le q, \\ y(t^{n+1}) = \Big[y(t^n) + h \sum_{i=1}^{q} b_i f(t^{n,i}, \zeta^{n,i}) \Big] - E^n. \end{cases}$$

According to Proposition 3.2, we then have, taking into account the fact that $y^0 = y_0 = y(t^0)$,

$$\max_{0\le n\le N} |y(t^n) - y^n| \le C_2 h^{-1} \max_{0\le n\le N-1} |E^n|,$$

see (3.26), from which the desired result follows, in view of (3.28). □

Under analogous assumptions (see Exercise 3.1) the corresponding result holds also for systems of o.d.e's. (We replace the absolute value by the norm $\|y(t^n) - y^n\|_\infty$ on the left-hand side of (3.30).)

Notice also that the consistency of Runge–Kutta methods is a necessary condition for convergence; see Exercise 3.10.

In the sequel in this paragraph, as well as in the next section, we will focus on the determination of the order of accuracy of various RK methods. In any case, we will define the quantities $\zeta^{n,i}$ and the consistency error E^n by the relations (3.27) and will try to determine the largest integer p for which the estimate (3.28) holds for all initial value problems (3.1) with smooth functions y and f. For simplicity, we will determine the order for scalar o.d.e's. (The order for systems is the same.)

Example 3.1 The explicit midpoint method (3.10)

For the explicit midpoint method we have, by simple substitution, $\zeta^{n,1} = y(t^n)$ and $\zeta^{n,2} = y(t^n) + \frac{h}{2} f\big(t^n, y(t^n)\big)$. Consequently, the second relation of (3.27) immediately yields

$$E^n = y(t^n) + hf\Big(t^n + \frac{h}{2}, y(t^n) + \frac{h}{2} f\big(t^n, y(t^n)\big)\Big) - y(t^{n+1}). \tag{3.31}$$

We will now determine the order of accuracy in two ways. The first way is to expand the consistency error in power series of h, using the Taylor theorem and the fact that

y is a solution of the o.d.e. $y'(t) = f\big(t, y(t)\big)$, substituting the derivatives of y by quantities including f and appropriate partial derivatives of f. First, by Taylor expanding a function of a single variable,

$$y(t^{n+1}) = y(t^n) + hy'(t^n) + \frac{h^2}{2}y''(t^n) + \frac{h^3}{6}y'''(\xi^n),$$

with $\xi^n \in (t^n, t^{n+1})$. From the o.d.e. $y' = f(t,y)$ we have $y'' = f_t(t,y) + f_y(t,y)y' = f_t(t,y) + f(t,y)f_y(t,y)$, whence the previous Taylor expansion can be written in the form

$$\begin{aligned} y(t^{n+1}) &= y(t^n) + hf\big(t^n, y(t^n)\big) \\ &\quad + \frac{h^2}{2}\big[f_t\big(t^n, y(t^n)\big) + f\big(t^n, y(t^n)\big)f_y\big(t^n, y(t^n)\big)\big] + O(h^3). \end{aligned} \tag{3.32}$$

Furthermore, by Taylor expanding a function of two variables, we have

$$\begin{aligned} &f\Big(t^n + \frac{h}{2}, y(t^n) + \frac{h}{2}f\big(t^n, y(t^n)\big)\Big) = f\big(t^n, y(t^n)\big) + \frac{h}{2}f_t\big(t^n, y(t^n)\big) + \\ &\frac{h}{2}f\big(t^n, y(t^n)\big)f_y\big(t^n, y(t^n)\big) + \\ &\frac{1}{2}\Big[\frac{h^2}{4}f_{tt}(P^n) + \frac{h^2}{2}f\big(t^n, y(t^n)\big)f_{ty}(P^n) + \frac{h^2}{4}f\big(t^n, y(t^n)\big)f_{yy}(P^n)\Big], \end{aligned} \tag{3.33}$$

with P^n a point in the segment joining the points $\big(t^n, y(t^n)\big)$ and $\big(t^n + \frac{h}{2}, y(t^n) + \frac{h}{2}f\big(t^n, y(t^n)\big)\big)$ in the plane ty. (For (3.32) and (3.33) to hold, we assume that y is three times continuously differentiable and f two times continuously differentiable as a function of two variables.) We thus infer from (3.31)–(3.33) that $E^n = O(h^3)$, i.e., that the order of accuracy of the method is at least two. Moreover, we consider the initial value problem

$$\begin{cases} y' = t^2, \quad 0 \le t \le 1, \\ y(0) = 0, \end{cases}$$

with solution $y(t) = t^3/3$, and notice that, according to (3.31),

$$E^n = \frac{1}{3}(t^n)^3 + h\big(t^n + \frac{h}{2}\big)^2 - \frac{1}{3}(t^n + h)^3 = -\frac{h^3}{12}.$$

We infer that the order of accuracy is in general at most two. Summarizing, the order of the method is exactly two.

Alternatively, we try, using Taylor expansions around appropriate points and the o.d.e., to express as many terms in the expansion of E^n as we can in terms of derivatives $y^{(k)}$ of y. In our case, we have

$$\begin{aligned}
E^n &= y(t^n) + h\, f\big(t^n + \frac{h}{2}, y(t^n) + \frac{h}{2}\, y'(t^n)\big) - y(t^{n+1}) \\
&= y(t^n + \frac{h}{2}) - \frac{h}{2}\, y'(t^n + \frac{h}{2}) + \frac{h^2}{8}\, y''(t^n + \frac{h}{2}) - \frac{h^3}{48}\, y'''(\xi^n) \\
&\quad + h\, f\big(t^n + \frac{h}{2}, y(t^n + \frac{h}{2}) - \frac{h^2}{8}\, y''(\vartheta^n)\big) \\
&\quad - \big[y(t^n + \frac{h}{2}) + \frac{h}{2}\, y'(t^n + \frac{h}{2}) + \frac{h^2}{8}\, y''(t^n + \frac{h}{2}) + \frac{h^3}{48}\, y'''(\tilde{\xi}^n)\big] \\
&= -h\, y'(t^n + \frac{h}{2}) - \frac{h^3}{48}\big[y'''(\xi^n) + y'''(\tilde{\xi}^n)\big] \\
&\quad + h\, f\big(t^n + \frac{h}{2}, y(t^n + \frac{h}{2}) - \frac{h^2}{8}\, y''(\vartheta^n)\big) \\
&= -h\, y'(t^n + \frac{h}{2}) - \frac{h^3}{48}\big[y'''(\xi^n) + y'''(\tilde{\xi}^n)\big] \\
&\quad + h\, f\big(t^n + \frac{h}{2}, y(t^n + \frac{h}{2})\big) - \frac{h^3}{8}\, y''(\vartheta^n)\, f_y\big(t^n + \frac{h}{2}, \tilde{y}^n\big),
\end{aligned}$$

i.e.,

$$E^n = -\frac{h^3}{48}\big[y'''(\xi^n) + y'''(\tilde{\xi}^n)\big] - \frac{h^3}{8}\, y''(\vartheta^n)\, f_y\big(t^n + \frac{h}{2}, \tilde{y}^n\big)$$

with $\xi^n, \vartheta^n, \tilde{\xi}^n \in (t^n, t^{n+1})$ and $\tilde{y}^n$ between $y(t^n + \frac{h}{2})$ and $y(t^n + \frac{h}{2}) - \frac{h^2}{8}\, y''(\vartheta^n)$. Applying the method to the special problem we chose before, we again get $E^n = -h^3/12$. Furthermore, the above relation yields the estimate

$$|E^n| \le \frac{h^3}{8}\Big(\frac{1}{3}\|y'''\|_\infty + L\|y''\|_\infty\Big),$$

where

$$\|y^{(k)}\|_\infty = \max_{a \le t \le b} \big|y^{(k)}(t)\big| \quad \text{και} \quad L = \sup_{t,y} \big|f_y(t,y)\big|.$$

The order of the method is thus two. □

We will now investigate the general $q-$stage RK method. If the method is *explicit*, and we express the intermediate stages $y^{n,i}, i = 1, \ldots, q$, in terms of y^n, t^n and h, we can write the step $y^n \mapsto y^{n+1}$ of the method in the form

$$y^{n+1} = y^n + h\Phi(t^n, y^n; h), \quad n = 0, \ldots, N-1, \tag{3.34}$$

with a suitable function Φ, depending on f, that can be easily determined. For example, in the case of the explicit midpoint method of Example 3.1, we have

$$\Phi(t^n, y^n; h) = f\Big(t^n + \frac{h}{2}, y^n + \frac{h}{2} f(t^n, y^n)\Big).$$

Consequently, the determination of the order of accuracy of the method amounts to the use of Taylor expansion in the formula of the consistency error of the method

$$E^n = y(t^n) + h\Phi\big(t^n, y(t^n); h\big) - y(t^{n+1}).$$

As the number of stages q increases, the calculation of the terms with derivatives of f in the Taylor expansions get extremely complicated, since the number of the terms increases exponentially. For example, we have

$$\begin{aligned} y' &= f \\ y'' &= f_t + f f_y \\ y''' &= f_{tt} + 2 f f_{ty} + f_t f_y + f f_y^2 + f^2 f_{yy} \\ &\vdots \end{aligned}$$

The calculations are considerably simplified by using notation and techniques of graph theory ("Butcher trees"), see [16, §II.2], and symbolic computation packages. For small stage numbers they can be done elementary; see Exercises 3.2–3.3.

It is easily seen that necessary and sufficient conditions for a $2-$stage method to be of order $p = 2$ are the following

$$a_{21} = \tau_2, \;\; b_1 + b_2 = 1, \;\; b_2\tau_2 = \frac{1}{2}; \tag{3.35}$$

these conditions lead to the existence of a one-parameter family of such methods. With a little more effort, it is possible to prove that sufficient and necessary conditions for a $3-$stage method to be of order $p = 3$ are the following

$$\begin{cases} a_{21} = \tau_2, \;\; a_{31} + a_{32} = \tau_3, \;\; b_1 + b_2 + b_3 = 1, \\ b_2\tau_2 + b_3\tau_3 = \dfrac{1}{2}, \;\; b_2\tau_2^2 + b_3\tau_3^2 = \dfrac{1}{3}, \;\; b_3 a_{32}\tau_2 = \dfrac{1}{6}, \end{cases} \tag{3.36}$$

that is a system of six equations with eight unknowns $a_{21}, a_{31}, a_{32}, \tau_2, \tau_3, b_1, b_2, b_3$. This system possesses a biparametric family of solutions.

The corresponding conditions for $p = q = 4$ are more complicated. We are led to a system of eleven equations

$$\begin{aligned}
&\sum_{j=1}^{i-1} a_{ij} = \tau_i,\ i = 2,3,4,\\
&b_1 + b_2 + b_3 + b_4 = 1,\\
&b_2\tau_2 + b_3\tau_3 + b_4\tau_4 = \frac{1}{2},\\
&b_2\tau_2^2 + b_3\tau_3^2 + b_4\tau_4^2 = \frac{1}{3},\\
&b_2\tau_2^3 + b_3\tau_3^3 + b_4\tau_4^3 = \frac{1}{4},\\
&b_3a_{32}\tau_2 + b_4a_{42}\tau_2 + b_4a_{43}\tau_3 = \frac{1}{6},\\
&b_3a_{32}\tau_2^2 + b_4a_{42}\tau_2^2 + b_4a_{43}\tau_3^2 = \frac{1}{12},\\
&b_3\tau_3a_{32}\tau_2 + b_4\tau_4a_{42}\tau_2 + b_4\tau_4a_{43}\tau_3 = \frac{1}{8},\\
&b_4a_{43}a_{32}\tau_2 = \frac{1}{24},
\end{aligned}$$

for the 13 unknowns $\tau_i, a_{ij},\ i = 2,3,4,\ j = 1,\ldots,i-1,\ b_i, i = 1,\ldots,4$, which, as can be shown, possesses again a biparametric family of solutions. Generally, it is known from Butcher's work (see [16, §II.5]) that the highest attainable order of an explicit $q-$stage RK method is q, if $q \leq 4$. The highest attainable order of an explicit $5-$stage, $6-$stage and $7-$stage method is, four, five and six, respectively.

In the case of *implicit* methods, it is theoretically feasible, solving the system (3.4) with respect to $y^{n,1}, \ldots, y^{n,q}$, for sufficiently small h, and substituting in (3.5), to symbolically write the method in the form (3.34). However, in practice this is in general impossible because it requires solving a nonlinear system in closed form. But in reality, we do not actually need Φ in closed form, but rather its expansion in powers of h. We can compute the terms of

this expansion that are needed each time by successive compositions of power series of the intermediate stages, usually of $k^{n,i}$; see (3.7). The calculations are considerably simplified by using Butcher trees and symbolic computation packages. For small stage numbers they can be done elementary. As we will see in the next section, the highest attainable order of an implicit $q-$stage RK methods is $2q$. Indeed, for any q, there exists exactly one $q-$stage RK method of order $p = 2q$, the so-called RK Gauss–Legendre method. The Gauss–Legendre methods with $q = 1$ and $q = 2$ stages are, respectively, the (implicit) midpoint method and the method (3.12).

There are criteria (necessary and sufficient conditions) for the order of a RK method to be p; in general they are not so easy to use. In the next section we will see some simplifying, sufficient conditions for the order of a RK method to be (at least) p. Their advantages are first that they are easy to use and, secondly and more importantly, that they yield, in fact, the exact order of some important families of RK methods.

3.4 Sufficient conditions for a certain order of accuracy

In this section we state a Theorem due to Butcher [4], that gives sufficient condition on the entries q, a_{ij}, b_i, τ_i in the Bucher tableau of a RK method for the order of the method to be (at least) p. The proof was later simplified by Crouzeix [7].

Theorem 3.2 (Simplifying conditions.) *Consider the initial value problem* (3.1) *and assume that f is sufficiently regular in $[a,b] \times \mathbb{R}$ and f as well as some appropriate derivatives of f are bounded in $[a,b] \times \mathbb{R}$. Assume that there exist $p, s, r \geq 0$ such that*

$$\sum_{i=1}^{q} b_i \tau_i^k = \frac{1}{k+1}, \quad \text{for } 0 \leq k \leq p-1, \tag{3.37}$$

$$\sum_{j=1}^{q} a_{ij} \tau_j^k = \frac{\tau_i^{k+1}}{k+1}, \quad 1 \leq i \leq q, \quad \text{for } 0 \leq k \leq s-1, \tag{3.38}$$

$$\sum_{i=1}^{q} b_i \tau_i^k a_{ij} = \frac{b_j(1-\tau_j^{k+1})}{k+1}, \quad 1 \le j \le q, \quad \text{for } 0 \le k \le r-1, \tag{3.39}$$

$$p \le r+s+1 \quad \text{and} \quad p \le 2s+2. \tag{3.40}$$

Then the order of the RK method (3.6) *is at least* p. *Conditions* (3.37) –(3.40) *are called 'simplifying' sufficient conditions for order of accuracy* p. □

The meaning of (3.40) is that the order of the method is at least $\min(p, r+s+1, 2s+2)$.

Obviously, conditions (3.37) and (3.38) of Theorem 3.2 express, respectively, that the corresponding quadrature rules, respectively, are exact for polynomials of degree at most $p-1$ and $s-1$, respectively. We say that these rules are of *order* $p-1$ and $s-1$, respectively. It is useful to see (without proof) a corollary of Theorem 3.2, in which sufficient conditions for the RK method (3.6) to be of order p are given, in which only the properties of the quadrature rules enter, that is condition (3.39) is irrelevant in this case.

If p and s are the largest integers for which (3.37) and (3.38) are valid, then the integer $\tilde{p} := \min(p, s)$ is the so-called *stage order* of the method.

Corollary 3.1 (Butcher–Crouzeix)
a) *If the quadrature rule with nodes* $\tau_1, \ldots, \tau_q$ *and weights* $b_1, \ldots, b_q$ *integrates polynomials of degree at most* $p-1$ *exactly, i.e., if* (3.37) *holds, and the quadrature rules with nodes* $\tau_1, \ldots, \tau_q$ *and weights* $a_{i1}, \ldots, a_{iq}$, $i = 1, \ldots, q$, *integrate polynomials of degree at most* $p-2$ *exactly, i.e., if* (3.38) *holds with* $s = p-1$, *then the order of the RK method is* p.
b) *Let* q' *be the number of the distinct points between* $\tau_1, \ldots, \tau_q$. *If the quadrature rule with nodes* $\tau_1, \ldots, \tau_q$ *and weights* $b_1, \ldots, b_q$ *integrates polynomials of degree at most* $p-1$ *exactly, i.e., if* (3.37) *holds, and the quadrature rules with nodes* $\tau_1, \ldots, \tau_q$ *and weights* $a_{i1}, \ldots, a_{iq}$, $i = 1, \ldots, q$, *integrate polynomials of degree at most* $q'-1$ *exactly, i.e., if* (3.38) *holds with* $s = q'$, *then the order of the RK method is* p.
c) *There exists exactly one* $q-$*stage RK method of order of accuracy* $p = 2q$, *for each* $q \in \mathbb{N}$. *This is the method with* τ_i *and* b_i *the nodes and the*

weights, respectively, of the Gauss–Legendre quadrature rule (*i.e., the Gauss quadrature rule with weight function* $w(x) = 1$) *in the interval* $[0,1]$. (*The entries* a_{ij} *are such that* (3.38) *is satisfied for* $s = q$. *For each* i, *this gives a* $q \times q$ *linear system for* $a_{i1}, \ldots, a_{iq}$, *with invertible matrix* (*Vandermonde matrix*).) □

Remark 3.1 There exist very interesting families of RK methods with pairwise distinct $\tau_1, \ldots, \tau_q \in [0,1]$ and such that (3.38) holds with $s = q$. One such family are the RK Gauss–Legendre methods, which are mentioned in Corollary 3.1c; another family are the RK Radau IIA, discussed in Remark 3.4 in the sequel. For such methods Corollary 3.1b (and in some other cases Corollary 3.1a) yields the exact order of accuracy: if p is the largest integer for which (3.37) is satisfied then, under the conditions mentioned above, the exact order of the method is p. As we will see in the sequel, see Proposition 3.3, the RK methods satisfying these conditions are equivalent to collocation methods; see the next section and, in particular, the discussion about the order of accuracy of collocation methods. □

3.5 Collocation methods

In this section we will study *collocation* methods for initial value problems. The approximations are now defined in the whole interval $[a,b]$ rather than at the nodes of a partition, in contrast to the case of Runge–Kutta or multistep methods, which we will study in the next chapter. The approximations are continuous piecewise polynomial functions and satisfy the differential equation exactly at some intermediate nodes. Collocation methods are closely related to a subclass of implicit Runge–Kutta methods, as we will see in the sequel, and this is why we are studying them in this chapter.

Let $q \in \mathbb{N}$ and $\tau_1, \ldots, \tau_q \in (0,1]$ be pairwise distinct points, the collocation nodes; without loss of generality we assume that $0 < \tau_1 < \cdots < \tau_q \leq 1$. (For simplicity we assumed here that $\tau_1 > 0$. We will see in the sequel a case with $\tau_1 = 0$, in the third of the examples we will discuss.) We consider the initial value problem (3.1) and a uniform partition of the interval $[a,b]$ with stepsize h, i.e., for $N \in \mathbb{N}$ and $h := (b-a)/N$, we set $t^n := a + nh$, $n = 0, \ldots, N$.

In addition, we consider the intermediate nodes $t^{n,i} := t^n + \tau_i h, i = 1, \ldots, q$. Notice that $t^{n,i} \in [t^n, t^{n+1}], i = 1, \ldots, q$. The *collocation* method is now the following: Seek a function Y, continuous in $[a, b]$ and polynomial of degree at most q in each subinterval $[t^n, t^{n+1}]$, such that $Y(a) = y_0$ and, for $n = 0, \ldots, N-1$,

$$Y'(t^{n,i}) = f\big(t^{n,i}, Y(t^{n,i})\big), \quad i = 1, \ldots, q. \tag{3.41}$$

In the subinterval $[t^0, t^1]$, besides (3.41), the approximation Y must also satisfy the initial condition $Y(a) = y_0$. Similarly, in all other subintervals, if we have already computed the approximation in the subinterval $[t^{n-1}, t^n]$, then the value of the approximation at the node t^n in the subinterval $[t^n, t^{n+1}]$ is the same as in the preceding interval, to ensure continuity in the whole interval $[a, b]$. Before we proceed, we have to check whether the number of conditions in every subinterval coincides with the degrees of freedom in the same subinterval, so that we may hope that the approximations are well defined. Indeed, the degrees of freedom are $q + 1$ in each subinterval, since the dimension of polynomials of degree up to q is $q + 1$, and the conditions are the q relations (3.41) and the condition at the left endpoint of the interval $[t^n, t^{n+1}]$, since the value of the solution there has been pre-asigned by its form in the previous subinterval. So we have in total $q+1$ conditions and $q+1$ degrees of freedom.

Remark 3.2 (Pointwise formulation of collocation methods.) Let I_{q-1} denote the interpolation operator at the collocation nodes $t^{n,1}, \ldots, t^{n,q}$, i.e., such that $I_{q-1}\varphi$ is a polynomial of degree at most $q-1$ satisfying

$$(I_{q-1}\varphi)(t^{n,i}) = \varphi(t^{n,i}), \quad i = 1, \ldots, q.$$

Then, (3.41) can, obviously, be also written in the form

$$Y'(t^{n,i}) = \Big(I_{q-1} f\big(\cdot, Y(\cdot)\big)\Big)(t^{n,i}), \quad i = 1, \ldots, q.$$

Now, since both Y' and $I_{q-1} f\big(\cdot, Y(\cdot)\big)$ are polynomials of degree at most $q-1$ coinciding at q distinct points, they coincide in the whole interval (t^n, t^{n+1}), i.e, (3.41) can be equivalently written in a pointwise formulation as

$$Y'(t) = \Big(I_{q-1} f\big(\cdot, Y(\cdot)\big)\Big)(t), \quad t \in (t^n, t^{n+1}). \tag{3.42}$$

This formulation has been useful in the a posteriori error analysis. □

We now proceed to study the relation between collocation methods and a subclass of implicit RK methods.

Proposition 3.3 (Equivalence of collocation to RK methods.) *We consider the collocation method with nodes $\tau_1, \dots, \tau_q$. Denote by $L_i \in \mathbb{P}_{q-1}, i = 1, \dots, q$, the Lagrange polynomials to the nodes $\tau_1, \dots, \tau_q$ and consider the Runge–Kutta method with the same $\tau_1, \dots, \tau_q$ and*

$$a_{ij} := \int_0^{\tau_i} L_j(\tau)\, d\tau, \quad b_i := \int_0^1 L_i(\tau)\, d\tau, \quad i, j = 1, \dots, q. \tag{3.43}$$

Then, with the usual notation, there holds

$$Y(t^n) = y^n \text{ and } Y(t^{n,i}) = y^{n,i}, \quad i = 1, \dots, q, \tag{3.44}$$

for $n = 0, \dots, N-1$.

Proof. Obviously $Y(t^0) = y^0$. Assume now that $Y(t^n) = y^n$. We will show that $Y(t^{n,i}) = y^{n,i}, i = 1, \dots, q$, and $Y(t^{n+1}) = y^{n+1}$. Recall that the Lagrange polynomials L_i satisfy the relations $L_i(\tau_j) = \delta_{ij}, i, j = 1, \dots, q$. For $i = 1, \dots, q$, we have, with $\widetilde{L}_j(s) := L_j\big((s - t^n)/h\big), s \in [t^n, t^{n+1}], j = 1, \dots, q$, since Y' is a polynomial of degree at most $q-1$ in the interval $[t^n, t^{n+1}]$,

$$\begin{aligned} Y(t^{n,i}) - y^n &= Y(t^{n,i}) - Y(t^n) = \int_{t^n}^{t^{n,i}} Y'(s)\, ds \\ &= \int_{t^n}^{t^{n,i}} \sum_{j=1}^{q} Y'(t^{n,j}) \widetilde{L}_j(s)\, ds = \sum_{j=1}^{q} Y'(t^{n,j}) \int_{t^n}^{t^{n,i}} \widetilde{L}_j(s)\, ds \\ &= \sum_{j=1}^{q} f\big(t^{n,j}, Y(t^{n,j})\big) h \int_0^{\tau_i} L_j(\tau)\, d\tau = h \sum_{j=1}^{q} a_{ij} f\big(t^{n,j}, Y(t^{n,j})\big). \end{aligned}$$

Hence, we indeed have $Y(t^{n,i}) = y^{n,i}, i = 1, \dots, q$. Furthermore, completely

analogously, we have

$$\begin{aligned} Y(t^{n+1}) - y^n = Y(t^{n+1}) - Y(t^n) &= \int_{t^n}^{t^{n+1}} Y'(s)\,ds \\ &= \int_{t^n}^{t^{n+1}} \sum_{i=1}^{q} Y'(t^{n,i})\widetilde{L}_i(s)\,ds = \sum_{i=1}^{q} Y'(t^{n,i}) \int_{t^n}^{t^{n+1}} \widetilde{L}_i(s)\,ds \\ &= \sum_{i=1}^{q} f\big(t^{n,i}, Y(t^{n,i})\big)h \int_0^1 L_i(\tau)\,d\tau = h\sum_{i=1}^{q} b_i f\big(t^{n,i}, Y(t^{n,i})\big), \end{aligned}$$

whence $Y(t^{n+1}) = y^{n+1}$. □

Relations (3.43) for a_{ij} mean simply that (3.38) holds with $s \geq q$. Similarly, relations (3.43) for b_i mean that (3.37) is satisfied with $p \geq q$. It is now easy to infer that, inversely, implicit RK methods with pairwise distinct $\tau_1, \ldots, \tau_q \in (0, 1]$ and such that relations (3.37) and (3.38) are satisfied with $p, s \geq q$ correspond, in the sense of Proposition 3.3, to collocation methods, i.e., are equivalent to collocation methods. Moreover, the order of accuracy of such RK methods is p, with p the largest integer for which (3.37) is satisfied; see Corollary 3.1b. Equivalently, the order p is characterized by the orthogonality property

$$\forall r \in \mathbb{P}_{p-1-q} \quad \int_0^1 (\tau - \tau_1)\cdots(\tau - \tau_q) r(\tau)\,d\tau = 0.$$

Therefore, in contrast to the general case of RK methods, the determination of the order of collocation methods is simple.

It is an easy consequence of Proposition 3.3 that, if we first compute the intermediate stages $y^{n,i}, i = 1, \ldots, q$, of the corresponding RK method, then the approximation Y, the result of the collocation method, is simply the interpolation polynomial passing through the points (t^n, y^n) and $(t^{n,i}, y^{n,i}), i = 1, \ldots, q$. We infer that existence and uniqueness of the collocation approximate solution follows easily from the corresponding results for Runge–Kutta methods; thus we will not repeat the analysis here.

As far as the approximation properties of collocation methods are concerned, we distinguish two issues here, the approximation at the nodes t^n and

the approximation in the whole interval $[a, b]$. The first issue is immediately settled through the equivalence of collocation and Runge–Kutta methods: If p is the order of the method and the solution y of initial value problem (3.1) is sufficiently smooth, then there exists a constant C such that, for sufficiently small h,

$$\max_{0\le n\le N} \left|y(t^n) - Y(t^n)\right| \le C h^p; \tag{3.45}$$

see (3.30). Concerning the approximation properties in the whole interval, we simply note that

$$\max_{a\le t\le b} \left|y(t) - Y(t)\right| \le C h^{\min(q+1,p)}. \tag{3.46}$$

This is due to the fact that when we approximate with polynomials of degree at most q, then the approximation order is at most $q + 1$, independently of the approximation procedure, that is even when y if given and we approximate it by its best approximation from the corresponding polynomial space; we assume here, of course, that y does not belong to the space from which we approximate, since otherwise there is no error at all. In case $p > q + 1$, the approximation order at the nodes is higher than in the whole interval, and this phenomenon is referred to in the literature as *superconvergence*. From what we know from the quadrature theory, we can easily construct a collocation method of order $p = 2q$, which is the highest possible order; see Corollary 3.1 and the results about Gaussian quadrature in numerical integration and/or elementary Numerical Analysis books.

Let us, however, emphasize that, having nodal approximations of order p at our disposal, we can compute approximations of the same order in the whole interval $[a, b]$, provided the solution y is sufficiently smooth, for instance by interpolating at $p + 1$ consecutive points $\big(t^n, Y(t^n)\big)$. In case N is not an integer multiple of p, the approximation in an interval of the form $[t^{mp}, t^N]$ can be computed by interpolating at the nodes of this subinterval and at as many of the preceding nodes $t^{mp-1}, t^{mp-2}, \ldots$, as are needed to have $p + 1$ interpolation points in total.

Examples of collocation methods

We will now see a few examples of collocation methods and their relation to implicit RK methods. Of course, this relation follows also from Proposition 3.3; in these simple examples we prefer to study this relation directly.
1. $q = 1, \tau_1 = 1$: In this case the approximate solution Y is a polynomial of degree at most one in each subinterval $[t^n, t^{n+1}]$ and the collocation method (3.41) takes the form

$$Y'(t^{n+1}) = f\big(t^{n+1}, Y(t^{n+1})\big). \tag{3.47}$$

In (3.47) the derivative is taken, obviously, as left-hand, and, since Y is a polynomial of degree at most one in $[t^n, t^{n+1}]$, we have

$$Y'(t^{n+1}) = \frac{1}{h}\big[Y(t^{n+1}) - Y(t^n)\big],$$

whence (3.47) takes the form

$$Y(t^{n+1}) = Y(t^n) + hf\big(t^{n+1}, Y(t^{n+1})\big). \tag{3.48}$$

It is thus obvious that the approximations $Y(t^n)$ at the nodes coincide with those of the implicit Euler method. Conversely, if we have computed the nodal approximations $Y(t^n)$ and $Y(t^{n+1})$ by the implicit Euler method, then we can recover Y in the interval $[t^n, t^{n+1}]$ by linearly interpolating at the endpoints, i.e., from the formula

$$Y(t) = Y(t^n) + \frac{1}{h}\big[Y(t^{n+1}) - Y(t^n)\big](t - t^n), \quad t \in [t^n, t^{n+1}]. \tag{3.49}$$

2. $q = 1, \tau_1 = 1/2$: In this case the approximate solution Y is a polynomial of degree at most one in each subinterval $[t^n, t^{n+1}]$ and the collocation method (3.41) takes the form

$$Y'\big(t^n + \frac{h}{2}\big) = f\big(t^n + \frac{h}{2}, Y\big(t^n + \frac{h}{2}\big)\big). \tag{3.50}$$

Now, in analogy to the previous example,

$$Y'\big(t^n + \frac{h}{2}\big) = \frac{1}{h}\big[Y(t^{n+1}) - Y(t^n)\big], \quad Y\big(t^n + \frac{h}{2}\big) = \frac{1}{2}\big[Y(t^n) + Y(t^{n+1})\big],$$

and (3.50) can be written as

$$Y(t^{n+1}) = Y(t^n) + hf\Big(t^n + \frac{h}{2}, \frac{1}{2}\big[Y(t^n) + Y(t^{n+1})\big]\Big). \tag{3.51}$$

It is thus obvious that the nodal approximations $Y(t^n)$ coincide with those given by the midpoint method. Conversely, if we have computed the nodal approximations $Y(t^n)$ and $Y(t^{n+1})$ by the midpoint method, then we can recover Y in the interval $[t^n, t^{n+1}]$ from formula (3.49).

3. $q = 2, \tau_1 = 0, \tau_2 = 1$: Until now we assumed, for simplicity, that $\tau_1 > 0$. Now we will look at an example with $\tau_1 = 0$. In our case, the approximate solution Y is a polynomial of degree at most two in each subinterval $[t^n, t^{n+1}]$ and the collocation method (3.41) takes the form

$$\begin{cases} Y'(t^n) = f\big(t^n, Y(t^n)\big), \\ Y'(t^{n+1}) = f\big(t^{n+1}, Y(t^{n+1})\big). \end{cases} \tag{3.52}$$

Knowing the values of the function and its first derivative at t^n, we write Y in the form

$$Y(t) = Y(t^n) + f\big(t^n, Y(t^n)\big)(t - t^n) + \frac{1}{2}C(t - t^n)^2, \quad t \in [t^n, t^{n+1}],$$

with a constant C which we will calculate in the sequel. From this relation we immediately infer that $Y'(t^{n+1}) = f\big(t^n, Y(t^n)\big) + Ch$, whence, taking (3.52) into account, we have $f\big(t^{n+1}, Y(t^{n+1})\big) = f\big(t^n, Y(t^n)\big) + Ch$, and therefore

$$C = \frac{1}{h}\big[f\big(t^{n+1}, Y(t^{n+1})\big) - f\big(t^n, Y(t^n)\big)\big].$$

Consequently, the approximate solution takes in the interval $[t^n, t^{n+1}]$ the form

$$\begin{aligned} Y(t) =& Y(t^n) + f\big(t^n, Y(t^n)\big)(t - t^n) \\ &+ \frac{1}{2h}\big[f\big(t^{n+1}, Y(t^{n+1})\big) - f\big(t^n, Y(t^n)\big)\big](t - t^n)^2. \end{aligned} \tag{3.53}$$

For $t = t^{n+1}$, we thus have

$$Y(t^{n+1}) = Y(t^n) + \frac{h}{2}\big[f\big(t^n, Y(t^n)\big) + f\big(t^{n+1}, Y(t^{n+1})\big)\big]. \tag{3.54}$$

It is now obvious that the nodal approximations $Y(t^n)$ coincide with those of the trapezoidal method. Conversely, if we have computed the nodal approximations $Y(t^n)$ and $Y(t^{n+1})$ by the trapezoidal method, then we can recover Y in the interval $[t^n, t^{n+1}]$ from formula (3.53).

3.6 Absolute stability of RK methods

In this section we will study absolute stability properties of RK methods. More precisely, in the first subsection we will investigate issues related to A–stability and in the second problems related to B–stability.

3.6.1 Absolute stability and rational approximations to the exponential function

In Chapter 2 we introduced the A–stability and B–stability concepts for single-step methods. In this section we will study stability properties of RK methods.

Consider a Runge–Kutta method described by the tableau (3.6). Applying the method with step-size h to the test problem (1.18), that is to

$$\begin{cases} y' = \lambda y, \quad t > 0, \\ y(0) = 1, \end{cases}$$

we obtain approximations $(y^n)_{n\in\mathbb{N}}$, given by $y^0 := 1$ and, for $n \geq 0$, by the relations

$$y^{n,i} = y^n + h\lambda \sum_{j=1}^{q} a_{ij}\, y^{n,j}, \quad 1 \leq i \leq q, \tag{3.55}$$

$$y^{n+1} = y^n + h\lambda \sum_{i=1}^{q} b_i\, y^{n,i}\,. \tag{3.56}$$

We will now solve (3.55) for $y^{n,1}, \ldots, y^{n,q}$ and will substitute these values into (3.56), to express y^{n+1} in terms of y^n and $h\lambda$. If I denote the unit $q \times q$ matrix, then (3.55) can be written as

$$(I - h\lambda A) \begin{pmatrix} y^{n,1} \\ \vdots \\ y^{n,q} \end{pmatrix} = \begin{pmatrix} y^n \\ \vdots \\ y^n \end{pmatrix}.$$

Consequently, if the matrix $I - h\lambda A$ is invertible, we have

$$\begin{pmatrix} y^{n,1} \\ \vdots \\ y^{n,q} \end{pmatrix} = (I - h\lambda A)^{-1} \begin{pmatrix} y^n \\ \vdots \\ y^n \end{pmatrix}.$$

Therefore, with $\boldsymbol{e} \in \mathbb{R}^q$, $\boldsymbol{e} := (1, 1, \ldots, 1)^T$, we obtain

$$\sum_{i=1}^{q} b_i y^{n,i} = y^n \, b^T (I - h\lambda A)^{-1} \boldsymbol{e},$$

whence (3.56) gives

$$y^{n+1} = y^n \big[1 + h\lambda b^T (I - h\lambda A)^{-1} \boldsymbol{e}\big], \quad n \geq 0.$$

We now introduce the function

$$r(z) := 1 + z b^T (I - zA)^{-1} \boldsymbol{e}, \tag{3.57}$$

and write the previous relation as

$$y^{n+1} = r(h\lambda) y^n, \quad n \geq 0. \tag{3.58}$$

The function r in (3.57) is well defined for all complex numbers z, but the ones for which $I - zA$ is singular, i.e., those z for which $\frac{1}{z}$ is an eigenvalue of A, whence for at most q complex numbers. Expressing $w = (I - zA)^{-1} \boldsymbol{e}$, that is, the solution of the linear system

$$(I - zA) w = \boldsymbol{e},$$

by Cramer's rule, and substituting in (3.57), we immediately see that r is a *rational* function, and both the numerator and the denominator are polynomials of degree at most q.

From (3.58) and Definition 2.3 we immediately infer that the stability region S of a Runge–Kutta method consists of the points $z \in \mathbb{C}$, at which the corresponding rational function r satisfies the inequality $|r(z)| \leq 1$, i.e.,

$$S = \{z \in \mathbb{C} : \ |r(z)| \leq 1\}. \tag{3.59}$$

The rational function r is a *rational approximation* of the exponential function e^z. Indeed, the solution y of problem (1.18) satisfies the relation (with $t^k := kh,\ k \in \mathbb{N}_0$)

$$y(t^{n+1}) = \mathrm{e}^{h\lambda} y(t^n), \quad n \geq 0. \tag{3.60}$$

Relation (3.58) is a discrete analogue of (3.60), for the RK method (3.6). According to (3.58) and (3.60), for the consistency error E^n we have

$$E^n = \left[r(h\lambda) - \mathrm{e}^{h\lambda}\right] y(t^n), \quad n \geq 0. \tag{3.61}$$

If the order of the method is p, then, obviously,

$$\mathrm{e}^z - r(z) = O\left(|z|^{p+1}\right) \quad \text{for } z \to 0. \tag{3.62}$$

Relation (3.62) is thus a necessary condition, for the RK method to be of order p. It is not, however, in general, true that (3.62) implies that the order of a RK method, corresponding to the rational function r, is p. We easily convince ourselves for this fact, if we notice that the function r is independent of the intermediate nodes $\tau_i, i = 1, \ldots, q$. More precisely, (3.62) is a sufficient condition, for the order to be p, when the method is applied to homogeneous linear systems of o.d.e's with constant coefficients.

The rational approximations r corresponding to the explicit Euler method, the implicit Euler method, and the trapezoidal method are, respectively,

$$r_1(z) := 1 + z, \quad r_2(z) := \frac{1}{1-z}, \quad r_3(z) := \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}},$$

and satisfy, respectively, the inequalities $|r_1(z)| \leq 1$, for $|z+1| \leq 1$ (stability region) and $|r_i(z)| \leq 1$ for all $z \in \mathbb{C}$ with $\mathrm{Re}\, z \leq 0$, if $i = 2, 3$ (A–stable methods). Notice that $\mathrm{e}^z = r_i(z) + O(z^2)$, if $i = 1, 2$, while $\mathrm{e}^z = r_3(z) + O(z^3)$, in accordance to (3.62) and the known accuracy orders of these three methods.

Consider now an *explicit* Runge–Kutta method. Expressing in (3.55) the quantities $y^{n,i}$ in terms of y^n and $h\lambda$ and replacing the results in (3.56), we immediately see that the corresponding r in (3.57) is a polynomial of degree

at most q. According to (3.62), for an explicit $q-$stage RK method of order p we have

$$r(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^p}{p!} + c_{p+1}z^{p+1} + \cdots + c_q z^q.$$

We infer that there is no explicit A–stable RK method, that is explicit method such that $|r(z)| \leq 1$, for all $z \in \mathbb{C}$ with $\mathrm{Re}\, z \leq 0$. Here we have assumed consistency of the method, i.e., $p \geq 1$, in which case r can not be constant equal to one. Furthermore, from the above relation we see that the order p of an explicit $q-$stage RK method is at most q. As already mentioned in section 3.3, it can be shown that only for $q \leq 4$ there exist explicit RK methods of order $p = q$. In these cases, that is for explicit RK methods with $p = q$, the function r is of the form

$$r(z) = 1 + z + \frac{z^2}{2!} + \cdots + \frac{z^p}{p!},$$

that is the Taylor polynomial of degree p, around the point $z = 0$, of the exponential function. For example, the function $r(z) = 1+z+\frac{z^2}{2}$ corresponds to all explicit RK methods with $p = q = 2$. (Consequently, the same rational approximation r to the exponential function may correspond to two different RK methods.) The stability regions, that is the regions in the complex plane in which $|r(z)| \leq 1$ holds true, of the explicit RK methods with $p = q = 1, 2, 3, 4$, are the gray regions in Figure 3.1.

From the examples of implicit RK methods we saw in section 3.1, the midpoint method (Example 3) coincides, in the case of test problem (1.18), with the trapezoidal method, and is, consequently, A–stable with rational approximation to the exponential function r_3. The semiimplicit methods (3.11) give the rational functions

$$r(z) = \big(1 + (1-2\mu)z + (1/2 - 2\mu + \mu^2)z^2\big)/(1-\mu z)^2,$$

and are A–stable for appropriate μ (see Exercise 3.17). The $q-$stage Runge–Kutta Gauss–Legendre methods (method (3.12) in case q=2; see Corollary 3.1c for the general definition) have very interesting properties, as we know, because their order of accuracy is the highest possible ($p = 2q$) for a given stage

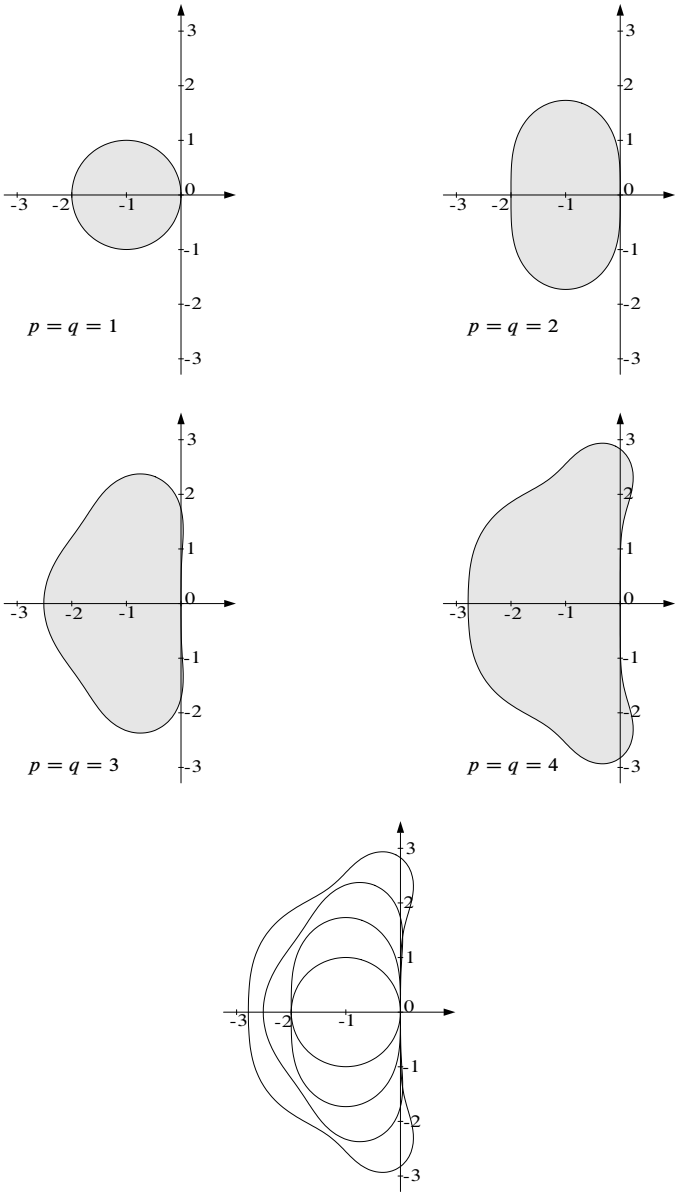

Figure 3.1: Stability regions in the complex plane of the explicit Runge–Kutta methods with q stages and order p, for $p = q = 1, 2, 3, 4$, for each case simultaneously as well as for all methods together.

number q. The rational function corresponding to (3.12) is

$$r(z) = \frac{1 + \frac{z}{2} + \frac{z^2}{12}}{1 - \frac{z}{2} + \frac{z^2}{12}},$$

and we infer that (3.12) is A–stable.

In the case of implicit Runge–Kutta methods, the rational function r is usually an element of the so-called *Padé 'table'* for the exponential function. Let $\ell, m \in \mathbb{N}_0$. We say that a rational function $\frac{P}{Q}$, with P, Q polynomials of degree at most ℓ and m, respectively, is a *Padé approximation* to the exponential function, if

$$\mathrm{e}^z - \frac{P(z)}{Q(z)} = O\big(|z|^{m+\ell+1}\big) \quad \text{as } z \to 0. \tag{3.63}$$

Let us first show, for any given pair (ℓ, m), uniqueness of the corresponding Padé approximation $\frac{P}{Q}$ to the exponential function. Indeed, if $\frac{\widetilde{P}}{\widetilde{Q}}$ was another Padé approximation corresponding to the same pair (ℓ, m), then we would have

$$\frac{P(z)}{Q(z)} - \frac{\widetilde{P}(z)}{\widetilde{Q}(z)} = O\big(|z|^{m+\ell+1}\big) \quad \text{as } z \to 0,$$

i.e.,

$$P(z)\widetilde{Q}(z) - \widetilde{P}(z)Q(z) = O\big(|z|^{m+\ell+1}\big) \quad \text{as } z \to 0.$$

Since the function on the left-hand side is a polynomial of degree at most $\ell+m$, we infer that it vanishes identically, i.e., $\frac{P}{Q} = \frac{\widetilde{P}}{\widetilde{Q}}$.

The Padé approximations to the exponential function are known in closed form; more precisely we have

$$\begin{cases} P(z) = \displaystyle\sum_{j=0}^{\ell} \frac{\ell!\,(\ell+m-j)!}{j!\,(\ell-j)!}\, z^j \\[2ex] Q(z) = \displaystyle\sum_{j=0}^{m} \frac{m!\,(\ell+m-j)!}{j!\,(m-j)!}\, (-z)^j. \end{cases} \tag{3.64}$$

Proposition 3.4 (Padé approximation to the exponential function.) *For the polynomials P and Q defined in* (3.64) *there holds*

$$1 + z + \cdots + \frac{z^{\ell+m}}{(\ell+m)!} - \frac{P(z)}{Q(z)} = O\big(|z|^{\ell+m+1}\big) \quad \text{as } z \to 0. \tag{3.65}$$

Proof. Since the constant coefficient of Q is not zero, it suffices to show that

$$\left[1 + z + \cdots + \frac{z^{\ell+m}}{(\ell+m)!}\right] Q(z) - P(z) = O\big(|z|^{\ell+m+1}\big) \quad \text{as } z \to 0. \tag{3.66}$$

Consequently, it suffices to show that the coefficients of the monomials z^j, $j = 0, \ldots, \ell+m$, in the left-hand side of (3.66) vanish. Let α_j be the coefficient of z^j of the product $\left[1 + z + \cdots + \frac{z^{\ell+m}}{(\ell+m)!}\right] Q(z)$. For $j = 0, \ldots, \ell+m$, and with $M := \min(j, m)$, we obviously have

$$\alpha_j = \sum_{i=0}^{M} \frac{m!\,(\ell+m-i)!}{i!\,(m-i)!}\,(-1)^i\,\frac{1}{(j-i)!},$$

i.e.,

$$\alpha_j = \frac{\ell!\,m!}{j!} \sum_{i=0}^{M} \binom{\ell+m-i}{m-i} (-1)^i \binom{j}{i}. \tag{3.67}$$

We recall that, for $\alpha \in \mathbb{R}$ and $n \in \mathbb{N}_0$, $\binom{\alpha}{n}$ is the coefficient of x^n of the Taylor expansion around zero of the function $(1+x)^\alpha$, and that there holds

$$\binom{\alpha}{n} = \frac{\alpha\,(\alpha-1)\cdots(\alpha-n+1)}{n!}\,.$$

To avoid splitting the proof in various cases, we *furthermore* set, for $\alpha \in \mathbb{R}$ and $-n \in \mathbb{N}$, $\binom{\alpha}{n} := 0$. Then (3.67) yields

$$\alpha_j = \frac{\ell!\,m!}{j!} \sum_{i=0}^{j} \binom{\ell+m-i}{m-i} (-1)^i \binom{j}{i}$$

or

$$\alpha_j = (-1)^m\,\frac{\ell!\,m!}{j!} \sum_{i=0}^{j} \binom{-\ell-1}{m-i} \binom{j}{i}.$$

Comparing the coefficients of x^m in both sides of the identity

$$(1+x)^{-\ell-1}\,(1+x)^j = (1+x)^{j-\ell-1}$$

we obtain

$$\sum_{i=0}^{j} \binom{-\ell-1}{m-i}\binom{j}{i} = \binom{j-\ell-1}{m} = (-1)^m \binom{\ell+m-j}{m},$$

i.e.,

$$\alpha_j = \frac{\ell!m!}{j!}\binom{\ell+m-j}{m}. \tag{3.68}$$

The desired result follows from (3.68) and the first relation in (3.64). □

The first entries of the Padé table for the exponential function are given in Table 3.1.

m \ ℓ	0	1	2	$\cdots$
0	1	$1+z$	$1+z+\dfrac{z^2}{2}$	$\cdots$
1	$\dfrac{1}{1-z}$	$\dfrac{1+\frac{z}{2}}{1-\frac{z}{2}}$	$\dfrac{1+\frac{2}{3}z+\frac{z^2}{6}}{1-\frac{z}{3}}$	$\cdots$
2	$\dfrac{1}{1-z+\frac{z^2}{2}}$	$\dfrac{1+\frac{z}{3}}{1-\frac{2}{3}z+\frac{1}{6}z^2}$	$\dfrac{1+\frac{z}{2}+\frac{z^2}{12}}{1-\frac{z}{2}+\frac{z^2}{12}}$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table 3.1: Padé table for the exponential function.

Notice that to the explicit Euler method, the implicit Euler method and the trapezoidal method, respectively, correspond the pairs $(\ell, m) = (0,1), (1,0)$ and $(1,1)$, respectively, of the Padé table. The two-stage RK Gauss–Legendre method, (3.12), corresponds to the entry $(2,2)$. In general, to the $q-$stage RK

Gauss–Legendre method corresponds the diagonal (q,q) entry of the Padé table of the exponential function, as we immediately infer, if we take into account the fact that this is the only $q-$stage method of order $2q$. In this case we have $\ell = m$, και $\lim_{|z|\to\infty} \left|r(z)\right| = 1$. Since $P(\mathrm{i}y) = Q(-\mathrm{i}y)$ for $y \in \mathbb{R}$, there holds

$$\forall\, y \in \mathbb{R} \quad \left|r(\mathrm{i}y)\right| = 1. \tag{3.69}$$

From this relation, the fact that the denominator of the diagonal entries of the Padé table does not vanish for $z \in \mathbb{C}$ with $\mathrm{Re}\, z < 0$ (which we do not prove here), and the maximum principle for analytic functions, we infer that

$$\forall z \in \mathbb{C} \quad \mathrm{Re}\, z < 0 \quad \left|r(z)\right| < 1,$$

i.e., all RK Gauss–Legendre methods are A–stable.

More generally, it can be shown that a RK method is A–stable, if and only if the corresponding rational function r does not have poles with negative real part, is bounded at infinity and satisfies the inequality $|r(\mathrm{i}y)| \leq 1$, for all $y \in \mathbb{R}$; see also (3.69). The proof uses again the maximum principle for analytic functions of a complex variable. In particular, r does not have poles with negative real part, if the eigenvalues of the matrix $A = (a_{ij})$ have nonnegative real parts. Notice also that the rational function r can be written in the form

$$r(z) = \frac{\det(I - zA + z\boldsymbol{e}b^T)}{\det(I - zA)};$$

in particular, a complex number z may be a pole of r, only if $1/z$ is an eigenvalue of the matrix A. Furthermore, if the rational function r is an element of the Padé table of the exponential function (which is usually the case), it is known that the corresponding RK method is A–stable, if and only if the degree of the denominator is equal to the degree of the numerator or exceeds it by one or by two, i.e., if r is an element of the diagonal, the below diagonal, or the second below diagonal of the Padé table.

A–stable RK methods are more generally suitable for linear systems of first order o.d.e's with constant coefficients. If we discretize initial value problems for systems of linear o.d.e's of the form $y' = My + g(t)$, with an $m \times m$

negative semidefinite real matrix M, then for the difference of approximations $y^\ell - z^\ell$ we have

$$y^{n+1} - z^{n+1} = r(hM)(y^n - z^n); \tag{3.70}$$

see Exercise 3.19. If the method is A–stable, then according to the von Neumann theorem, see Remark 2.7, we have $\|r(hM)\| \leq 1$, and infer that

$$\|y^{n+1} - z^{n+1}\| \leq \|y^n - z^n\|; \tag{3.71}$$

see Remark 2.7 for details. For an elementary proof of the estimate (3.71), in the case of a symmetric matrix M, we refer to Exercise 3.20.

3.6.2 B–stability

In Proposition 3.5 we give a sufficient condition for the B–stability of a RK method. This proposition is particularly useful because it can be used to establish error estimates for initial values problems with f satisfying the one-sided Lipschitz condition; see Exercise 3.29. But first we need a definition.

Definition 3.2 (Algebraic stability.) A Runge–Kutta method, with Butcher tableau (3.2), is called *algebraically stable*, if it satisfies the conditions:

$$\begin{cases} b_i \geq 0, \quad 1 \leq i \leq q. \\ \text{The symmetric } q \times q \text{ matrix with entries } m_{ij}, \\ \qquad m_{ij} := b_i a_{ij} + b_j a_{ji} - b_i b_j, \quad 1 \leq i, j \leq q, \\ \text{is positive semidefinite, } \displaystyle\sum_{i,j=1}^{q} m_{ij} x_i x_j \geq 0 \ \ \forall x \in \mathbb{R}^q. \end{cases} \tag{3.72}$$

Notice that this kind of stability can be checked by algebraic means, and this is why it is called algebraic; furthermore, the quantities $\tau_1, \ldots, \tau_q$ are irrelevant here. As we will see in the sequel, the algebraic stability implies B–stability. It is also known (Butcher, Crouzeix) that if the intermediate nodes $\tau_i, i = 1, \ldots, q$, are pairwise distinct, then the algebraic stability and the B–stability are equivalent.

Proposition 3.5 (Algebraic stability implies B–stability.) *An algebraically stable Runge–Kutta method is B–stable.*

Proof. Let $\{y^n, y^{n,i}\}, \{z^n, z^{n,i}\}$ be solutions of (3.6) and (3.20). Subtracting, we have

$$\begin{cases} y^{n,i} - z^{n,i} = y^n - z^n + \sum_{j=1}^{q} a_{ij}\varphi^j, & 1 \le i \le q, \\ y^{n+1} - z^{n+1} = y^n - z^n + \sum_{i=1}^{q} b_i\varphi^i, \end{cases} \tag{3.73}$$

with $\varphi^j := h\big[f(t^{n,j}, y^{n,j}) - f(t^{n,j}, z^{n,j})\big]$, $j = 1, \dots, q$. Taking here the Euclidean inner product of each member of the second relation of (3.73) by itself, we obtain

$$\|y^{n+1} - z^{n+1}\|^2 = \|y^n - z^n\|^2 + 2\sum_{i=1}^{q} b_i(\varphi^i, y^n - z^n) + \Big\|\sum_{i=1}^{q} b_i\varphi^i\Big\|^2. \tag{3.74}$$

Using the first relations of (3.73) we get

$$\sum_{i=1}^{q} b_i(\varphi^i, y^n - z^n) = \sum_{i=1}^{q} b_i(\varphi^i, y^{n,i} - z^{n,i}) - \sum_{i,j=1}^{q} b_i a_{ij}(\varphi^i, \varphi^j).$$

Consequently, (3.74) can be written as

$$\begin{aligned} \|y^{n+1} - z^{n+1}\|^2 = \|y^n - z^n\|^2 &+ 2\sum_{i=1}^{q} b_i(\varphi^i, y^{n,i} - z^{n,i}) \\ &+ \sum_{i,j=1}^{q} b_i b_j(\varphi^i, \varphi^j) - 2\sum_{i,j=1}^{q} b_i a_{ij}(\varphi^i, \varphi^j). \end{aligned} \tag{3.75}$$

But, due to the one-sided Lipschitz condition (1.24), for each $i, 1 \le i \le q$, we have

$$(\varphi^i, y^{n,i} - z^{n,i}) = h\big(f(t^{n,i}, y^{n,i}) - f(t^{n,i}, z^{n,i}), y^{n,i} - z^{n,i}\big) \le 0.$$

Furthermore, due to the symmetry of (φ^i, φ^j),

$$2\sum_{i,j=1}^{q} b_i a_{ij}(\varphi^i, \varphi^j) = \sum_{i,j=1}^{q} (b_i a_{ij} + b_j a_{ji})(\varphi^i, \varphi^j).$$

Therefore, we obtain from (3.75)

$$\|y^{n+1} - z^{n+1}\|^2 \le \|y^n - z^n\|^2 - \sum_{i,j=1}^{q} (b_i a_{ij} + b_j a_{ji} - b_i b_j)(\varphi^i, \varphi^j),$$

i.e.,

$$\|y^{n+1} - z^{n+1}\|^2 \le \|y^n - z^n\|^2 - \sum_{i,j=1}^{q} m_{ij}(\varphi^i, \varphi^j). \tag{3.76}$$

Now, if $\{e^1, \dots, e^m\}$ is the canonical basis of $\mathbb{R}^m$, then

$$\varphi^i = \sum_{\ell=1}^{m} \alpha_{i\ell} e^\ell$$

and we have

$$(\varphi^i, \varphi^j) = \sum_{\ell=1}^{m} \alpha_{i\ell}\alpha_{j\ell},$$

whence

$$\sum_{i,j=1}^{q} m_{ij}(\varphi^i, \varphi^j) = \sum_{\ell=1}^{m} \Big(\sum_{i,j=1}^{q} m_{ij}\alpha_{i\ell}\alpha_{j\ell} \Big).$$

But, in view of conditions (3.72), we have

$$\sum_{i,j=1}^{q} m_{ij}\alpha_{i\ell}\alpha_{j\ell} \ge 0, \quad \ell = 1, \dots, m,$$

and (3.76) yields the desired estimate

$$\|y^{n+1} - z^{n+1}\|^2 \le \|y^n - z^n\|^2. \qquad \square$$

Let us now see some examples of algebraically stable and consequently B–stable methods. For the implicit Euler method, we have $q = 1, b_1 = 1 > 0$

και $m_{11} = 1 + 1 - 1 = 1 \geq 0$, whence the method is algebraically stable. Similarly, for the (implicit) midpoint method we have $q = 1, b_1 = 1 > 0$ και $m_{11} = \frac{1}{2} + \frac{1}{2} - 1 = 0$, whence this method is also algebraically stable. Furthermore, the DIRK methods (3.11) for $\mu \geq 1/4$, and all RK Gauss–Legendre methods are algebraically stable, whence also B–stable. (Actually, for the Gauss–Legendre methods we have $m_{ij} = 0$.) See also Exercise 3.24.

For the trapezoidal method, on the other hand, we have $m_{11} = -0.25 < 0$, and this information is sufficient to infer that this method is *not* algebraically stable. Indeed, the method is not even B–stable; see Remark 2.6 for a direct proof.

Remark 3.3 (B–stability of the RK Gauss–Legendre methods.) We have already mentioned, without a complete proof, that the RK Gauss–Legendre methods are A–stable. As an application of the equivalence of collocation methods to RK methods, we prove here that RK Gauss–Legendre methods are B–stable. This simple proof was given by Wanner; see [17, Example 12.3]. Let, now, $q \in \mathbb{N}, \tau_1, \ldots, \tau_q \in (0,1)$ be the nodes and $b_1, \ldots, b_q$ the corresponding (positive) weights of the Gauss quadrature formula with weight function $w, w(x) = 1, x \in [0,1]$. We consider the initial value problems (1.19) with right-hand side f satisfying the one-sided Lipschitz condition (1.24), and, with the usual notation, for given approximations $Y(t^n)$ and $Z(t^n)$, we consider the approximations Y and Z that the collocation method with these nodes yield. We set

$$m(t) := \|Y(t) - Z(t)\|^2, \quad t \in [t^n, t^{n+1}],$$

and notice that function m is a polynomial of degree at most $2q$. Furthermore, obviously, $m'(t) = 2\big(Y'(t) - Z'(t), Y(t) - Z(t)\big)$, whence, in view of (1.24),

$$m'(t^{n,i}) = 2\big(f\big(t^{n,i}, Y(t^{n,i})\big) - f\big(t^{n,i}, Z(t^{n,i})\big), Y(t^{n,i}) - Z(t^{n,i})\big) \leq 0.$$

Now,

$$\|Y(t^{n+1}) - Z(t^{n+1})\|^2 = m(t^{n+1}) = m(t^n) + \int_{t^n}^{t^{n+1}} m'(t)\,dt.$$

But, the polynomial $m' \in \mathbb{P}_{2q-1}$ is integrated exactly by the Gauss formula, whence

$$\int_{t^n}^{t^{n+1}} m'(t)\,dt = h\sum_{i=1}^{q} b_i m'(t^{n,i}) \leq 0,$$

and thus

$$\|Y(t^{n+1}) - Z(t^{n+1})\|^2 \leq m(t^n) = \|Y(t^n) - Z(t^n)\|^2,$$

i.e.,

$$\|Y(t^{n+1}) - Z(t^{n+1})\| \leq \|Y(t^n) - Z(t^n)\|. \qquad \square$$

Remark 3.4 (RK Radau IIA methods.) Let $q \geq 1$ and $\tau_q = 1$. It is well known from the theory of numerical integration that for exactly one choice of nodes $\tau_1, \ldots, \tau_{q-1}$ there are weights $b_1, \ldots, b_q$, such that the quadrature rule Q_q,

$$Q_q(f) = \sum_{i=1}^{q} b_i f(\tau_i),$$

integrates in $[0, 1]$ polynomials of degree up to $2q - 2$ exactly, i.e.,

$$\forall\, p \in \mathbb{P}_{2q-2} \quad \int_0^1 p(\tau)\,d\tau = Q_q(p).$$

The rule is called *Radau quadrature rule*. The weights $b_1, \ldots, b_q$ are positive and the nodes $\tau_1, \ldots, \tau_{q-1}$ belong to $(0, 1)$.

We consider now the collocation method with nodes $\tau_1, \ldots, \tau_q$. According to our previous discussion and section 3.5, the order of accuracy of the method is $2q - 1$. The RK method corresponding to this collocation method is called *RK Radau IIA method.* The first member of this family, i.e., for $q = 1$, is the implicit Euler method; for the second member we refer to Exercise 3.25. The corresponding rational approximation to the exponential of the $q-$stage RK Radau IIA method is the entry in $(q+1, q)$ of the Padé table of the exponential function. The methods of this family have very good stability properties and are widely used in applications. It can be easily shown that the RK Radau IIA methods are B–stable; see Exercise 3.37. $\square$

The interested reader can find more details about the stability properties of Runge–Kutta methods in the monographs [5], [17] and [10].

Exercises

3.1 Consider the $m \times m$ system of o.d.e's (3.8) and suppose that the function f satisfies the global Lipschitz condition

$$\exists L \geq 0 \ \ \forall t \in [a,b] \ \ \forall y_1, y_2 \in \mathbb{R}^m \quad \|f(t,y_1) - f(t,y_2)\|_\infty \leq L\|y_1 - y_2\|_\infty$$

(which, in view of the equivalence of norms in $\mathbb{R}^m$, is then valid also with respect to any other norm, with different Lipschitz constant $\widetilde{L}$). Assume that $\gamma h < 1$, with γ as in Proposition 3.1. Prove that system (3.4) possesses a unique solution $y^{n,1}, \ldots, y^{n,q}$. Moreover, prove an analogue of (3.23), i.e., if $z^n, z^{n,i} \in \mathbb{R}^m$ satisfy the relations (3.22) with given $z^0, \rho^n \in \mathbb{R}^m$, then

$$\max_{1\leq n\leq N} \|y^n - z^n\|_\infty \leq C_1\|y^0 - z^0\|_\infty + \frac{C_2}{h} \max_{0\leq n\leq N-1} \|\rho^n\|_\infty,$$

with C_1 and C_2 the constants introduced in the proof of Proposition 3.2. Deduce that the RK methods are stable also or systems, i.e., for $\rho^n = 0$ in (3.22), and any norm $\|\cdot\|$ in $\mathbb{R}^m$ there exists a constant C, independent of h, such that

$$\max_{1\leq n\leq N} \|y^n - z^n\| \leq C\,\|y^0 - z^0\|.$$

3.2 Consider the initial value problem (3.1) and assume that f is sufficiently regular in $[a,b] \times \mathbb{R}$, and that f and appropriate partial derivatives of it are bounded. Determine all Runge–Kutta methods of the form

$$\begin{array}{cc|c} 0 & 0 & 0 \\ a_{21} & 0 & \tau_2 \\ \hline b_1 & b_2 & \end{array},$$

with order of accuracy (at least) $p = 2$.

3.3 Under the assumptions of Exercise 3.2 prove that the order of accuracy of the explicit RK methods described by the tableaus (3.13) and (3.14) is three.
[*Hint*: With the usual notation, show first for the method (3.13) that

$$\zeta^{n,2} = y(t^n + \frac{h}{2}) - \frac{h^2}{8}y''(t^n) + O(h^3),$$
$$\zeta^{n,3} = y(t^{n+1}) + \frac{h^2}{2}y''(t^n) + O(h^3).$$

Analogously, show for the method (3.14) that

$$\zeta^{n,3} = y(t^n + \frac{2h}{3}) + O(h^3).$$

Consider also the particular initial value problem

$$\begin{cases} y' = y, \quad 0 \le t \le 1, \\ y(0) = 1.] \end{cases}$$

3.4 Under the assumptions of Exercise 3.2 show that the order of accuracy of the ("classical") explicit RK method (3.16) is four.

3.5 Prove that the order of accuracy of the explicit midpoint method for systems of o.d.e's of the form (3.8), with $f : [a, b] \times \mathbb{R}^m \to \mathbb{R}^m$ sufficiently regular, is two.

3.6 Prove that the only one-stage RK method of order two is the (implicit) midpoint method (Example 3 is section 3.1).

3.7 Prove that the order of all two-stage Runge–Kutta methods, different from (3.12), satisfying $\sum_{j=1}^{2} a_{ij} = \tau_i$, $i = 1, 2$, is at most three.
[*Hint*: Apply the method to the initial value problems

$$\begin{cases} y' = f(t, y), \quad 0 \le t \le 1, \\ y(0) = y_0 \end{cases}$$

with $f(t, y) = it^{i-1}$, $i = 1, 2, 3, 4$, and $y_0 = 0$, as well as with $f(t, y) = y$ and $y_0 = 1$. Utilize the fact that the only quadrature rule with two nodes that integrates the functions $q_i(t) := t^i, i = 0, 1, 2, 3$, exactly (i.e., polynomials of degree at most three exactly), is the Gauss rule.]

3.8 Show that the order of accuracy of a $q-$stage Runge–Kutta method is at most $2q$.

3.9 Show that a sufficient and necessary condition for the consistentcy of a Runge–Kutta method (3.2), for problem (3.1), is $b_1 + b_2 + \cdots + b_q = 1$.

3.10 Consider the initial value problem

$$\begin{cases} y'(t) = 1, \quad 0 \le t \le 1, \\ y(0) = 0. \end{cases}$$

Let $N \in \mathbb{N}$, $h := \frac{1}{N}$, and y^N be the approximation of the solution at the point 1, that a Runge–Kutta method yields when applied to our problem with step-size h. Assuming that

$$y^N \to 1 = y(1), \quad N \to \infty,$$

show that the Runge–Kutta method is consistent.
[*Hint*: Use Exercise 3.9.]

3.11 Consider the quantities $\zeta^{n,i}$ defined in (3.27). Show, for the initial value problem (3.1), under the hypotheses of Exercise 3.2, that

$$\max_{n,i} \left| y(t^{n,i}) - \zeta^{n,i} \right| \leq C h$$

and

$$\max_{n} \left| y(t^{n,i}) - \zeta^{n,i} \right| \leq C h^2 \iff \sum_{j=1}^{q} a_{ij} = \tau_i, \quad i = 1, \ldots, q.$$

3.12 For each one of the RK methods of the Examples 1–9 of section 3.1 compare their (real) order of accuracy with the order p that the (sufficient) conditions of Theorem 3.2 and of a) and b) of Corollary 3.1 yield, to get an idea of the power of these conditions for common RK methods.

3.13 Let $a = t^0 < t^1 < \cdots < t^N = b$ be a partition of $[a, b]$, $h_n := t^{n+1} - t^n$, $n = 0, \ldots, N-1$, and $h := \max_{0 \leq n \leq N-1} h_n$. If the solution of the initial value problem

$$\begin{cases} y' = f(t, y), & a \leq t \leq b, \\ y(a) = y_0 \end{cases}$$

is sufficiently smooth, then for the error $\varepsilon^n := y(t^n) - y^n$, $n = 0, \ldots, N$, of a Runge–Kutta method of order p there holds

$$|\varepsilon^{n+1}| \leq (1 + C_1 h_n) |\varepsilon^n| + C_2 h_n^{p+1}, \quad n = 0, \ldots, N-1,$$

with constants C_1 and C_2, independent of N and of the partition. Suppose that this is true, and prove that there exists a constant C, independent of h, such that

$$\max_{0 \leq n \leq N} |\varepsilon^n| \leq C h^p.$$

3.14 (Explicit RK methods as collocation methods.)

a) Consider the collocation method with $q = 1, \tau_1 = 0$. Show that this method is equivalent to the explicit Euler method, in the sense that both methods yield the same nodal approximations.

b) Consider an *explicit* $q-$stage RK method, with $q \geq 2$. Show that in (3.38) we necessarily have $s \leq 1$; in particular, these methods are not equivalent to collocation methods; see the relevant discussion subsequent to the proof of Proposition 3.3.

[*Hint*: In order that (3.38) be satisfied with $s \geq 1$ for $i = 1$, it is necessary that $\tau_1 = 0$. Show now that (3.38) can not be satisfied with $s \geq 2$, for any choice of a_{21}, for $i = 2$ with $\tau_2 \neq 0$.]

3.15 With the notation of section 3.5, consider the collocation method with q nodes such that $\tau_q = 1$. Show that for the corresponding Runge–Kutta method there holds $y^{n,q} = y^{n+1}$.

3.16 Let $q \in N$ and $\tau_1, \ldots, \tau_q \in (0,1)$ be the nodes of the Gauss quadrature rule in the interval $[0,1]$, with weight function $w(x) = 1, x \in [0,1]$. Consider the collocation method with these nodes. If the solution is smooth enough, what is the order of the approximations at the nodes of a uniform partition and what globally in the whole interval $[a,b]$? Which Runge–Kutta method corresponds to this collocation method?

3.17 Consider the semiimplicit RK methods (3.11) with a parameter $\mu \in \mathbb{R}$. For which values of μ are the methods A–stable? What is the order of accuracy p that (3.62) yields in this case?

3.18 Prove that the method (3.12) is A–stable. Expanding the corresponding rational approximation to the exponential r in power series of z, determine the order of accuracy p that (3.62) yields for this method.

3.19 With the standard notation, see in particular subsection 3.6.1, show that the step $y^n \mapsto y^{n+1}$ of a Runge–Kutta method applied to inhomogeneous linear o.d.e's with constant coefficients, $y' = \lambda y + f(t)$, can be written in the form

$$y^{n+1} = r(h\lambda)y^n + hb^T\left(I + h\lambda A(I - h\lambda A)^{-1}\right)\begin{pmatrix} f(t^{n,1}) \\ \vdots \\ f(t^{n,q}) \end{pmatrix}.$$

3.20 Let r be the rational approximation to the exponential of an A–stable RK method. Consider the system of o.d.e's $y' = My + g(t)$ with a symmetric and negative semidefinite $m \times m$ matrix M. With the standard notation, discretize the problem

by the RK method and consider two sequences of approximations y^n and z^n. Express the vectors $y^n - z^n$ and $y^{n,i} - z^{n,i}$ as expansions of the eigenvectors of M, and prove that $y^{n+1} - z^{n+1} = r(hM)(y^n - z^n)$, and, consequently, that

$$\|y^n - z^n\|_2 \leq \|r(hM)\|_2^n \, \|y^0 - z^0\|_2, \quad n \geq 0.$$

Infer, in view of Exercise 1.17a, that, since the method is A–stable, there holds

$$\|r(hM)\|_2 = \max_{1 \leq i \leq m} \left|r(h\lambda_i)\right| \leq 1 \quad \forall h > 0,$$

and thus

$$\|y^n - z^n\|_2 \leq \|y^0 - z^0\|_2, \quad n \geq 0,$$

i.e., a discrete analogue of the result of Exercise 1.17b holds true in this case.

3.21 Under the assumptions of Exercise 1.17, and supposing that the rational function r of a RK method satisfies the condition

$$|r(x)| \leq 1 \quad \forall x \in [-\alpha, 0],$$

for some positive constant α, prove the stability inequality $\|y^n - z^n\|_2 \leq \|y^0 - z^0\|_2$, provided $h \max_{1 \leq i \leq m} |\lambda_i| \leq \alpha$.

3.22 We approximate the solution of a system of o.d.e's (3.8) by the A–stable (see Exercise 3.17) semiimplicit RK method (3.11). At every time level, to advance in time we need to solve two linear systems. What do you observe corcerning these two systems?

3.23 As a prototype of systems of o.d.e's, with matrix M with *imaginary* eigenvalues (such systems arise from the discretization of hyperbolic p.d.e's, of oscillation problems without attenuation etc.), we consider the initial value problem of Exercise 2.11. Prove that the $q-$stage RK Gauss–Legendre methods are particularly suitable for such systems, in the sense that they yield approximations y^n such that $|y^n| = 1$, for all $n \geq 0$.

3.24 a) Prove that the semiimplicit RK methods (3.11) are B–stable for $\mu \geq 1/4$.
b) Prove that the two-stage RK Gauss–Legendre method (3.12) is B–stable, with $m_{ij} = 0$.
c) Show that the method described by the tableau

$$\begin{array}{cc|c} \frac{1}{8} & \frac{1}{8} & \frac{1}{4} \\ \frac{3}{8} & \frac{3}{8} & \frac{3}{4} \\ \hline \frac{1}{2} & \frac{1}{2} & \end{array}$$

is A–stable, but does not satisfy the algebraic stability conditions (3.72). (It can be shown that the method is not B–stable.)

3.25 Use Corollary 3.1 to prove that the order of accuracy of the Runge–Kutta–Radau method with tableau

$$\begin{array}{cc|c} \frac{5}{12} & -\frac{1}{12} & \frac{1}{3} \\ \frac{3}{4} & \frac{1}{4} & 1 \\ \hline \frac{3}{4} & \frac{1}{4} & \end{array}$$

is three.

3.26 Prove that the Runge–Kutta–Radau method of Exercise 3.25 is algebraically stable.
[*Hint*: With the notation of Definition 3.2 check that $m_{11} = m_{22} = 1/16$ και $m_{12} = m_{21} = -1/16$, whence

$$\sum_{i,j=1}^{2} m_{ij} x_i x_j = \frac{1}{16} (x_1 - x_2)^2 .]$$

3.27 Prove that the RK method of Exercise 3.25, see Proposition 3.3, corresponds to the collocation method with collocation nodes $\tau_1 = 1/3$ and $\tau_2 = 1$. Prove that

$$\int_0^1 (\tau - \frac{1}{3})(\tau - 1) r(\tau)\, d\tau = 0 \quad \forall r \in \mathbb{P}_0$$

and infer again that the order of accuracy of the RK method is three.

3.28 α) Prove that the RK method with tableau

$$\begin{array}{cc|c} 0 & 0 & 0 \\ \frac{\tau_2}{2} & \frac{\tau_2}{2} & \tau_2 \\ \hline 1 - \frac{1}{2\tau_2} & \frac{1}{2\tau_2} & \end{array}$$

is A–stable, if and only if $\tau_2 = 1$, i.e., if and only if it coincides with the trapezoidal method.
[*Hint*: Prove that the corresponding rational approximation to the exponential r is

$$r(z) = \frac{2 + (2 - \tau_2)z + (1 - \tau_2)z^2}{2 - \tau_2 z}.\,]$$

β) Consider the collocation method with collocation nodes $\tau_1 = 0$ and $0 < \tau_2 \leq 1$. Prove that the method is A–stable, if and only if $\tau_2 = 1$.
[*Hint*: Prove that the corresponding RK method is the one given in part a).]

3.29 (Algebraically stable Runge–Kutta methods: Error estimate independent of the Lipschitz constant.) We consider the initial value problem

$$\begin{cases} y' = f(t, y), \quad a \leq t \leq b, \\ y(a) = y_0, \end{cases}$$

and assume that f satisfies the condition

$$\big(f(t, y) - f(t, z)\big)(y - z) \leq 0, \quad a \leq t \leq b,\ y, z \in \mathbb{R}.$$

We discretize this problem by an algebraically stable Runge–Kutta method of order p, described by the tableau

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

assuming a uniform partition of the interval $[a, b]$ with step-size h. Assume that the solution is sufficiently regular and prove the error estimate

$$\max_{0 \leq n \leq N} |y(t^n) - y^n| \leq C h^p,$$

with a constant C independent of h.
[*Hint*: With the usual notation, prove first that

$$\Big(y^{n+1} - y(t^{n+1}) - E^n\Big)^2 \leq \big(y^n - y(t^n)\big)^2,$$

i.e.,

$$|y(t^{n+1}) - y^{n+1}| \leq |y(t^n) - y^n| + |E^n|\,.]$$

3.30 (In Proposition 3.1 we proved that the Runge–Kutta approximations are, for sufficiently small step-size h, well defined. As we will see in this Exercise, and in the next three, under some conditions, the approximations are well defined, without any conditions on h or under much milder conditions than the ones in Proposition 3.1.)

We consider a Runge–Kutta method, described by a tableau

$$\begin{array}{ccc|c} a_{11} & \dots & a_{1q} & \tau_1 \\ \vdots & & \vdots & \vdots \\ a_{q1} & \dots & a_{qq} & \tau_q \\ \hline b_1 & \dots & b_q & \end{array},$$

and assume that $\tau_1, \dots, \tau_q \in [0,1]$ and that the matrix $A = \big(a_{ij}\big)_{i,j=1,\dots,n}$ is positive definite. Prove that the RK approximations are well defined, for any h, for initial value problems

$$\begin{cases} y' = f(t,y), \quad a \le t \le b, \\ y(a) = y_0, \end{cases}$$

with $f : [a,b] \times \mathbb{R} \to \mathbb{R}$ a continuous function satisfying the one-sided Lipschitz condition, with respect to its second variable,

$$\big(f(t,y) - f(t,z)\big)(y-z) \le 0, \quad a \le t \le b,\ y,z \in \mathbb{R}.$$

[*Hint*: Obviously, the matrix A^{-1} is also positive definite; thus

$$(A^{-1}x, x) \ge c_1 \|x\|^2 \quad \forall x \in \mathbb{R}^q,$$

for an appropriate positive constant c_1. Let $h > 0$ and $y^n \in \mathbb{R}$. It suffices to show that the nonlinear, in general, system

$$(\star) \qquad y^{n,i} = y^n + h \sum_{j=1}^{q} a_{ij} f(t^{n,j}, y^{n,j}), \quad i = 1, \dots, q,$$

is uniquely solvable. With $Y := (y^{n,1}, \dots, y^{n,q})^T$ και $\boldsymbol{e} := (1, \dots, 1)^T \in \mathbb{R}^q$, we first write $(\star)$ in the form

$$A^{-1}(Y - y^n \boldsymbol{e}) = h \begin{pmatrix} f(t^{n,1}, y^{n,1}) \\ \vdots \\ f(t^{n,q}, y^{n,q}) \end{pmatrix}.$$

Now, the uniqueness is easily shown. For the existence, consider the mapping $G : \mathbb{R}^q \to \mathbb{R}^q$,

$$G(x) := A^{-1}(x - y^n \boldsymbol{e}) - h \begin{pmatrix} f(t^{n,1}, x_1) \\ \vdots \\ f(t^{n,q}, x_q) \end{pmatrix},$$

with $x = (x_1, \ldots, x_q)^T$. Then

$$\begin{aligned} \big(G(x), x\big) &= (A^{-1}x, x) - y^n(A^{-1}\boldsymbol{e}, x) \\ &\quad - h\sum_{i=1}^{q} \big[f(t^{n,i}, x_i) - f(t^{n,i}, 0)\big]x_i - h\sum_{i=1}^{q} f(t^{n,i}, 0)x_i. \end{aligned}$$

Using our assumption on f, we infer that

$$\big(G(x), x\big) \geq (A^{-1}x, x) - y^n(A^{-1}\boldsymbol{e}, x) - h\sum_{i=1}^{q} f(t^{n,i}, 0)x_i,$$

whence

$$\big(G(x), x\big) \geq c_1\|x\|^2 - c_2\|x\|$$

with a constant c_2 depending on the matrix A, the approximation y^n, the step-size h and the values $f(t^{n,1}, 0), \ldots, f(t^{n,q}, 0)$. Choose now x appropriately and utilize Brouwer's fixed-point theorem, in the version given in Exercise 2.13, to obtain the desired result.]

3.31 We use the notation of Exercise 3.30. We assume that the function f satisfies the condition mentioned there, while for the matrix A we assume this time that there exists an invertible diagonal matrix D, such that the matrix $C := DAD^{-1}$ is positive definite. (Notice that in the previous exercise D was the identity matrix.) Prove that the result of Exercise 3.30 is valid also under this more general condition
[*Hint*: Write now relation ($\star$) in the Hint of Exercise 3.30 in the form

$$C^{-1}D(Y - y^n\boldsymbol{e}) = hD \begin{pmatrix} f(t^{n,1}, y^{n,1}) \\ \vdots \\ f(t^{n,q}, y^{n,q}) \end{pmatrix},$$

define a suitable mapping, take the inner product with Dx and use the fact that the matrix C^{-1} is positive definite.]

3.32 Assume that a Runge–Kutta satisfies the hypotheses of Exercise 3.30 and that the continuous function $f : [a,b] \times \mathbb{R} \to \mathbb{R}$ is such that

$$\big(f(t,y) - f(t,z)\big)(y-z) \le \nu(y-z)^2, \quad a \le t \le b,\ y,z \in \mathbb{R},$$

with a positive constant ν. Show that the approximations are well defined, if the product νh is sufficiently small.

3.33 Generalize Exercise 3.32 to the case that a Runge–Kutta method satisfies the assumptions of Exercise 3.31.

3.34 Consider a Runge–Kutta method and the associated rational function r of the exponential function, $r(z) = 1 + zb^T(I - zA)^{-1}\boldsymbol{e}$; see (3.57). Show that the method is consistent, if and only if $r(0) = r'(0) = 1$; see Exercise 3.9.

3.35 Assume that the order of a Runge–Kutta method is p. Show that

$$b^T A^{\ell-1}\boldsymbol{e} = \frac{1}{\ell!}, \quad \ell = 1, \dots, p.$$

[*Hint*: With the notation of the previous Exercise, we have

$$r(z) = 1 + \sum_{j=0}^{\infty} \big(b^T A^j \boldsymbol{e}\big) z^{j+1}.$$

Use now (3.62).]

3.36 We consider a RK method and use the notation of (3.37) and (3.38).

a) Let $\nu := \min(p,s)$. Prove that, if the solution y of the initial value problem (3.1) is a polynomial of degree at most ν, then the RK method yields as approximations the exact values of the solution at the nodes, i.e., the method integrates the problem *exactly*.

[*Hint*: Obviously $y' \in \mathbb{P}_{\nu-1}$. Assume $y^n = y(t^n)$. Then

$$y(t^{n,i}) = y(t^n) + \int_{t^n}^{t^{n,i}} y'(t)\,dt = y(t^n) + h\sum_{j=1}^{q} a_{ij}\, y'(t^{n,i})$$

$$= y(t^n) + h\sum_{j=1}^{q} a_{ij}\, f\big(t^{n,j}, y(t^{n,j})\big),$$

whence, for sufficiently small h such that the quantities $y^{n,i}$ are uniquely defined, we have $y(t^{n,i}) = y^{n,i}, i = 1, \dots, q$. Furthermore,

$$\begin{aligned} y(t^{n+1}) &= y(t^n) + \int_{t^n}^{t^{n+1}} y'(t)\,dt = y(t^n) + h\sum_{i=1}^{q} b_i\, y'(t^{n,i}) \\ &= y(t^n) + h\sum_{i=1}^{q} b_i\, f\big(t^{n,i}, y(t^{n,i})\big) \\ &= y^n + h\sum_{i=1}^{q} b_i\, f\big(t^{n,i}, y^{n,i}\big) = y^{n+1}.\,] \end{aligned}$$

b) Consider the explicit midpoint method, Example 5 is section 3.1. Check that in this case we have $s = 1$ και $p = 2$. Apply the method to the example

$$\begin{cases} y'(t) = 2y(t)/t, \quad 1 \le t \le 2, \\ y(1) = 1, \end{cases}$$

with solution $y(t) = t^2$, with step-size $h = 1$, to see that $y^1 = 11/3$ while $y(2) = 4$, i.e., the problem is not integrated exactly. Thus, in general, in part a) we can not replace ν be the order of accuracy p of the method.

3.37 (B–stability of RK Radau IIA methods.) Prove that the RK Radau IIA methods are B–stable.

[*Hint*: Consider the initial value problem (1.19) and, with the usual notation, for given approximations $Y(t^n)$ and $Z(t^n)$, consider the approximations Y and Z given be the collocation method corresponding to the nodes $0 < \tau_1 < \cdots < \tau_q = 1$; see Remark 3.4. As in Remark 3.3, we set

$$m(t) := \|Y(t) - Z(t)\|^2, \quad t \in [t^n, t^{n+1}].$$

Assume now that the one-sided Lipschitz condition (1.24) is fulfilled and prove that $m'(t^{n,i}) \le 0$, whence

$$h\sum_{i=1}^{q} b_i m'(t^{n,i}) \le 0,$$

with b_i as in Remark 3.4. Now, as is well known, for the error of the Radau quadrature rule the following representation holds

$$\forall\, f \in C^{2q-1}[0,1]\ \ \exists \xi \in (0,1) \quad \int_0^1 f(x)\,dx - Q_q(f) = -\frac{q\big[(q-1)!\big]^4}{2\big[(2q-1)!\big]^3} f^{(2q-1)}(\xi).$$

Check that $m^{(2q)} \ge 0$ and combine the above results to infer that $m(t^{n+1}) \le m(t^n)$.]

3.38 Let p be the order of an explicit RK method. If the polynomial r associated with the method is of degree higher that p, show that r is *not* an entry of the Padé table of the exponential function.

3.39 Give an example of an implicit RK method with corresponding rational approximation to the exponential r that is not an element of the Padé table of the exponential. [*Hint*: Consider the theta-method with $\vartheta \neq 0, 1/2, 1$. Alternatively, consider the method (3.11) with appropriate μ. (The corresponding rational approximation to the exponential r was given a little before (3.63).)]

3.40 Assume that the matrix A in the tableau (3.2) is invertible. We introduce the vector $c \in \mathbb{R}^q$, $c := (A^T)^{-1}b$. Show that (3.5) can be written in the form

$$y^{n+1} = y^n + \sum_{i=1}^{q} c_i (y^{n,i} - y^n).$$

[*Hint*: We write (3.4) as

$$\begin{pmatrix} y^{n,1} - y^n \\ \vdots \\ y^{n,q} - y^n \end{pmatrix} = hA \begin{pmatrix} f(t^{n,1}, y^{n,1}) \\ \vdots \\ f(t^{n,q}, y^{n,q}) \end{pmatrix}$$

and have

$$b^T A^{-1} \begin{pmatrix} y^{n,1} - y^n \\ \vdots \\ y^{n,q} - y^n \end{pmatrix} = hb^T \begin{pmatrix} f(t^{n,1}, y^{n,1}) \\ \vdots \\ f(t^{n,q}, y^{n,q}) \end{pmatrix} = h \sum_{i=1}^{q} b_i f(t^{n,i}, y^{n,i}),$$

whence, with $b^T A^{-1} = c^T$,

$$\sum_{i=1}^{q} c_i (y^{n,i} - y^n) = h \sum_{i=1}^{q} b_i f(t^{n,i}, y^{n,i}).$$

Now, relation $b^T A^{-1} = c^T$ can be equivalently written as $\left((A^{-1})^T b\right)^T = c^T$, i.e., $c = (A^{-1})^T b$. Finally, as is well known, $(A^{-1})^T = (A^T)^{-1}$.]

4. Multistep methods

Two are the most important and most widely used classes of numerical methods for initial value problems. The Runge–Kutta methods, which we studied in the previous chapter, and the multistep methods, which we will investigate in this chapter. The multistep methods have a long history, basically because their implementation is not expensive. This is why they were implemented even before computers were invented, starting from the work of Bashforth and Adams from 1883, but their theory was developed in the 1950's, mainly due to G. Dahlquist's contributions, who proved that a multistep method converges, if and only if it is consistent and stable[1]. (We will define these concepts and will reproduce Dahlquist's proof in sections 4.3 and 4.4.) The drawback of multistep methods is that they do not have as good stability properties as the Runge–Kutta methods, and also that their implementation and theory for variable step-sizes is more involved. They are superior to Runge–Kutta methods as far as the computational cost is concerned.

In section 4.1 we will present some preliminaries and examples of multistep methods. Section 4.2 is devoted to some basic properties of linear difference equations, which will be helpful in section 4.3, where we will study the stability of multistep methods. (We notice already here that in contrast to Runge–Kutta methods (which are stable, see Proposition 3.2), not all multistep methods are stable.) In section 4.4 we will investigate the consistency and convergence of multistep methods, and in section 4.5 their absolute stability properties.

[1] G. Dahlquist: *Convergence and stability in the numerical integration of ordinary differential equations*. Math. Scand. **4** (1956) 33–53.

4.1 Preliminaries: Notation and examples

In this section we will introduce a second class of numerical methods for initial value problems, the so-called multistep methods. More precisely, we will study *linear* multistep methods.

We will again consider the initial value problem: Seek a function $y : [a,b] \to \mathbb{R}^m$ such that

$$\begin{cases} y' = f(t,y), \quad a \le t \le b, \\ y(a) = y_0 \end{cases} \tag{4.1}$$

with given $y_0 \in \mathbb{R}^m$ and $f : [a,b] \times \mathbb{R}^m \to \mathbb{R}^m$. Let $N \in \mathbb{N}, h := \frac{b-a}{N}$ and $t^n := a + nh, n = 0, \ldots, N$. For brevity, in the sequel we write f^k instead of $f(t^k, y^k)$. An example of a multistep (more precisely, a two-step) method is the scheme

$$\begin{cases} y^0, y^1 \quad \text{given}, \\ y^{n+2} - y^n = 2hf^{n+1}, \quad n = 0, \ldots, N-2. \end{cases} \tag{4.2}$$

On way that leads to it is by approximating $y'(t^{n+1})$ by the difference quotient $\frac{y(t^{n+2})-y(t^n)}{2h}$. If we approximate this derivative by other difference quotients, we obtain other multistep methods. We will come back to this in the sequel.

Another useful technique to construct multistep methods is by means of numerical integration. For instance, integrating the o.d.e. $y'(t) = f\big(t, y(t)\big)$ in the interval $[t^n, t^{n+2}]$, we obtain

$$y(t^{n+2}) - y(t^n) = \int_{t^n}^{t^{n+2}} f\big(t, y(t)\big)\, dt.$$

Approximating the integral on the right-hand side by the midpoint rule, we obtain again method (4.2). Approximating the same integral by the Simpson rule, we obtain the two-step method

$$\begin{cases} y^0, y^1 \quad \text{given}, \\ y^{n+2} - y^n = \dfrac{h}{3}\big(f^{n+2} + 4f^{n+1} + f^n\big), \quad n = 0, \ldots, N-2, \end{cases} \tag{4.3}$$

which, for obvious reasons, is referred to as *Simpson's method*. Simpson's method is *implicit*, since the calculation of y^{n+2} requires at each time level solving an $m \times m$ nonlinear system. In contrast, method (4.2) is *explicit*.

The general (linear) $k-$*step* method for the numerical solution of problem (4.1) is described by $2k+2$ constants $\alpha_0, \ldots, \alpha_k$, $\beta_0, \ldots, \beta_k$, and is of the form

$$(4.4)\quad \begin{cases} y^0, y^1, \ldots, y^{k-1} \quad \text{given,} \\ \alpha_k y^{n+k} + \alpha_{k-1} y^{n+k-1} + \cdots + \alpha_0 y^n = h\big(\beta_k f^{n+k} + \cdots + \beta_0 f^n\big), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad n = 0, \ldots, N-k. \end{cases}$$

We will usually assume that $\alpha_k = 1$ and that $|\alpha_0| + |\beta_0| > 0$, such that we will indeed have a $k-$step method. If $\beta_k = 0$, then the method is called *explicit*: In this case y^{n+k} can be determined by a simple substitution of the already known values $y^{n+i}, i = 0, \ldots, k-1$. If $\beta_k \neq 0$, then the method is called *implicit*: To compute y^{n+k} we need to solve an $m \times m$ nonlinear system of the form

$$y^{n+k} = h\beta_k f(t^{n+k}, y^{n+k}) + g^n,$$

with known g^n. If L is the Lipschitz constant of the function f, with respect to its second variable y, and $h|\beta_k|L < 1$, then, according to the contraction theorem, y^{n+k} is uniquely defined. As far as the computational cost is concerned, multistep methods are by far less expensive than Runge–Kutta methods. In the case of an explicit multistep method only one evaluation of the function f is needed (all other evaluations have been already done at previous time levels), and in the case of implicit methods we need, in addition, to solve an $m \times m$ nonlinear system (in contrast to Runge–Kutta methods, where the system is $qm \times qm$). The advantage of implicit Runge–Kutta methods is that they combine high order of accuracy with excellent stability properties. The starting values $y^0, \ldots, y^{k-1}$ that are needed in the case of a $k-$step method are usually computed by Runge–Kutta methods, with given initial value y^0.

There are various ways to systematically construct multistep methods via, e.g., numerical integration, numerical differentiation, Taylor expansions etc. Utilizing polynomial interpolation and numerical differentiation, we are led, for instance, to a particularly useful class of multistep methods, the so-called

backward difference formulas (BDF): We assume for the time being that $m = 1$. Let $P_{n,k}$ be the polynomial of degree at most k such that

$$P_{n,k}(t^{n+i}) = y(t^{n+i}), \quad i = 0, \ldots, k,$$

that is the Lagrange interpolating polynomial at the values $y(t^{n+i}), 0 \le i \le k$. Approximating $y'(t^{n+k})$ in relation $y'(t^{n+k}) = f\big(t^{n+k}, y(t^{n+k})\big)$ by the derivative of the interpolating polynomial at the same point, $P'_{n,k}(t^{n+k})$, and expressing the last quantity in terms of the values $y(t^{n+i}), 0 \le i \le k$, we are led to the method

$$\text{(4.5)} \qquad \begin{cases} y^0, y^1, \ldots, y^{k-1} \quad \text{given,} \\ \displaystyle\sum_{j=1}^{k} \frac{1}{j} \nabla^j y^{n+k} = h f^{n+k}, \quad n = 0, \ldots, N-k, \end{cases}$$

where we used the usual notation $\nabla^1 y^n := y^n - y^{n-1}$, $\nabla^j y^n := \nabla^1(\nabla^{j-1} y^n)$. Obviously, (4.5) is a $k-$step method that can be used also for systems of o.d.e's. Method (4.5) is called *$k-$step backward difference formula*. For $k = 1, 2, 3, \ldots$, (4.5) can be written in the form (4.4) as:

$$\begin{aligned} &k = 1\colon \ \alpha_1 = 1, \ \alpha_0 = -1, \quad \beta_1 = 1 \quad \text{(implicit Euler)} \\ &k = 2\colon \ \alpha_2 = 1, \ \alpha_1 = -\frac{4}{3}, \quad \alpha_0 = \frac{1}{3}, \quad \beta_2 = \frac{2}{3} \\ &k = 3\colon \ \alpha_3 = 1, \ \alpha_2 = -\frac{18}{11}, \ \alpha_1 = \frac{9}{11}, \ \alpha_0 = -\frac{2}{11}, \ \beta_3 = \frac{6}{11} \end{aligned}$$

etc. Notice that multiplying by an appropriate factor, such that $\beta_k = 1$, see (4.5), the quantities α_j, $j = 0, \ldots, k$, are the coefficients of ζ^j, $j = 0, \ldots, k$, of the polynomial α,

$$\alpha(\zeta) := \sum_{i=1}^{k} \frac{1}{i} \zeta^{k-i} (\zeta - 1)^i.$$

For a systematic study as well as for many examples of multistep methods we refer to the classic monograph of Henrici [18]. Usually $k-$step methods

of the form

$$\text{(4.6)}\qquad \begin{cases} y^0, y^1, \dots, y^{k-1} \quad \text{given,} \\ y^{n+k} - y^{n+k-1} = h \sum_{j=0}^{k} \beta_j f^{n+j}, \quad n = 0, \dots, N-k, \end{cases}$$

are called *Adams methods*. Explicit methods of this class ($\beta_k = 0$) are called *Adams–Bashforth methods*, while implicit ($\beta_k \neq 0$) *Adams–Moulton methods*. Examples of Adams–Bashforth methods are

$$\text{(4.7)}\qquad k = 2: \quad y^{n+2} - y^{n+1} = h\big(\frac{3}{2} f^{n+1} - \frac{1}{2} f^n\big),$$

$$\text{(4.8)}\qquad k = 3: \quad y^{n+3} - y^{n+2} = h\big(\frac{23}{12} f^{n+2} - \frac{4}{3} f^{n+1} + \frac{5}{12} f^n\big),$$

while the methods

$$\text{(4.9)}\quad k = 2: \quad y^{n+2} - y^{n+1} = h\big(\frac{5}{12} f^{n+2} + \frac{2}{3} f^{n+1} - \frac{1}{12} f^n\big),$$

$$\text{(4.10)}\quad k = 3: \quad y^{n+3} - y^{n+2} = h\big(\frac{9}{24} f^{n+3} + \frac{19}{24} f^{n+2} - \frac{5}{24} f^{n+1} + \frac{1}{24} f^n\big)$$

are of Adams–Moulton–type.

Methods of the form

$$\text{(4.11)}\qquad \begin{cases} y^0, y^1, \dots, y^{k-1} \quad \text{given,} \\ y^{n+k} - y^{n+k-2} = h \sum_{j=0}^{k} \beta_j f^{n+j}, \quad n = 0, \dots, N-k, \end{cases}$$

are called *Nyström methods*, if $\beta_k = 0$, and *Milne–Simpson methods*, if $\beta_k \neq 0$. Thus, method (4.2) is of Nyström-type, while (4.3) is of Milne–Simpson-type, with $k = 2$.

We now proceed to the analysis of multistep methods. To simplify the notation, we restrict ourselves to the case $m = 1$, i.e., to scalar o.d.e's. Analogue results hold also for systems of o.d.e's, $m > 1$.

We start with a brief review of linear difference equations; elements of this theory will be useful in the sequel.

4.2 Linear difference equation

The homogeneous difference equations with constant coefficients are of the form

$$\alpha_k y^{n+k} + \alpha_{k-1} y^{n+k-1} + \cdots + \alpha_0 y^n = 0, \quad n \geq 0, \tag{4.12}$$

with $\alpha_0, \ldots, \alpha_k \in \mathbb{R}$ given coefficients. We say that (4.12) has constant coefficients, since the coefficients $\alpha_k, \ldots, \alpha_0$ are independent of n. Every sequence $(y^n)_{n\in\mathbb{N}_0} \subset \mathbb{C}$, satisfying (4.12), is called a solution of (4.12). In the sequel, without loss of generality, we assume that the coefficients α_k and α_0 do not vanish. The solution space of (4.12) is a linear space, i.e., if $\left(y_1^n\right)_{n\in\mathbb{N}_0}, \left(y_2^n\right)_{n\in\mathbb{N}_0} \subset \mathbb{C}$ are two solutions and $\alpha, \beta \in \mathbb{C}$, then $\left(\alpha y_1^n + \beta y_2^n\right)_{n\in\mathbb{N}_0}$ is also a solution of (4.12). The dimension of the solution space of (4.12) can be easily determined. We first observe that, for any starting data $y^0, \ldots, y^{k-1}$, there exists exactly one sequence $(y^n)_{n\in\mathbb{N}_0}$ (with the first k components the starting values), which solves (4.12). Let $\left(y_j^n\right)_{n\in\mathbb{N}_0}$, $j = 1, \ldots, m$, be solutions of (4.12). We say that these solutions are linearly dependent, if there exist constants $\gamma_1, \ldots, \gamma_m \in \mathbb{C}$, not all vanishing, such that

$$\gamma_1 y_1^n + \cdots + \gamma_m y_m^n = 0, \quad n \in \mathbb{N}_0.$$

If the solutions are not linearly dependent, then we say they are linearly independent. Consider now the following solutions $\left(y_j^n\right)_{n\in\mathbb{N}_0} \subset \mathbb{R}$, $j = 0, \ldots, k-1$, of (4.12): The starting values $\left(y_j^n\right)_{n\in\mathbb{N}_0}$ are such that $y_j^m = \delta_{jm}$, $j, m = 0, \ldots, k-1$. Obviously, these solutions of (4.12) are linearly independent. Furthermore, every solution $(y^n)_{n\in\mathbb{N}_0}$ of (4.12) can be written in the form

$$y^n = y^0 y_0^n + \cdots + y^{k-1} y_{k-1}^n, \quad n \in \mathbb{N}_0. \tag{4.13}$$

Indeed, both sides of this relation yield solutions of (4.12), and their first k starting components coincide. Since our problem is uniquely solvable, if we specify the first k components, these two solutions coincide. The dimension of the solution space of (4.12) is, therefore, k. Every basis of the solution space of (4.12) is called *fundamental system* of (4.12). Let $\left(y_j^n\right)_{n\in\mathbb{N}_0}$, $j = 1, \ldots, k$,

be a fundamental system of (4.12), and $(\gamma^n)_{n\in\mathbb{N}_0}$ an arbitrary solution. Let $c_1,\dots,c_k$ be the unique solution of the system

$$\sum_{j=1}^{k} y_j^n c_j = \gamma^n, \quad n = 0,\dots,k-1.$$

Then, obviously,

$$\gamma^n = \sum_{j=1}^{k} c_j y_j^n, \quad n \in \mathbb{N}_0.$$

We presented above a way to determine a fundamental system of (4.12). Theoretically more elegant, and also more useful for our subsequent analysis, is the following approach to determine a fundamental system of (4.12): With the coefficients $\alpha_k,\dots,\alpha_0$ of (4.12), consider the polynomial ρ,

$$\rho(z) := \alpha_k z^k + \cdots + \alpha_0.$$

The roots of ρ do not vanish, since the constant coefficient α_0 does not vanish. We now want to determine nontrivial solutions of (4.12) (nontrivial meaning that at least one of its components does not vanish) of the form $(y^n)_{n\in\mathbb{N}_0}$, with $y^n = z^n$ (n is an index in y^n and an exponent in z^n). Then $z \neq 0$ and $\alpha_k z^{n+k} + \cdots + \alpha_0 z^n = 0$, i.e., $\rho(z) = 0$. Consequently, z is a solution of ρ. We now distinguish two cases.

First case. The roots $z_1,\dots,z_k$ of ρ are pairwise distinct. We can then easily see (Vandermonde determinant!), that the solutions $\left(y_j^n\right)_{n\in\mathbb{N}_0}$, $y_j^n := z_j^n$, $j = 1,\dots,k$, of (4.12) constitute a fundamental system.

Second case. Polynomial ρ has multiple roots. Let z be a solution of ρ of multiplicity ν. Then, the sequences $\left(y_j^n\right)_{n\in\mathbb{N}_0}$, $j = 1,\dots,\nu$,

$$\text{(4.14)} \qquad \begin{cases} y_1^n := z^n \\ y_2^n := nz^n \\ \quad\vdots \\ y_\nu^n := n(n-1)\cdots(n-\nu+2)z^n \end{cases}, \quad n \in \mathbb{N}_0,$$

are solutions of (4.12). Indeed, for every $n \in \mathbb{N}_0$, z is a root of multiplicity ν of the polynomial r_n, $r_n(x) := x^n\rho(x)$. Consequently, $r_n(z) = r_n'(z) = \cdots =$

$r_n^{(\nu-1)}(z) = 0$. This means that

$$\begin{aligned}
&\alpha_k z^{n+k} + \alpha_{k-1} z^{n+k-1} + \cdots + \alpha_0 z^n = 0,\\
&\alpha_k (n+k) z^{n+k-1} + \alpha_{k-1}(n+k-1) z^{n+k-2} + \cdots + \alpha_0 n z^{n-1} = 0,\\
&\qquad \vdots\\
&\alpha_k (n+k)(n+k-1)\cdots(n+k-\nu+2) z^{n+k-\nu+1} + \cdots\\
&\qquad\qquad \cdots + \alpha_0 n(n-1)\cdots(n-\nu+2) z^{n-\nu+1} = 0,
\end{aligned}$$

whence the sequences given in (4.14) are in fact solutions of (4.12). Let now $z_1, \ldots, z_m, m < k$, be the pairwise distinct roots of ρ and $p_1, \ldots, p_m$ their multiplicities. Then, according to our previous remarks, the sequences $\left(y_j^n\right)_{n\in\mathbb{N}_0}$, $j = 1, \ldots, k$,

$$\left\{\begin{aligned}
y_1^n &:= z_1^n\\
y_2^n &:= n z_1^n\\
&\vdots\\
y_{p_1}^n &:= n(n-1)\cdots(n-p_1+2) z_1^n\\
y_{p_1+1}^n &:= z_2^n\\
&\vdots\\
y_{p_1+p_2}^n &:= n(n-1)\cdots(n-p_2+2) z_2^n\\
&\vdots\\
y_k^n &:= n(n-1)\cdots(n-p_m+2) z_m^n,
\end{aligned}\right.$$

are solutions of (4.12). Proving that an appropriate matrix is invertible, one can infer that these solutions constitute a fundamental system of (4.12). We will not give the details here.

As we just saw, we can in both cases give a fundamental system of (4.12) in closed form.

4.3 Stability of multistep methods

Consider the initial value problem

$$\begin{cases} y' = f(t,y), \quad a \le t \le b, \\ y(a) = y_0, \end{cases} \tag{4.15}$$

with continuous $f : [a,b] \times \mathbb{R} \to \mathbb{R}$, satisfying, as usually, the Lipschitz condition with respect to its second variable, uniformly with respect to its first variable. Let $N \in \mathbb{N}$, $h := \frac{b-a}{N}$, and $t^n := a + nh, n = 0, \dots, N$.

Definition 4.1 (Stability of multistep methods.) We say that a $k-$step method, described by the constants $\alpha_k, \dots, \alpha_0, \beta_k, \dots, \beta_0$, is *stable*, if there exists a constant C, depending on f but independent of N, such that for sequences $(y^n), (z^n)$ satisfying

$$\begin{cases} y^0, \dots, y^{k-1} \quad \text{given}, \\ \alpha_k y^{n+k} + \alpha_{k-1} y^{n+k-1} + \dots + \alpha_0 y^n = \\ \quad h\big[\beta_k f(t^{n+k}, y^{n+k}) + \dots + \beta_0 f(t^n, y^n)\big], \ \ n = 0, \dots, N-k, \end{cases} \tag{4.16}$$

and

$$\begin{cases} z^0, \dots, z^{k-1} \quad \text{given}, \\ \alpha_k z^{n+k} + \alpha_{k-1} z^{n+k-1} + \dots + \alpha_0 z^n = \\ \quad h\big[\beta_k f(t^{n+k}, z^{n+k}) + \dots + \beta_0 f(t^n, z^n)\big], \quad n = 0, \dots, N-k, \end{cases} \tag{4.17}$$

there holds

$$\max_{0 \le n \le N} |y^n - z^n| \le C \max_{0 \le j \le k-1} |y^j - z^j|. \tag{4.18}$$

Definition 4.2 (The root condition.) We say that a multistep method (4.4) satisfies the *root condition*, if for its *characteristic polynomial* ρ, defined by

$$\rho(z) := \alpha_k z^k + \dots + \alpha_0,$$

there holds

$$\begin{aligned} \rho(z) = 0 &\implies |z| \le 1, \\ \rho(z) = \rho'(z) = 0 &\implies |z| < 1, \end{aligned}$$

that is, all roots of ρ have absolute value at most one (are in the unit disc) while multiple roots are in the interior of the unit disc (i.e., the roots on the unit circle are simple.) □

Our aim in this section is to show that a multistep method is stable, if and only if it satisfies the root condition. Let us note here that to check the root condition we do not really need to know the roots of a polynomial, as this can be done using some criteria, like the Schur criterion or the Routh–Hurwitz criterion; see Remark 4.1 at the end of this section.

The necessity of the root condition can be easily shown, by utilizing the theory of linear difference equations. The converse is essentially more involved and requires some preparationn.

Let us begin with the easy part. We set $f = 0$ in (4.15) and $z^i = 0, i \in \mathbb{N}_0$, in (4.17), in which case (4.18) takes the form

$$\max_{0\le n\le N} |y^n| \le C \max_{0\le j\le k-1} |y^j|;$$

this relation must hold for the solutions of the difference equation

$$\alpha_k y^{n+k} + \alpha_{k-1} y^{n+k-1} + \cdots + \alpha_0 y^n = 0, \quad n \in \mathbb{N}_0,$$

see (4.16) for $f = 0$, with a constant C independent of N. Let z be a root of the characteristic polynomial ρ. Then, the sequence $y^n := z^n$ (n is an index in y^n and an exponent in z^n), $n \in \mathbb{N}_0$, is a solution of the above equation, and, consequently, there must hold

$$\max_{0\le n\le N} |z|^n \le C \max_{0\le j\le k-1} |z^j|.$$

This relation can not be valid for $|z| > 1$, since then $|z|^N \to \infty$, $N \to \infty$. So, if the method is stable, then we necessarily have $|z| \le 1$. Assume now that z is a multiple root of ρ. Then, the sequence $y^n := nz^n, n \in \mathbb{N}_0$, is a solution of the difference equation, whence the estimate

$$\max_{0\le n\le N} \left(n|z|^n\right) \le C \max_{0\le j\le k-1} |jz^j|$$

must hold true. If $|z| = 1$, this relation can obviously not hold for any constant C, and we infer that $|z| < 1$. Summarizing, we proved that if the method is stable, then its characteristic polynomial satisfies the root condition.

To prove the converse direction, i.e., that the root condition ensures stability, there are two techniques. Dahlquist proved the result first, and later Butcher gave an alternative proof. Both proofs are of interest. Butcher's proof is simpler, but Dahlquist's proof gives more information, at least in some difficult cases. Hence, we will present both proofs here. We will first follow Butcher's technique, and for that we will need a preliminary result. We will present the other technique later in this section.

Lemma 4.1 (Preliminary result for Butcher's technique.) *Let* ρ, $\rho(z) := \alpha_0 + \alpha_1 z + \cdots + \alpha_{k-1} z^{k-1} + \alpha_k z^k$, *be a polynomial, with* $\alpha_k = 1$, *satisfying the root condition. Consider the* $k \times k$ *matrix*

$$A := \begin{pmatrix} -\alpha_{k-1} & -\alpha_{k-2} & \dots & -\alpha_0 \\ 1 & 0 & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & 0 \end{pmatrix}.$$

Then, there exists a norm $\|\cdot\|$ *in* $\mathbb{C}^k$, *such that for the induced norm for* $k \times k$ *matrices there holds*

$$\|A\| \leq 1. \tag{4.19}$$

Proof. Denote by $\lambda_1, \dots, \lambda_m$ the simple and by $\lambda_{m+1}, \dots, \lambda_k$ the multiple roots of ρ. Then, according to the root condition, we have

$$|\lambda_i| \leq 1,\ 1 \leq i \leq m, \quad |\lambda_i| < 1,\ m+1 \leq i \leq k. \tag{4.20}$$

Now, it is not difficult to check that the eigenvalues of the matrix A are exactly the roots of ρ. (Indeed, it can be easily seen, e.g., by expanding the determinant with respect to its last column, that $\det(A - \lambda I) = (-1)^k \rho(\lambda)$, or, alternatively, to check that any root λ of ρ is an eigenvalue of A with corresponding eigenvector $(\lambda^{k-1}, \lambda^{k-2}, \dots, 1)^T$.) Consequently, there exists an invertible matrix $T \in \mathbb{C}^{k,k}$ such that $T^{-1}AT = J$, where J is the *Jordan*

canonical form of A,

$$J = \begin{pmatrix} \lambda_1 & & 0 & & & & 0 \\ & \ddots & & & & & \\ 0 & & \lambda_m & & & & \\ & & & \lambda_{m+1} & \sigma_{m+1} & & 0 \\ & & & & \ddots & \ddots & \\ & & & & & & \sigma_{k-1} \\ 0 & & & 0 & & & \lambda_k \end{pmatrix},$$

and where the entries $\sigma_i, m+1 \leq i \leq k-1$, are either zero or one. Consider now the $k \times k$ diagonal matrix $D = \text{diag}\,(1, \varepsilon, \varepsilon^2, \ldots, \varepsilon^{k-1})$, with ε a positive number. Then, it is not difficult to see that the matrix $\widetilde{J} = D^{-1}JD$ is of the form

$$\widetilde{J} = \begin{pmatrix} \lambda_1 & & 0 & & & & 0 \\ & \ddots & & & & & \\ 0 & & \lambda_m & & & & \\ & & & \lambda_{m+1} & \tilde{\sigma}_{m+1} & & 0 \\ & & & & \ddots & \ddots & \\ & & & & & & \tilde{\sigma}_{k-1} \\ 0 & & & 0 & & & \lambda_k \end{pmatrix},$$

with $\tilde{\sigma}_i, m+1 \leq i \leq k-1$, equal to zero or to ε. Consequently, since $\|\widetilde{J}\|_\infty \leq \max\{\max_{1\leq i\leq m} |\lambda_i|, \max_{m+1\leq i\leq k} |\lambda_i| + \varepsilon\}$, in view of the root condition (4.20) we will have

$$\|\widetilde{J}\|_\infty \leq 1, \tag{4.21}$$

if we choose $\varepsilon > 0$ such that $\max_{m+1\leq i\leq k} |\lambda_i| + \varepsilon \leq 1$.

Now to prove the desired inequality (4.19), we notice that the similarity transformation $\widetilde{J} = Q^{-1}AQ$, with $Q = TD$, holds true, since $\tilde{J} = D^{-1}JD = D^{-1}T^{-1}ATD$. Introducing the norm $\|\cdot\|$ in $\mathbb{C}^k$ by $\|x\| := \|Q^{-1}x\|_\infty$

for $x \in \mathbb{C}^k$, we thus have, in view of (4.21),

$$\|A\| = \sup_{\substack{x\in\mathbb{C}^k \\ x\neq 0}} \frac{\|Ax\|}{\|x\|} = \sup_{\substack{x\in\mathbb{C}^k \\ x\neq 0}} \frac{\|Q^{-1}Ax\|_\infty}{\|Q^{-1}x\|_\infty} = \sup_{\substack{x\in\mathbb{C}^k \\ x\neq 0}} \frac{\|\widetilde{J}Q^{-1}x\|_\infty}{\|Q^{-1}x\|_\infty}$$
$$= \sup_{\substack{y\in\mathbb{C}^k \\ y\neq 0}} \frac{\|\widetilde{J}y\|_\infty}{\|y\|_\infty} = \|\widetilde{J}\|_\infty \leq 1. \qquad \square$$

We are now in a position to prove that, if a multistep method satisfies the root condition, then it is stable. We will formulate the result a little bit more generally, in a form we will use later in the error estimate, to avoid repetitions.

Proposition 4.1 (Stability of multistep methods. Butcher) *Assume that the $k-$step method* (4.4) *satisfies the root condition. Let $\lambda^n, n = 0, \ldots, N-k$, be given constants, and let $\beta_i^n, i = 0, \ldots, k$, $n = 0, \ldots, N-k$, be such that $|\beta_i^n| \leq B < \infty$. For $h = \frac{b-a}{N}$ we consider the difference equation*

$$\alpha_k\psi^{n+k} + \alpha_{k-1}\psi^{n+k-1} + \cdots + \alpha_0\psi^n = h\big(\beta_k^n\psi^{n+k} + \cdots + \beta_0^n\psi^n\big) + \lambda^n, \quad 0 \leq n \leq N-k. \tag{4.22}$$

There, there exists $h_0 > 0$ such that for $h \leq h_0$ there holds

$$\max_{0\leq n\leq N} |\psi^n| \leq C\big[N \max_{0\leq n\leq N-k} |\lambda^n| + \max_{0\leq j\leq k-1} |\psi^j|\big]; \tag{4.23}$$

here the constant C depends on $b-a$, h_0, B, but is independent of h, λ^n, ψ^n, N and β_i^n.

Proof. Without loss of generality, we assume that $\alpha_k = 1$ and write (4.22) in the form

$$\psi^{n+k} + \alpha_{k-1}\psi^{n+k-1} + \cdots + \alpha_0\psi^n = h\big(\beta_k^n\psi^{n+k} + \cdots + \beta_0^n\psi^n\big) + \lambda^n.$$

These relations can be written as single-step *vectorial* recursive relation,

$$Y^{n+1} = AY^n + G^n, \quad 0 \leq n \leq N-k,$$

with A the $k \times k$ matrix of Lemma 4.1, and with the $k-$vectors Y^j, G^j defined as

$$Y^j = \begin{pmatrix} \psi^{j+k-1} \\ \psi^{j+k-2} \\ \vdots \\ \psi^j \end{pmatrix}, \quad G^j = \begin{pmatrix} h\big(\beta_k^j \psi^{j+k} + \cdots + \beta_0^j \psi^j\big) + \lambda^j \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Consequently, for the norm $\|\cdot\|$ of $\mathbb{C}^k$ ensured by Lemma 4.1, we have

$$\|Y^{n+1}\| \le \|A\|\,\|Y^n\| + \|G^n\| \le \|Y^n\| + \|G^n\|, \quad 0 \le n \le N-k. \tag{4.24}$$

In view of the equivalence of the norms in $\mathbb{C}^k$, there exist positive constants C_1 and C_2 such that

$$\begin{aligned} \|G^n\| &\le hC_1|\psi^{n+k}| + hC_1\big(|\psi^{n+k-1}| + \cdots + |\psi^n|\big) + C_1|\lambda^n| \\ &\le hC_2\|Y^{n+1}\| + hC_2\|Y^n\| + C_1|\lambda^n|. \end{aligned}$$

Therefore, (4.24) leads to the inequality

$$(1 - C_2h)\|Y^{n+1}\| \le (1 + C_2h)\|Y^n\| + C_1|\lambda^n|,$$

from which, for $h \le h_0 < 1/C_2$, we obtain

$$\|Y^{n+1}\| \le \frac{1 + C_2h}{1 - C_2h}\|Y^n\| + \frac{C_1}{1 - C_2h}|\lambda^n|, \quad 0 \le n \le N-k.$$

Notice now that there exists a positive constant C_3 such that the previous relation yields

$$\|Y^{n+1}\| \le (1 + C_3h)\|Y^n\| + C_3|\lambda^n|, \quad 0 \le n \le N-k. \tag{4.25}$$

(Indeed, we have

$$\frac{1 + C_2h}{1 - C_2h} = 1 + \frac{2C_2h}{1 - C_2h} \le 1 + \frac{2C_2h}{1 - C_2h_0}.$$

Thus, we can choose $C_3 = \max(\frac{2C_2}{1-C_2h_0}, \frac{C_1}{1-C_2h_0})$.)

Repeated application of (4.25) yields now (see Lemma 2.1)

$$\max_{0\le n\le N-k}\|Y^{n+1}\| \le C_4\big(\|Y^0\| + N\max_{0\le n\le N-k}|\lambda^n|\big), \tag{4.26}$$

with a constant C_4. Again due to the equivalence of norms in $\mathbb{C}^k$ we have $|\psi^{n+k}| \le \|Y^{n+1}\|_\infty \le C_5\|Y^{n+1}\|$ and $\|Y^0\| \le C_6\|Y^0\|_\infty = C_6\max_{0\le j\le k-1}|\psi^j|$, for appropriate constants C_5 and C_6. Thus, the desired estimate (4.23) follows from (4.26). □

The stability of a multistep method satisfying the root condition follows easily from Proposition 4.1. Indeed, assume that the $k-$step method (4.4) satisfies the root condition. Setting $\psi^j := y^j - z^j$, $j = 0,\dots,N$, with y^j, z^j satisfying (4.16) and (4.17), respectively, and subtracting (4.17) from (4.16) , we obtain

$$\begin{aligned}\alpha_k\psi^{n+k} + \alpha_{k-1}\psi^{n+k-1} + \cdots + \alpha_0\psi^n = \\ h\Big[\beta_k\big[f(t^{n+k},y^{n+k}) - f(t^{n+k},z^{n+k})\big] + \cdots \\ + \beta_0\big[f(t^n,y^n) - f(t^n,z^n)\big]\Big]\end{aligned} \tag{4.27}$$

for $n = 0,\dots,N-k$. Setting now, for $m = 0,\dots,N$,

$$g^m := \begin{cases}\dfrac{f(t^m,y^m) - f(t^m,z^m)}{y^m - z^m}, & \text{if } y^m \ne z^m,\\ 0, & \text{if } y^m = z^m,\end{cases}$$

we obviously have $|g^m| \le L$, with L the Lipschitz constant of f with respect to its second variable. With $\beta_i^n := \beta_i g^{n+i}$, $i = 0,\dots,k$, $n = 0,\dots,N-k$, (4.27) takes the form

$$\alpha_k\psi^{n+k} + \alpha_{k-1}\psi^{n+k-1} + \cdots + \alpha_0\psi^n = h\big(\beta_k^n\psi^{n+k} + \cdots + \beta_0^n\psi^n\big), \quad n = 0,\dots,N-k. \tag{4.28}$$

Now (4.23) leads immediately to (4.18).

To demonstrate the importance of the stability, we will now introduce the notion of convergence of a multistep method and will prove that stability is a necessary condition for convergence.

Definition 4.3 (Convergent multistep methods.) Let $y^0, \dots, y^{k-1}$ be such that

$$\lim_{h\to 0} y^j = y_0, \quad j = 0, \dots, k-1. \tag{4.29}$$

Let y^n be the approximation of $y(t^n)$, given by the $k-$step method (4.4). We say that the method is *convergent*, if, for all $t \in [a, b]$, there holds

$$\lim y^n = y(t),$$

as $h \to 0, n \to \infty$, in a way that $a + nh \to t$, and if this holds for any initial value problem of the form (4.15), with f satisfying the conditions mentioned there. □

Proposition 4.2 (Convergence implies stability.) *Every convergent multistep method is stable.*

Proof. (Henrici, [18].) We consider the initial value problem

$$\begin{cases} y'(t) = 0, \quad 0 \le t \le T, \\ y(0) = 0. \end{cases} \tag{4.30}$$

The choice of this problem is motivated by the fact that the coefficients $\beta_k, \dots, \beta_0$ are irrelevant for the stability of the method, and these coefficients do not appear in the method when $f = 0$.

The unique solution of (4.30) is obviously $y(t) = 0, 0 \le t \le T$. According to (4.4) we then have

$$\begin{cases} y^0, y^1, \dots, y^{k-1} \quad \text{given}, \\ \alpha_k y^{n+k} + \dots + \alpha_0 y^n = 0. \end{cases} \tag{4.31}$$

Since the method is convergent, there holds

$$\lim_{n\to\infty} y^n = 0, \quad h = \frac{T}{n}, \tag{4.32}$$

provided that

$$\lim_{h\to 0} y^j = 0, \quad j = 0, \dots, k-1. \tag{4.33}$$

Let z be a root of the characteristic polynomial ρ of the method. We write z in the form $z = re^{i\varphi}, r \geq 0, 0 \leq \varphi < 2\pi$. Since ρ has real coefficients, the sequence $\left(y^\ell\right)_{\ell\in\mathbb{N}_0}$, $y^\ell := h\,\mathrm{Re}(z^\ell) = hr^\ell\cos(\ell\varphi), \ell \in \mathbb{N}_0$, satisfies (4.31). Furthermore, due to the factor h in its definition, the sequence obviously satisfies (4.33) as well. If $\varphi = 0$ or $\varphi = \pi$ (that is $z \in \mathbb{R}$), then a necessary condition for (4.32) is $|y^n| = \frac{T}{n}r^n \to 0, n \to \infty$, i.e., $r \leq 1$, whence $|z| \leq 1$. Let now $\varphi \neq 0, \varphi \neq \pi$. Then

$$\begin{aligned}(y^n)^2 - y^{n+1}y^{n-1} &= h^2r^{2n}\cos^2(n\varphi) - h^2r^{2n}\cos[(n+1)\varphi]\cos[(n-1)\varphi]\\ &= h^2r^{2n}\sin^2\varphi,\end{aligned}$$

whence

$$\frac{(y^n)^2 - y^{n+1}y^{n-1}}{\sin^2\varphi} = h^2r^{2n}.$$

The left-hand side of this equality tends to zero, for $n \to \infty$, in view of (4.32). Consequently,

$$h^2r^{2n} = T^2\Big(\frac{r^n}{n}\Big)^2 \to 0, \quad n \to \infty,$$

whence $|r| \leq 1$. So, until now, we proved that every root z of the characteristic polynomial ρ of a convergent multistep method is in the unit disc, $|z| \leq 1$.

Let now $z \in \mathbb{C}$ be a multiple root of ρ; then $\rho(z) = \rho'(z) = 0$. We know from our study of difference equations that the sequence $(y^\ell)_{\ell\in\mathbb{N}_0}$,

$$y^\ell := h\ell\,\mathrm{Re}(z^\ell) = h\ell r^\ell\cos(\ell\varphi), \quad \ell \in \mathbb{N}_0,$$

is a solution of (4.31). Furthermore, it satisfies (4.33), and, therefore, it satisfies also (4.32). If $\varphi = 0$ or $\varphi = \pi$, then

$$|y^n| = hnr^n = Tr^n \to 0, \quad n \to \infty,$$

whence $r < 1$. If $\varphi \neq 0, \varphi \neq \pi$, then for $x^\ell := \frac{1}{\ell h}y^\ell$ we have

$$\frac{(x^n)^2 - x^{n+1}x^{n-1}}{\sin^2\varphi} = r^{2n}.$$

Since $x^n = \frac{1}{T}y^n \to 0$, from the previous relation we infer that $r^{2n} \to 0$, i.e., $r < 1$. Thus we proved that every multiple root of the characteristic polynomial ρ of a convergent multistep method is in the interior of the unit disc, i.e., it has absolute value strictly less that one.

We infer that a convergent multistep method satisfies the root condition, and is thus stable. □

For the explicit Euler method, the implicit Euler method and the trapezoidal method, we have $\rho(z) = z - 1$, whence these methods are stable. For (4.2) and for the Simpson method, we have $\rho(z) = z^2 - 1$, whence these methods are also stable. Furthermore, the Adams methods ($\rho(z) = z^{k-1}(z-1)$), the Nyström methods and the Milne–Simpson methods ($\rho(z) = z^{k-2}(z^2-1)$) are obviously stable. Also, it is known that the backward difference methods (4.5) are stable, if and only if $1 \leq k \leq 6$.

In practice, if we compute with an unstable method, we soon observe a 'blow-up' of the numerical solution y^n. Due to roundoff errors the solution will always have components that do not remain bounded, as n increases. If we choose smaller step-size h, things get worse.

Second stability proof of multistep methods. We will conclude this section with a second proof of Proposition 4.1, due to Dahlquist. As we mentioned before, this proof is interesting both for historical reasons, since it was given prior the proof based on Lemma 4.1, but mainly because of the fact that it gives more information than the other proof, information that leads to improved results in some difficult cases.

Similarly to the first proof, the second is also based on an auxiliary result.

Lemma 4.2 (Preliminary result for Dahlquist's technique.) *Let $\rho, \rho(z) := \alpha_0 + \alpha_1 z + \cdots + \alpha_{k-1} z^{k-1} + \alpha_k z^k$, be a polynomial, with $\alpha_k = 1$, satisfying the root condition. Consider the constants γ_j, $j \in \mathbb{N}_0$, defined by the expansion*

$$\frac{1}{1 + \alpha_{k-1} z + \cdots + \alpha_0 z^k} = \gamma_0 + \gamma_1 z + \gamma_2 z^2 + \cdots . \tag{4.34}$$

Then, we have

$$\Gamma := \sup_{j \in \mathbb{N}_0} |\gamma_j| < \infty \tag{4.35}$$

and

$$\text{(4.36)}\qquad \begin{cases} \gamma_0 = 1, \\ \gamma_j + \alpha_{k-1}\gamma_{j-1} + \cdots + \alpha_{k-j}\gamma_0 = 0, \quad 1 \le j \le k, \\ \gamma_j + \alpha_{k-1}\gamma_{j-1} + \cdots + \alpha_0\gamma_{k-j} = 0, \quad j > k. \end{cases} \qquad \square$$

Let us note that setting $\gamma_{-1} := \cdots := \gamma_{-k} := 0$, relations (4.36) can be written in the form

$$\gamma_j + \alpha_{k-1}\gamma_{j-1} + \cdots + \alpha_0\gamma_{k-j} = 0, \quad j \neq 0.$$

In the sequel we will use this notation.

The proof of (4.35) requires tools from the theory of complex analysis and will not be given here; the interested reader can find it in the book of Henrici [18]. We are led to (4.36) by multiplying both sides of (4.34) by the denominator of the left-hand side and comparing the coefficients of the same powers of z on both sides.

The coefficients γ_j, $j \in \mathbb{N}_0$, are useful, since they allow us to "solve" inhomogeneous difference equations of the form

$$\text{(4.37)}\qquad y^{n+k} + \alpha_{k-1}y^{n+k-1} + \cdots + \alpha_0 y^n = \rho^n, \quad n \in \mathbb{N}_0,$$

that is, to express the components y^m, for $m \ge k$, in terms of the starting components $y^0, \ldots, y^{k-1}$ and $\rho^0, \ldots, \rho^{m-k}$. To this end, let us consider relations (4.37) for $n = 0, \ldots, m-k$. For $n = m-k-j$, $j = 0, \ldots, m-k$, we multiply (4.37) by γ_j and add the resulting relations. Denote by S_m the resulting sum on each side. The right-hand side gives, obviously,

$$\text{(4.38)}\qquad S_m = \sum_{j=0}^{m-k} \gamma_j \rho^{m-k-j}.$$

As far as the left-hand side is concerned, we have

$$\begin{aligned} S_m = {} & \gamma_0\big(y^m + \alpha_{k-1}y^{m-1} + \cdots + \alpha_0 y^{m-k}\big) \\ & + \gamma_1\big(y^{m-1} + \alpha_{k-1}y^{m-2} + \cdots + \alpha_0 y^{m-k-1}\big) \\ & + \\ & \vdots \\ & + \gamma_{m-k}\big(y^k + \alpha_{k-1}y^{k-1} + \cdots + \alpha_0 y^0\big). \end{aligned}$$

Therefore,

$$\begin{aligned} S_m = y^m &+ \big(\gamma_1 + \alpha_{k-1}\gamma_0\big)y^{m-1} + \cdots \\ &+ \big(\gamma_{m-k} + \alpha_{k-1}\gamma_{m-k-1} + \cdots + \alpha_0\gamma_{m-2k}\big)y^k \\ &+ \big(\alpha_{k-1}\gamma_{m-k} + \cdots + \alpha_0\gamma_{m-2k+1}\big)y^{k-1} + \cdots + \alpha_0\gamma_{m-k}y^0. \end{aligned}$$

Utilizing here (4.36), we easily infer that the above sum can be written in the form

$$S_m = y^m + \big(\alpha_{k-1}\gamma_{m-k} + \cdots + \alpha_0\gamma_{m-2k+1}\big)y^{k-1} + \cdots + \alpha_0\gamma_{m-k}y^0. \tag{4.39}$$

The equality of the left-hand sides of (4.38) and (4.39) leads to the "solution" of the difference equation (4.37),

$$\begin{aligned} y^m = &- \Big[\big(\alpha_{k-1}\gamma_{m-k} + \cdots + \alpha_0\gamma_{m-2k+1}\big)y^{k-1} + \cdots + \alpha_0\gamma_{m-k}y^0\Big] \\ &+ \sum_{j=0}^{m-k} \gamma_j \rho^{m-k-j}, \end{aligned} \tag{4.40}$$

in which the component y^m, for $m \geq k$, is expressed in terms of the data, i.e., the starting components $y^0, \ldots, y^{k-1}$ and the inhomogeneous terms $\rho^0, \ldots, \rho^{m-k}$. (We write "solution", because we use the constants γ_j, $j \in \mathbb{N}_0$, which we have to determine from the expansion (4.34).)

Now in the second proof of Proposition 4.1 we start from relations (4.40) with

$$\rho^n := h\big(\beta_k^n \psi^{n+k} + \cdots + \beta_0^n \psi^n\big) + \lambda^n,$$

see (4.22), and obtain

$$\begin{aligned} \psi^m = &- \Big[\big(\alpha_{k-1}\gamma_{m-k} + \cdots + \alpha_0\gamma_{m-2k+1}\big)\psi^{k-1} + \cdots + \alpha_0\gamma_{m-k}\psi^0\Big] \\ &+ h\sum_{j=0}^{m-k} \gamma_j\big(\beta_k^{m-k-j}\psi^{m-j} + \cdots + \beta_0^{m-k-j}\psi^{m-k-j}\big) + \sum_{j=0}^{m-k} \gamma_j \lambda^{m-k-j} \end{aligned}$$

or

$$\begin{aligned}\big(1-h\beta_k^{m-k}\big)\psi^m = h\Big[\big(\gamma_0\beta_{k-1}^{m-k}+\gamma_1\beta_k^{m-k-1}\big)\psi^{m-1}+\cdots+\gamma_{m-k}\beta_0^0\psi^0\Big]\\ -\Big[\big(\alpha_{k-1}\gamma_{m-k}+\cdots+\alpha_0\gamma_{m-2k+1}\big)\psi^{k-1}+\cdots+\alpha_0\gamma_{m-k}\psi^0\Big]\\ +\sum_{j=0}^{m-k}\gamma_j\lambda^{m-k-j}.\end{aligned}$$

Using here (4.35), we immediately get

$$|1-h\beta_k^{m-k}||\psi^m| \le C_1 h\sum_{j=0}^{m-1}|\psi^j| + C_2\sum_{j=0}^{k-1}|\psi^j| + \Gamma N\max_{0\le j\le m-k}|\lambda^j|,$$

for $m \le N$, with $C_1 := (k+1)B\Gamma$ and $C_2 := \Gamma\big(|\alpha_0|+\cdots+|\alpha_k|\big)$.

Let now $h_0 > 0$ be such that $Bh_0 < 1$. We immediately deduce from the previous estimate that there exists a constant C' such that, for $h < h_0$,

$$|\psi^m| \le C'\Big[h\sum_{j=0}^{m-1}|\psi^j| + N\max_{0\le j\le m-k}|\lambda^j| + C_2\sum_{j=0}^{k-1}|\psi^j|\Big], \tag{4.41}$$

$m = k,\dots,N$. The result follows now using the discrete Gronwall inequality; see Exercise 2.21. We will complete the proof, solving at the same time Exercise 2.21. We set

$$E := C'\Big[N\max_{0\le j\le N-k}|\lambda^j| + C_2\sum_{j=0}^{k-1}|\psi^j|\Big].$$

Then, obviously, there holds

$$|\psi^j| \le A(1+C'h)^j, \quad 0 \le j \le k-1.$$

We will now inductively show this inequality for $j = 0,\dots,N$. Assume that $k \le m < N$ and that there holds

$$|\psi^j| \le A(1+C'h)^j, \quad 0 \le j \le m-1.$$

Then, according to (4.41),

$$|\psi^j| \le C'hA\sum_{j=0}^{m-1}(1+C'h)^j + A = C'hA\frac{(1+C'h)^m - 1}{hC'} + A = A(1+C'h)^m.$$

Consequently, we indeed have

$$|\psi^j| \le A(1+C'h)^j, \quad 0 \le j \le N. \tag{4.42}$$

Now, for $j = 0, \dots, N$, we have

$$(1+C'h)^j \le \mathrm{e}^{jhC'} \le \mathrm{e}^{NhC'} = \mathrm{e}^{(b-a)C'},$$

whence (4.42) yields

$$|\psi^j| \le A\mathrm{e}^{(b-a)C'}, \quad k \le j \le N, \tag{4.43}$$

from which (4.23) follows immediately. □

Remark 4.1 (Schur and Routh–Hurwitz criteria.) The root condition can be easily checked, for instance, for quadratic polynomials ρ. It would be useful to have analytic conditions on the coefficients ensuring that a polynomial ρ of degree $k > 2$ satisfies the root condition. A useful theory is the so-called *Schur theory*[2]. We say that a polynomial π of degree k with complex coefficients a_i,

$$\pi(z) = a_0 + a_1 z + \cdots + a_k z^k, \tag{4.44}$$

where $a_0 \ne 0, a_k \ne 0$, is a *Schur polynomial*, if all its roots are contained in the open unit disc $D = \{z \in \mathbb{C} \colon |z| < 1\}$ in the complex plane. We say that π is a *simple von Neumann polynomial*, if all its roots are in the closed unit disc $\bar{D}$ and multiple roots are in the interior D, i.e., only simple root may have absolute value one. In other words, the root condition is equivalent to ρ being a simple von Neumann polynomial. Given a polynomial π, we consider the polynomial

$$\pi^\star(z) = \bar{a}_k + \bar{a}_{k-1} z + \cdots + \bar{a}_0 z^k, \tag{4.45}$$

[2] J. J. H. Miller: *On the location of zeros of certain classes of polynomials with applications to numerical analysis*. J. Inst. Math. Applic. **8** (1971) 397–406.

with $\bar{z}$ the conjugate of z. Obviously, we have $\pi^\star(z) = z^k \bar{\pi}(z^{-1})$. Furthermore, we consider the "reduced polynomial" π_1,

$$\pi_1(z) := \frac{1}{z}\big[\pi^\star(0)\pi(z) - \pi(0)\pi^\star(z)\big], \tag{4.46}$$

of degree at most $k-1$. Then, the following holds true:

i. The polynomial π is a Schur polynomial, if and only if $|\pi^\star(0)| > |\pi(0)|$ and π_1 is a Schur polynomial.

ii. The polynomial π is a simple von Neumann polynomial, if and only if
   a) either $|\pi^\star(0)| > |\pi(0)|$ and π_1 is a simple von Neumann polynomial,
   b) or $\pi_1 = 0$ and the derivative $\frac{d\pi}{dz}$ of π is a Schur polynomial.

With this theory we reduce the problem whether a polynomial is Schur or simple von Neumann to an analogue problem for a polynomial of degree by one less than the degree of the original polynomial, and we can proceed in the same way. These criteria are easy to use for $k = 3$ or 4, say.

Let us by the way mention also the *Routh–Hurwitz criterion* (see [20, p. 6]), that is also useful (mainly for $k = 2, 3, 4$, in practice). We consider a polynomial π, with *real coefficients*, given by (4.44), and perform the change of variables

$$w = \frac{1+z}{1-z}, \quad z = \frac{w-1}{w+1}, \tag{4.47}$$

that maps the unit circle $|w| = 1$ onto the imaginary axis $\operatorname{Re} z = 0$, the unit disc $D = \{w \in \mathbb{C} : |w| < 1\}$ onto the left complex half-plane $\operatorname{Re} z < 0$, the point $w = 1$ to $z = 0$, and the point $w = -1$ to the point at infinity $z = \infty$. Then, we can easily see that π is a Schur polynomial, if and only if the polynomial $\tilde{\pi}$,

$$\tilde{\pi}(z) = (1-z)^k \pi\big(\frac{1+z}{1-z}\big) = b_0 z^k + \cdots + b_k,$$

has the property that all its roots have *negative real parts*. Sufficient and necessary conditions for this are the *Routh–Hurwitz conditions* on the coefficients

b_i, which, in the case of $k = 2, 3, 4$, are:

$$\begin{aligned} k = 2: &\quad b_i > 0, \quad 0 \le i \le 2, \\ k = 3: &\quad b_i > 0, \quad 0 \le i \le 3, \quad b_1 b_2 - b_3 b_0 > 0, \\ k = 4: &\quad b_i > 0, \quad 0 \le i \le 4, \quad b_1 b_2 b_3 - b_0 b_3^2 - b_4 b_1^2 > 0. \end{aligned}$$

4.4 Order of accuracy, consistency and convergence of multistep methods

We consider problem (4.15) and assume that the solution y is sufficiently regular. For $t \in [a, b - kh]$ we define the quantity

$$\big(L_h y\big)(t) := \sum_{j=0}^{k} \big[\alpha_j y(t + jh) - h\beta_j y'(t + jh)\big]. \tag{4.48}$$

Comment. Substituting the approximate solution in (4.16) by the corresponding nodal values of the exact solution, we do not have an equality any more; instead an error occurs, the so-called *consistency error*,

$$\sum_{j=0}^{k} \big[\alpha_j y(t^{n+j}) - h\beta_j f\big(t^{n+j}, y(t^{n+j})\big)\big],$$

which we can also write in the form

$$\sum_{j=0}^{k} \big[\alpha_j y(t^{n+j}) - h\beta_j y'(t^{n+j})\big] = \sum_{j=0}^{k} \big[\alpha_j y(t^n + jh) - h\beta_j y'(t^n + jh)\big].$$

Replacing t^n by t in this relation we are led to $\big(L_h y\big)(t)$. Thus, the quantity $L_h y(t)$ is the amount by which the exact solution misses being approximate solution, in the sense that it misses satisfying the scheme (4.4) at the point t. Consequently, the quantity $L_h y(t^n)$ is the exact analogue of the consistency error (local error)

$$y(t^{n+1}) - \big[y(t^n) + h\Phi\big(t^n, y(t^n); h\big)\big]$$

in the case of Runge–Kutta methods.

Definition 4.4 (Order of accuracy of multistep methods.) Let $y : [a,b] \to \mathbb{R}$ be an arbitrary, sufficiently regular function. If p is the largest integer, for which an estimate of the form

$$\exists C = C(y) \ \ \forall t \in [a, b-kh] \quad |(L_h y)(t)| \leq C h^{p+1},$$

holds, with $L_h y$ as defined in (4.48), then we say that the *order of accuracy* of the multistep method (4.4) is p. If the order of accuracy is at least one, the method is called *consistent*. □

In the case of implicit Runge–Kutta methods, the determination of the order of accuracy was a rather difficult problem; this is due to the fact that the error could not be expressed in terms of one function y only; the function f was also involved. In contrast, in the case of multistep methods, and due to the fact that the consistency error is expressed in terms of y only, this is a rather trivial problem. Indeed, by Taylor expanding $y(t+jh)$ and $y'(t+jh)$ around t, we obtain

$$\big(L_h y\big)(t) = C_0 y(t) + C_1 h y'(t) + C_2 h^2 y''(t) + \cdots$$

with constants C_j independent of y, t and h, and depending only on the concrete method. Obviously, the order of the method (4.4) is p, if and only if

$$C_0 = C_1 = \cdots = C_p = 0 \quad \text{και} \quad C_{p+1} \neq 0.$$

We can easily see that

$$\begin{aligned} C_0 &= \alpha_0 + \alpha_1 + \cdots + \alpha_k \\ C_1 &= \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k - (\beta_0 + \cdots + \beta_k) \end{aligned}$$

and, for $j \geq 2$,

$$\begin{aligned} C_j = \frac{1}{j!}\big(\alpha_1 + 2^j\alpha_2 + 3^j\alpha_3 + \cdots + k^j\alpha_k\big) \\ - \frac{1}{(j-1)!}\big(\beta_1 + 2^{j-1}\beta_2 + 3^{j-1}\beta_3 + \cdots + k^{j-1}\beta_k\big). \end{aligned}$$

We infer that a sufficient and necessary condition for the consistency of a method (that is for $p \geq 1$) is that the following relations

$$\begin{cases} \alpha_0 + \alpha_1 + \cdots + \alpha_k = 0 \\ \alpha_1 + 2\alpha_2 + \cdots + k\alpha_k - (\beta_0 + \cdots + \beta_k) = 0 \end{cases} \tag{4.49}$$

hold true. Utilizing the characteristic polynomial ρ, that we already introduced, and the polynomial $\sigma, \sigma(z) := \beta_k z^k + \cdots + \beta_0$, we write (4.49) in the form

$$\rho(1) = 0, \quad \rho'(1) = \sigma(1).$$

Calculating the constants C_j for the multistep methods we saw as examples in section 4.1, we easily see that the order of accuracy of the Euler methods is one, of the trapezoidal method it is two, of method (4.2) also two, of the Simpson method (4.3) four, while for the $k-$step backward difference method is k.

Let us note that relations $C_0 = C_1 = \cdots = C_p = 0$ can be equivalently written in the form

$$\sum_{i=0}^{k} i^j \alpha_i = j \sum_{i=0}^{k} i^{j-1} \beta_i, \quad j = 0, 1, \ldots, p. \tag{4.50}$$

Now, we have

$$\rho(\mathrm{e}^x) = \sum_{i=0}^{k} \alpha_i \Big(\sum_{j=0}^{p} \frac{(ix)^j}{j!} \Big) + O(x^{p+1}) = \sum_{j=0}^{p} \frac{1}{j!} \Big(\sum_{i=0}^{k} i^j \alpha_i \Big) x^j + O(x^{p+1}),$$

for $x \to 0$, and easily infer that (4.50) can be equivalently written as

$$\rho(\mathrm{e}^x) = x\sigma(\mathrm{e}^x) + O(x^{p+1}) \quad \text{as } x \to 0. \tag{4.51}$$

If the order of the method is p, then we obviously have

$$\big(L_h y\big)(t) = C_{p+1} h^{p+1} y^{p+1}(t) + O(h^{p+2}); \tag{4.52}$$

so C_{p+1} is a factor of the leading term in the power expansion of $L_h y$. This constant is thus important and it is usually referred to as *error constant* of the method. Of course, since we can multiply the method by any non-vanishing

constant and obtain still the same method, while C_{p+1} is multiplied by the constant, we need somehow to "normalize" the method; the most common normalizations are $\beta_k + \cdots + \beta_0 = 1$ or $\alpha_k = 1$.

As we will next see, the consistency of a method is necessary for its convergence.

Proposition 4.3 (Convergence implies consistency.) *If method* (4.4) *converges, then it is consistent.*

Proof. Let us first show that $\rho(1) = 0$. Since the coefficients $\beta_k, \ldots, \beta_0$ do not enter in this relation, we can choose a problem with $f = 0$. So, we consider the initial value problem

$$\begin{cases} y'(t) = 0, \quad 0 \le t \le T, \\ y(0) = 1; \end{cases}$$

its solution is, obviously, $y(t) = 1, t \in [0, T]$. In this case the method (4.4) takes the form

$$\begin{cases} y^0, y^1, \ldots, y^{k-1} \quad \text{given}, \\ \alpha_k y^{n+k} + \alpha_{k-1} y^{n+k-1} + \cdots + \alpha_0 y^n = 0, \quad n = 0, \ldots, N-k. \end{cases} \tag{4.53}$$

Choosing $y^0 := \cdots := y^{k-1} := 1$ in (4.53), we get, in view of the convergence of the method, $\alpha_k + \cdots + \alpha_0 = 0$, i.e., that indeed $\rho(1) = 0$.

Next, we will show that $\rho'(1) = \sigma(1)$. In this relation the coefficients $\beta_k, \ldots, \beta_0$ do enter, and we need to choose a problem with $f \neq 0$. The simplest such f is $f(t, y) = 1$. So, we consider the initial value problem

$$\begin{cases} y'(t) = 1, \quad 0 \le t \le T, \\ y(0) = 0; \end{cases}$$

its solution is $y(t) = t, 0 \le t \le T$. In this case the method (4.4) reads

$$\begin{cases} y^0, y^1, \ldots, y^{k-1} \quad \text{given}, \\ \alpha_k y^{n+k} + \cdots + \alpha_0 y^n = h(\beta_k + \cdots + \beta_0), \quad n = 0, \ldots, N-k. \end{cases} \tag{4.54}$$

Since our method converges, every solution of (4.54) for which

$$\lim_{h\to 0} y^j = 0, \quad j = 0, \dots, k-1, \tag{4.55}$$

must also satisfy the relation

$$\lim_{N\to\infty} y^N = T, \quad Nh = T. \tag{4.56}$$

Let $M := \frac{\sigma(1)}{\rho'(1)}$. This constant is well defined, since the method is convergent, whence also stable, see Proposition 4.2, whence since $\rho(1) = 0$ we have $\rho'(1) \neq 0$. We now define

$$y_h^n := nhM, \quad n = 0, \dots, N, \quad h := \frac{T}{N}. \tag{4.57}$$

Then

$$\begin{aligned}\sum_{j=0}^{k} \alpha_j y_h^{n+j} &= hM \sum_{j=0}^{k} (n+j)\alpha_j = hM \sum_{j=0}^{k} j\alpha_j \\ &= hM\rho'(1) = h\sigma(1) = h\sum_{j=0}^{k} \beta_j;\end{aligned}$$

in the second equality we used the fact that $\alpha_k + \cdots + \alpha_0 = 0$. Consequently, (4.54) is satisfied. It is also obvious that (4.55) is satisfied. Therefore, according to (4.56),

$$T = \lim_{N\to\infty} (hNM) = TM,$$

whence $M = 1$, i.e., $\rho'(1) = \sigma(1)$, and thus $C_1 = 0$. We infer that the method is indeed consistent. □

According to Propositions 4.2 and 4.3, stability and consistency are necessary conditions for the convergence of a multistep method. We will now see that these two conditions are also *sufficient* for the convergence of a method.

Theorem 4.1 (Error estimate.) *Let the $k-$step method* (4.4) *be stable and have order of accuracy $p \geq 1$. Let $y \in C^{p+1}[a,b]$ be the solution of initial value problem* (4.15)*. Then, there exists $h_0 > 0$, such that, for $0 < h \leq h_0$, there holds*

$$\max_{0\leq n\leq N} |y(t^n) - y^n| \leq C\Big[\max_{0\leq j\leq k-1} |y(t^j) - y^j| + h^p \max_{a\leq t\leq b} |y^{p+1}(t)|\Big] \tag{4.58}$$

with a constant C independent of h, N, y. Of course, we assumed here $hN = b - a, t^n = a + nh, n = 0, \dots, N$, and that the approximations y^n satisfy (4.4).

Proof. Let

$$E^n := \sum_{j=0}^{k} \big[\alpha_j y(t^n + jh) - h\beta_j y'(t^n + jh)\big], \quad n = 0, \dots, N-k,$$

be the consistency error of the method. Since, according to our hypothesis, the order of the method is p, we have

$$\max_{0 \le n \le N-k} |E^n| \le C' h^{p+1} \|y^{(p+1)}\|_\infty, \tag{4.59}$$

with a constant C', independent of h and y, as we easily see by Taylor expanding; see (4.48) and Definition 4.4. With $\varepsilon^n := y(t^n) - y^n, n = 0, \dots, N$, we have

$$\begin{aligned}
&\alpha_k \varepsilon^{n+k} + \alpha_{k-1}\varepsilon^{n+k-1} + \cdots + \alpha_0 \varepsilon^n = \\
&= \big[\alpha_k y(t^{n+k}) + \cdots + \alpha_0 y(t^n)\big] - \big(\alpha_k y^{n+k} + \cdots + \alpha_0 y^n\big) \\
&= h\big[\beta_k y'(t^{n+k}) + \cdots + \beta_0 y'(t^n)\big] - h\big(\beta_k f^{n+k} + \cdots + \beta_0 f^n\big) + E^n \\
&= h\Big[\beta_k\big[f\big(t^{n+k}, y(t^{n+k})\big) - f^{n+k}\big] + \cdots + \beta_0\big[f\big(t^n, y(t^n)\big) - f^n\big]\Big] + E^n.
\end{aligned}$$

Now, we set, for $m = 0, \dots, N$,

$$g^m := \begin{cases} \dfrac{f\big(t^m, y(t^m)\big) - f^m}{\varepsilon^m}, & \text{if } \varepsilon^m \neq 0, \\ 0 \qquad\qquad\quad , & \text{if } \varepsilon^m = 0, \end{cases}$$

and write the above relation in the form

$$\alpha_k \varepsilon^{n+k} + \cdots + \alpha_0 \varepsilon^n = h\big(\beta_k g^{n+k}\varepsilon^{n+k} + \cdots + \beta_0 g^n \varepsilon^n\big) + E^n, \tag{4.60}$$

for $n = 0, \dots, N-k$. Now, obviously,

$$|g^n| \le L, \quad n = 0, \dots, N,$$

with L the Lipschitz condition of f with respect to its second variable. Thus, with

$$\beta_i^n := \beta_i g^{n+i}, \quad i = 0, \ldots, k, \quad n = 0, \ldots, N-k,$$

there holds

$$\max_{i,n} |\beta_i^n| \le L \max_i |\beta_i| =: B < \infty. \tag{4.61}$$

Now (4.60) can be written as

$$\alpha_k \varepsilon^{n+k} + \cdots + \alpha_0 \varepsilon^n = h\big(\beta_k^n \varepsilon^{n+k} + \cdots + \beta_0^n \varepsilon^n\big) + E^n, \quad n = 0, \ldots, N-k,$$

and Proposition 4.1 gives

$$\max_{0 \le n \le N} |\varepsilon^n| \le C\Big[N \max_{0 \le n \le N-k} |E^n| + \max_{0 \le j \le k-1} |\varepsilon^j|\Big].$$

Using here the consistency estimate (4.59) and the fact that $Nh = b - a$, we obtain the desired error estimate (4.58). □

The result in Theorem 4.1 suggest how to compute the starting approximations $y^0, \ldots, y^{k-1}$, needed in the multistep method (4.4), in such a way that the order of the method will not be reduced. If we let $y^0 := y_0$ and compute $y^1, \ldots, y^{k-1}$ with a Runge–Kutta method of order at least $p-1$ ($p-1$ suffices, since the method is applied only $k-1$ times), then, according to (4.58), we will have

$$\max_{0 \le n \le N} |y(t^n) - y^n| \le C h^p,$$

with a constant C, independent of h.

We close this section mentioning an important result of Dahlquist (the proof can be found in the book of Henrici [18]), namely that the highest attainable order of accuracy of a *stable* $k-$step method is $p = k+1$, if k is odd, and $p = k+2$, if k is even. Therefore, the trapezoidal method ($k = 1$, $p = 2$) and Simpson's method ($k = 2$, $p = 4$) are stable method of highest order of accuracy, for $k = 1$ and 2 steps, respectively. Notice that stable methods of highest accuracy are always implicit. The highest order of accuracy of an explicit $k-$step method is $p = k$.

4.5 Absolute stability of multistep methods

In this section we will briefly investigate absolute stability properties of multistep methods. More precisely, in the first subsection, stability properties that make the methods suitable for the discretization of stiff linear equations with constant coefficients, and in the second, corresponding properties for nonlinear equations.

4.5.1 Absolute stability

Since in the case of multistep methods it does not make sense to compare two consecutive approximations, we use the interpretation of Remark 2.8 for the A–stability, i.e., we require the method to yield *bounded* approximations, when applied to test problem (1.18).

Applying a $k-$step method, described by the constants $\alpha_k, \ldots, \alpha_0$, $\beta_k, \ldots, \beta_0$, to the test initial value problem

$$\begin{cases} y' = \lambda y, \quad t \in [0, \infty), \\ y(0) = 1, \end{cases} \tag{4.62}$$

with $\lambda \in \mathbb{C}$, we obtain the approximations $(y^n)_{n\in\mathbb{N}_0}$, satisfying

$$\sum_{j=0}^{k} (\alpha_j - h\lambda\beta_j) y^{n+j} = 0, \quad n \geq 0. \tag{4.63}$$

According to the theory of homogeneous linear difference equations with constant coefficients, which we presented in section 4.2, the approximations y^n, $n \in \mathbb{N}$, remain bounded, if and only if, for the given h, the polynomial π,

$$\pi(\zeta, h\lambda) := \rho(\zeta) - h\lambda\sigma(\zeta), \tag{4.64}$$

satisfies the roof condition,

(4.65) polynomial $\pi(\cdot, h\lambda)$ satisfies the root condition.

Therefore, the problem of the determination of the stability region of method (4.4) reduces to the study of the roots of π as functions of the complex parameter $h\lambda$ and in the formulation of conditions ensuring the root condition (4.65).

Let us see some examples. For the method (4.2) we have $\rho(\zeta) = \zeta^2 - 1, \sigma(\zeta) = 2\zeta$, whence $\pi(\zeta) = \zeta^2 - 2h\lambda\zeta - 1$, and infer easily that π satisfies the root condition only for $h\lambda = 0$. (Indeed, if $h\lambda \neq 0$, the roots of π are not on the unit circle and their product is -1. Consequently, it is not possible that both belong to the unit disc, $|z| \leq 1$.) Therefore, the stability region of the method consists only of the origin, $z = 0$. The same holds true for the Simpson method (4.3). We already know that the implicit Euler method and the trapezoidal method are A–stable.

Example 4.1 The two-step backward difference method (BDF) $\frac{3}{2}y^{n+2} - 2y^{n+1} + \frac{1}{2}y^n = hf^{n+2}$ is A–stable.

We will now see that the two-step BDF method is A–stable. According to our remarks above, the proof reduces to proving that the polynomial π,

$$\pi(\zeta) := \big(\frac{3}{2} - a - b\mathrm{i}\big)\zeta^2 - 2\zeta + \frac{1}{2},$$

with $a, b \in \mathbb{R}, a \leq 0$, satisfies the root condition. (We set $h\lambda = a + b\mathrm{i}$.) We distinguish two cases, $a = b = 0$ and $(a, b) \neq (0, 0)$. In the first case π coincides with the characteristic polynomial ρ of the method; its roots are 1 and $1/3$, whence the root condition is satisfied. We assume now that $(a, b) \neq (0, 0)$ and utilizing the Schur theory (see Remark 4.1) will again show that the root condition is satisfied. More precisely, we will now show that the roots of π are in the interior of the unit disc in the complex plane, i.e., that π is a *Schur polynomial*. This is equivalent to two conditions, $|\pi^\star(0)| > |\pi(0)|$, with $\pi^\star(\zeta) := \frac{1}{2}\zeta^2 - 2\zeta + \big(\frac{3}{2} - a + b\mathrm{i}\big)$, i.e.,

$$\big|\frac{1}{2}\big| < \big|\frac{3}{2} - a + b\mathrm{i}\big|,$$

which is obviously valid in our case, and the condition that the reduced polynomial π_1, $\pi_1(\zeta) = \big[\pi^\star(0)\pi(\zeta) - \pi(0)\pi^\star(\zeta)\big]/\zeta$, is a Schur polynomial. Now

$$\pi_1(\zeta) = (2 - 3a + a^2 + b^2)\zeta - 2(1 - a + b\mathrm{i}),$$

and it is Schur polynomial, if and only if $|2(1 - a + b\mathrm{i})| < |2 - 3a + a^2 + b^2|$. The last condition can be equivalently written in the form

$$a^4 + b^4 + 9a^2 + 2a^2b^2 - 6a^3 - 6ab^2 - 4a > 0,$$

relation that is obviously satisfied for the values of a and b we are considering here. □

More generally, the $k-$step backward difference methods (BDF) (4.5), $1 \leq k \leq 6$, have the following stability properties. For $k = 1$ (implicit Euler method) and $k = 2$ they are, as we have seen, A–stable. For $3 \leq k \leq 6$ they are $A(\vartheta)-$stable (see Definition 2.3) with angles $\vartheta_k, 3 \leq k \leq 6$, approximately equal to $86°, 73°, 52°, 18°$, for $k = 3, 4, 5, 6$, respectively. Therefore, they are suitable for stiff systems with eigenvalues lying in the corresponding regions (sectors) S_{ϑ_k}. Program ode15s of MATLAB is based on the BDF methods and has a mechanism incorporated to control the step-size and the choice of the order of accuracy of the method from $p = 1$ to $p = 5$.

There exists an extensive theory of the absolute stability properties of multistep methods, initiated in the work of G. Dahlquist in the 1960's. Dahlquist proved that there are no (consistent) explicit A–stable multistep methods and that the order of accuracy p of an A–stable multistep method can not exceed two (Dahlquist barrier). Therefore, we have a severe restriction in the order of an A–stable multistep method, a fact that highlights, e.g., the importance of implicit Runge–Kutta methods, among which there exist A–stable methods of any order of accuracy (for instance, the $q-$stage Gauss–Legendre methods of order $p = 2q$ and the RK Radau IIA methods of order $p = 2q - 1$).

The situation is much better if we do not impose A–stability condition but confine ourselves with $A(\vartheta)-$stability, for $0 < \vartheta < \pi/2$, or A_0-stability. (For the discretization in time of parabolic p.d.e's, usually A_0- or $A(\vartheta)-$stability suffices. This is not the case for hyperbolic equations though, in which case the eigenvalues of A are imaginary; see Exercise 3.23.) It is known (Widlund) that there do not exist explicit $A(\vartheta)-$stable methods and that the only $A(\vartheta)-$stable $k-$step method of order p exceeding k is the trapezoidal method $(k = 1, p = 2)$. It is also known (Grigorieff–Schroll) that for *any* $\vartheta \in (0, \pi/2)$ and $k \in \mathbb{N}$, there exist $A(\vartheta)-$stable methods of order $p = k$. If we restrict ourselves to A_0-stable methods, it is known (Cryer) that again A_0-stable methods are implicit, and that their order p can not exceed k with only one exception, the trapezoidal method.

4.5.2 G–stability

Let us now turn our attention to systems of nonlinear o.d.e's. As in the case of RK methods, we consider initial value problems for systems of o.d.e's,

$$\begin{cases} y' = f(t,y), \quad t \geq 0, \\ y(0) = y_0, \end{cases} \tag{4.66}$$

with $f : [0,\infty) \times \mathbb{R}^m \to \mathbb{R}^m$ a function satisfying the one-sided Lipschitz condition, with respect to its second variable,

$$\big(f(t,y) - f(t,z), y - z\big) \leq 0 \quad \forall t \geq 0,\ y,z \in \mathbb{R}^m. \tag{4.67}$$

We denoted here by $(\cdot,\cdot)$ the Euclidean inner product in $\mathbb{R}^m$. In the sequel, as we did up to now, we will denote by $\|\cdot\|$ the Euclidean norm in $\mathbb{R}^m$.

In the case of RK methods the concept of B–stability is based on the inequality

$$\|y^{n+1} - z^{n+1}\| \leq \|y^n - z^n\|,$$

which we require to hold whenever (4.67) is satisfied. In the case of multistep methods such an inequality does not make much sense, since the approximation y^{n+1}, for instance, does not only depend on the approximation y^n, but also on previous approximations.

An analogue to the B–stability concept in the case of RK methods, is the so-called G–stability in the case of multistep methods. The definition of G–stability is based on the so-called *one-leg version* of multistep methods. Thus, we will briefly discuss this version. With the usual notation, we consider a $k-$step method, described by the constants $\alpha_k,\ldots,\alpha_0,\ \beta_k,\ldots,\beta_0$, the step of which is given as

$$\sum_{i=0}^{k} \alpha_i y^{n+i} = h \sum_{i=0}^{k} \beta_i f(t^{n+i}, y^{n+i}). \tag{4.68}$$

We assume that the constants have been normalized in such a way that $\beta_0 + \cdots + \beta_k = 1$. We note that this is always possible for stable and consistent

multistep methods. Furthermore, we assume that the method is *irreducible*, i.e., the polynomials ρ and σ,

$$\rho(z) = \sum_{i=0}^{k} \alpha_i z^i, \quad \sigma(z) = \sum_{i=0}^{k} \beta_i z^i,$$

do not have roots in common, that is, as we say, they are *mutually irreducible*. Then the method (4.68) corresponds to the one-leg (multistep) method

$$\sum_{i=0}^{k} \alpha_i y^{n+i} = hf\Big(\sum_{i=0}^{k} \beta_i t^{n+i}, \sum_{i=0}^{k} \beta_i y^{n+i}\Big). \tag{4.69}$$

Notice that f is evaluated only at one point in (4.69). In general, methods (4.68) and (4.69) yield different approximations. However, when applied to linear o.d.e's with constant coefficients, like problem (4.62), then the two methods coincide. Furthermore, in the particular case of BDF methods (4.68) and (4.69) coincide.

Example 4.2 The trapezoidal method.
The one-leg version corresponding to the trapezoidal method

$$y^{n+1} - y^n = \frac{h}{2}\big[f(t^n, y^n) + f(t^{n+1}, y^{n+1})\big] \tag{4.70}$$

is

$$y^{n+1} - y^n = hf\Big(\frac{1}{2}(t^n + t^{n+1}), \frac{1}{2}(y^n + y^{n+1})\Big), \tag{4.71}$$

that is the midpoint method. Between the approximations of methods (4.70) and (4.71) we have the following relation: If (y^n) is a solution of the midpoint method (4.71), then the quantities

$$\tilde{y}^n := \frac{1}{2}(y^n + y^{n+1}), \quad \tilde{t}^n := \frac{1}{2}(t^n + t^{n+1})$$

satisfy the trapezoidal method (4.70). Conversely, if $(\tilde{t}^n, \tilde{y}^n)$ is a solution of the trapezoidal method (4.70), then the quantities

$$y^n := \tilde{y}^n - \frac{h}{2} f(\tilde{t}^n, \tilde{y}^n), \quad t^n := \tilde{t}^n - \frac{h}{2}$$

satisfy (4.71). □

For elements $X \in (\mathbb{R}^m)^k$ we use the notation $X = (x^1, \ldots, x^k)^T$ with $x^i \in \mathbb{R}^m, i = 1, \ldots, k$. Furthermore, for a $k \times k$ real matrix $G = (g_{ij})$ we denote by $GX \in (\mathbb{R}^m)^k$ the vector

$$GX = \Big(\sum_{j=1}^{k} g_{1j} x^j, \ldots, \sum_{j=1}^{k} g_{kj} x^j \Big)^T.$$

If $(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$ denote the Euclidean inner products in $\mathbb{R}^m$ and in $(\mathbb{R}^m)^k$, respectively, then we immediately infer that

$$\langle GX, Y \rangle = \sum_{i,j=1}^{k} g_{ij} (x^i, y^j).$$

We assume now that the matrix G is symmetric and positive definite. Then $\langle G\cdot, \cdot \rangle$ is an inner product in $(\mathbb{R}^m)^k$, and, in particular, $\| \cdot \|_G$,

$$\|X\|_G := \langle GX, X \rangle^{1/2},$$

is a norm in $(\mathbb{R}^m)^k$. We are now in a position to define the G–stability concept for one-leg (multistep) methods.

Definition 4.5 (G–stability. Dahlquist) The one-leg method (4.69) is called *G–stable*, if there exists a symmetric and positive definite matrix $G \in \mathbb{R}^{k,k}$, such that

$$\|Y^{n+1} - Z^{n+1}\|_G \leq \|Y^n - Z^n\|_G$$

for all $h > 0$ and for all initial value problems (4.66) for which (4.67) is satisfied. We used here the notation $Y^n = (y^{n+k-1}, \ldots, y^n)^T$. □

Example 4.3 The two-step backward difference method is G–stable.
With the usual notation, the two-step BDF method for problem (4.66) is

$$\begin{cases} y^0, \ y^1 \text{ given,} \\ \dfrac{3}{2} y^{n+2} - 2y^{n+1} + \dfrac{1}{2} y^n = h f(t^{n+2}, y^{n+2}), \quad n = 0, \ldots, N-2. \end{cases} \tag{4.72}$$

Considering now another approximate solution $(z^n)_{0 \leq n \leq N}$ and setting $\varepsilon^m := y^m - z^m$, we have

$$\frac{3}{2} \varepsilon^{n+2} - 2\varepsilon^{n+1} + \frac{1}{2} \varepsilon^n = h \big[f\big(t^{n+2}, y^{n+2}\big) - f\big(t^{n+2}, z^{n+2}\big) \big].$$

Taking here the inner product with ε^{n+2} and using (4.67), we get

$$(\frac{3}{2}\varepsilon^{n+2} - 2\varepsilon^{n+1} + \frac{1}{2}\varepsilon^{n}, \varepsilon^{n+2}) \le 0. \tag{4.73}$$

Now,

$$\begin{aligned}(\frac{3}{2}\varepsilon^{n+2} - 2\varepsilon^{n+1} + \frac{1}{2}\varepsilon^{n}, \varepsilon^{n+2}) = \frac{5}{4}\|\varepsilon^{n+2}\|^2 - \|\varepsilon^{n+1}\|^2 - \frac{1}{4}\|\varepsilon^{n}\|^2 \\ - \big[(\varepsilon^{n+2}, \varepsilon^{n+1}) - (\varepsilon^{n+1}, \varepsilon^{n})\big] + \frac{1}{4}\|\varepsilon^{n+2} - 2\varepsilon^{n+1} + \varepsilon^{n}\|^2,\end{aligned}$$

whence

$$\begin{aligned}(\frac{3}{2}\varepsilon^{n+2} - 2\varepsilon^{n+1} + \frac{1}{2}\varepsilon^{n}, \varepsilon^{n+2}) \ge \big[\frac{5}{4}\|\varepsilon^{n+2}\|^2 - (\varepsilon^{n+2}, \varepsilon^{n+1}) + \frac{1}{4}\|\varepsilon^{n+1}\|^2\big] \\ - \big[\frac{5}{4}\|\varepsilon^{n+1}\|^2 - (\varepsilon^{n+1}, \varepsilon^{n}) + \frac{1}{4}\|\varepsilon^{n}\|^2\big].\end{aligned} \tag{4.74}$$

Now (4.73) and (4.74) imply

$$\frac{5}{4}\|\varepsilon^{n+2}\|^2 - (\varepsilon^{n+2}, \varepsilon^{n+1}) + \frac{1}{4}\|\varepsilon^{n+1}\|^2 \le \frac{5}{4}\|\varepsilon^{n+1}\|^2 - (\varepsilon^{n+1}, \varepsilon^{n}) + \frac{1}{4}\|\varepsilon^{n}\|^2. \tag{4.75}$$

With $\mathcal{E}^m := (\varepsilon^m, \varepsilon^{m-1})^T$ and the symmetric and positive definite matrix G,

$$G := \frac{1}{4}\begin{pmatrix} 5 & -2 \\ -2 & 1 \end{pmatrix},$$

relation (4.75) takes the final form

$$\big\langle G\mathcal{E}^{n+2}, \mathcal{E}^{n+2}\big\rangle \le \big\langle G\mathcal{E}^{n+1}, \mathcal{E}^{n+1}\big\rangle, \tag{4.76}$$

which is the desired property, where $\langle\cdot,\cdot\rangle$ is the Euclidean inner product in $(\mathbb{R}^m)^2 = \mathbb{R}^{2m}$,

$$\Big\langle \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} z \\ \omega \end{pmatrix} \Big\rangle = (x, z) + (y, \omega).$$

Therefore, the two-step BDF method is indeed G–stable. □

In view of the fact that a multistep method (4.68) coincides with its corresponding one-leg version (4.69) for the initial value problem (4.62), we immediately infer that the G–stability implies A–stability; see also the discussion following (4.69). It is in fact known that an irreducible multistep method

(4.68) is A–stable, if and only if the corresponding one-leg version (4.69) is G–stable. We recall that there do not exist A–stable multistep methods of order higher than two.

A final remark concerning the assumption that the multistep method is irreducible: Assume that the method (4.68) is reducible. Let the polynomials ρ and σ have a polynomial τ as common divisor. We set

$$\tilde{\rho} := \frac{\rho}{\tau}, \quad \tilde{\sigma} := \frac{\sigma}{\tau}.$$

Polynomials $\tilde{\rho}$ and $\tilde{\sigma}$ produce now a new, simpler multistep method. With the *shift operator* E, $Ex^n = x^{n+1}$, and the usual notation $f^n = f(t^n, y^n)$, we write the new method in the form

$$\tilde{\rho}(E)y^n = h\tilde{\sigma}(E)f^n \tag{4.77}$$

and (4.68) as

$$\rho(E)y^n = h\sigma(E)f^n. \tag{4.78}$$

Multiplying (4.77) by $\tau(E)$ we immediately infer that every solution (y^n) of the simpler method (4.77) is also a soluten of (4.78). The two methods are essentially the same. It can also be easily seen that (4.77) and (4.78) have the same order of accuracy. Therefore, reducible multistep methods are not of interest; thus our assumption that the method (4.68) is irreducible is not restrictive.

Exercises

4.1 Determine the coefficients $\alpha_0, \alpha_1, \alpha_2$, such that the two-step method

$$\alpha_2 y^{n+2} + \alpha_1 y^{n+1} + \alpha_0 y^n = hf^{n+2}$$

has order of accuracy two. Is the resulting method stable?

4.2 Prove the relation

$$C_j = \frac{1}{j!}\left(\alpha_1 + 2^j\alpha_2 + \cdots + k^j\alpha_k\right) - \frac{1}{(j-1)!}\left(\beta_1 + 2^{j-1}\beta_2 + \cdots + k^{j-1}\beta_k\right), \quad j \geq 2,$$

given a little bit before (4.49).

4.3 Prove that the error constant C_{p+1} of method (4.4) is the coefficient of $(z-1)^{p+1}$ of the expansion of function $\rho(z) - \sigma(z)\ln(z)$ in powers of $z-1$. As an example, show that $C_3 = -1/12$ for the trapezoidal method.
[*Hint*: Choosing $y(t) = \mathrm{e}^t, t = t^n$, in (4.52), we have

$$\sum_{j=0}^{k} \left(\alpha_j \mathrm{e}^{t+jh} - h\beta_j \mathrm{e}^{t+jh}\right) \approx C_{p+1} h^{p+1} \mathrm{e}^t.$$

Setting $\mathrm{e}^h = z$, i.e., $h = \ln z, z > 1$, we obtain

$$\rho(z) - \ln z\ \sigma(z) = \sum_{j=0}^{k} \left(\alpha_j z^j - \ln z\ \beta_j z^j\right) \approx C_{p+1}(\ln z)^{p+1},\ z \downarrow 1.$$

The result follows by expanding $\ln z$ in powers of $z-1$: $\ln z = (z-1) - (z-1)^2/2 + (z-1)^3/3 + \cdots$.]

4.4 Let $y^0 := 0$, $y^1 := 1$, and $y^{n+2} := y^{n+1} + y^n, n \in \mathbb{N}_0$. Give in closed form the nth term of the sequence $(y^n)_{n\in\mathbb{N}_0}$. This sequence is known as *Fibonacci sequence*.

4.5 Determine the coefficients $\alpha_1, \alpha_0, \beta_2, \beta_1, \beta_0$ of the general two-step method

$$y^{n+2} + \alpha_1 y^{n+1} + \alpha_0 y^n = h\left(\beta_2 f^{n+2} + \beta_1 f^{n+1} + \beta_0 f^n\right)$$

such that its order p be at least two, at least three, and at least four, respectively. Does there exist a method with $p = 4$? Does there exist a method with $p = 5$? Are the resulting methods stable?

4.6 If $\rho(z) = z^4 - 1$, determine a polynomial σ of degree four, such that the method (ρ, σ) has the highest possible order. What is the order and what the error constant?

4.7 Determine the coefficients α_j, β_j of the three-step method of the highest possible order. What is its order? Is the method stable?

4.8 Express the coefficients of all two-step methods of order at least three in terms of the parameter β_0. For which values of β_0 are the methods stable? Express the error constant in terms of β_0 (for the values of β_0 yielding stable methods). What property distinguished Simpson's method, among all these methods?

4.9 Determine the order of accuracy of methods (4.7) – (4.10).

4.10 Determine the stable $4-$step methods of order $p = 6$.

4.11 Consider the quadratic polynomial $\alpha z^2 + \beta z + \gamma, \alpha, \beta, \gamma \in \mathbb{C}$. Show that sufficient and necessary conditions for it to be a simple von Neumann polynomial are $|\alpha| > |\gamma|$ and $|\bar{\alpha}\beta - \bar{\beta}\gamma| \leq |\alpha|^2 - |\gamma|^2$.

4.12 Prove that the order of the $k-$step BDF method is k.

4.13 Prove that the $k-$step BDF methods are stable for $k = 1, 2, 3, 4$. (It is known that these methods are stable, if and only if $1 \leq k \leq 6$.)
[*Hint*: Let ρ_k be the characteristic polynomial of the $k-$step method. Then, as mentioned in section 4.1,

$$\rho_k(\zeta) = \sum_{i=1}^{k} \frac{1}{i} \zeta^{k-i} (\zeta - 1)^i.$$

Check that

$$\rho_3(\zeta) = \frac{1}{6}(\zeta - 1) r_2(\zeta), \quad \rho_4(\zeta) = \frac{1}{12}(\zeta - 1) r_3(\zeta),$$

with $r_2(\zeta) = 11\zeta^2 - 7\zeta + 2$ και $r_3(\zeta) = 25\zeta^3 - 23\zeta^2 + 13\zeta - 3$. Determine the roots of r_2. Obviously $r_3(1) \neq 0$. Utilize the Schur theory or the Routh–Hurwitz criterion to check that r_3 is a simple von Neumann polynomial.]

4.14 Is the three-step method described by the constants

$$\alpha_3 = 1,\ \alpha_2 = -\frac{11}{6},\ \alpha_1 = 1,\ \alpha_0 = -\frac{1}{6},\ \beta_3 = \frac{1}{12},\ \beta_2 = \frac{1}{6},\ \beta_1 = -\frac{1}{2},\ \beta_0 = \frac{7}{12}$$

stable? Why?

4.15 Is the three-step method of Exercise 4.14 consistent? Why?

4.16 Determine the stability region of the Adams–Moulton method

$$y^{n+3} - y^{n+2} = h(9f^{n+3} + 19f^{n+2} - 5f^{n+1} + f^n)/24.$$

(Show that the region is symmetric with respect to the real axis and that the stability interval is $(-3, 0)$, and sketch the stability region in the complex plane. What happens near the imaginary axis?)

4.17 Show that the three-step BDF method is A_0-stable but not A–stable.
[*Hint*: The method can not be A–stable, since its order is three, while A–stable multistep methods are of order at most two. For the A_0-stability, it suffices to show, utilizing the Schur theory, for instance, that the polynomials π,

$$\pi(\zeta) = \left(\frac{11}{6} - \alpha\right)\zeta^3 - 3\zeta^2 + \frac{3}{2}\zeta - \frac{1}{3},$$

with $\alpha < 0$, satisfy the root condition.]

4.18 Let $a = t^0 < t^1 < \cdots < t^N = b$ be a partition (not necessarily uniform) of the interval $[a, b]$. The two-step BDF method is then

$$\frac{h_{n+2} + h_{n+1}}{h_{n+1}} \frac{y^{n+2} - y^{n+1}}{h_{n+2}} - \frac{h_{n+2}}{h_{n+1}} \frac{y^{n+2} - y^n}{h_{n+2} + h_{n+1}} = f^{n+2},$$

with $h_n := t^n - t^{n-1}$. Prove that its order is two, with respect to $h := \max_n h_n$. (It is known that for $\frac{h_n}{h_{n-1}} \leq \gamma$, with a constant $\gamma < 1 + \sqrt{2}$, the method is stable.)
[*Hint*: Taylor expand around the point t^{n+2} to check that

$$\frac{h_{n+2} + h_{n+1}}{h_{n+1} h_{n+2}} \big[y(t^{n+2}) - y(t^{n+1}) \big] - \frac{h_{n+2}}{h_{n+1}\big(h_{n+2} + h_{n+1}\big)} \big[y(t^{n+2}) - y(t^n) \big]$$
$$= -\frac{1}{6} h_{n+2}\big(h_{n+2} + h_{n+1}\big) y^{(3)}(t^{n+2}) + O(h^3).$$

Obviously, $h_{n+2}\big(h_{n+2} + h_{n+1}\big) = O(h^2)$.]

4.19 Consider a $k-$step method, described by the constants $\alpha_k, \ldots, \alpha_0$ and $\beta_k, \ldots, \beta_0$. If the method is stable and consistent, show that

$$\beta_k + \cdots + \beta_0 \neq 0.$$

4.20 We consider an initial value problem and assume that f is continuous and satisfies condition (1.13). Prove that the approximations of the two-step BDF method, with given starting approximations y^0 and y^1, are well defined, for any step-size h.

4.21 Determine the values of the parameter α, for which the three-step method

$$y^{n+3} - (\alpha + 1)^2 y^{n+2} + \alpha(\alpha^2 + 2\alpha + 2) y^{n+1} - \alpha^2(\alpha + 1) y^n = h f(t^{n+3}, y^{n+3})$$

is stable.

4.22 Let $k \in \mathbb{N}$. Prove the followng:

a) For given constants $\beta_0, \ldots, \beta_k \in \mathbb{R}$, there exists exactly one choice of constants $\alpha_0, \ldots, \alpha_k \in \mathbb{R}$, such that the order of accuracy of the corresponding $k-$step method is at least k.

b) For given constants $\alpha_0, \ldots, \alpha_k \in \mathbb{R}$ such that $\alpha_0 + \cdots + \alpha_k = 0$, there exists exactly one choice of constants $\beta_0, \ldots, \beta_{k-1}$, such that the order of the corresponding *explicit* $k-$step method is at least k.

[*Hint*: For $p = k$, consider the relations (4.50) as linear systems with unknowns $\alpha_0, \ldots, \alpha_k$ in the first case, and $\beta_0, \ldots, \beta_{k-1}$ in the second case. Consider also the implicit and the explicit Euler method to convince yourself that the assumption that the method is explicit is indeed necessary in the second case. For details see Exercise 4.26.]

4.23 (Implicit–explicit multistep methods.) Consider two $k-$step methods, one implicit and one explicit, described by the constants $\alpha_0, \ldots, \alpha_k$, $\beta_0, \ldots, \beta_k$ and $\alpha_0, \ldots, \alpha_k$, $\gamma_0, \ldots, \gamma_{k-1}$, respectively. For simplicity, we assume that the order of both methods is p. Combining these two methods, we discretize the initial value problem of Exercise 2.27 by the method

$$(\star) \qquad \sum_{j=0}^{k} \alpha_j y^{n+j} = h \sum_{j=0}^{k} \beta_j f(t^{n+j}, y^{n+j}) + h \sum_{j=0}^{k-1} \gamma_j g(t^{n+j}, y^{n+j}),$$

$n = 0, \ldots, N-1$, with given starting approximations $y^0, \ldots, y^{k-1}$. If the implicit method has good stability properties, methods of this form exhibit advantages similar to the ones mentioned in Exercise 2.27. Prove that the order of accuracy of the method $(\star)$ is p.

[*Hint*: Utilize the o.d.e. $y' = f(t, y) + g(t, y)$ to check that

$$\begin{aligned} &\sum_{j=0}^{k} \alpha_j y(t^{n+j}) - h \sum_{j=0}^{k} \beta_j f\big(t^{n+j}, y(t^{n+j})\big) - h \sum_{j=0}^{k-1} \gamma_j g\big(t^{n+j}, y(t^{n+j})\big) \\ &= \sum_{j=0}^{k} \big[\alpha_j y(t^{n+j}) - h\beta_j y'(t^{n+j})\big] + h\Big[\sum_{j=0}^{k} \beta_j G(t^{n+j}) - \sum_{j=0}^{k-1} \gamma_j G(t^{n+j})\Big], \end{aligned}$$

with $G(t) := g\big(t, y(t)\big)$. The first term on the right-hand side is of order h^{p+1}, since the order of the implicit method is p. To check that the second term is of the same order, Taylor expand around the point t^n and use the relations (4.50) between the coefficients β_j, α_j and γ_j, α_j, that determine the order of the original methods.]

4.24 (Implicit–explicit BDF methods.) Let $k \in \mathbb{N}, \beta_0 = \cdots = \beta_{k-1} = 0, \beta_k = 1$, and α_j, $j = 0, \ldots, k$, be the coefficients of ζ^j of the polynomial α,

$$\alpha(\zeta) = \sum_{i=1}^{k} \frac{1}{i} \zeta^{k-i} (\zeta - 1)^i.$$

As mentioned in section 4.1, these constants describe the $k-$step BDF method. Consider now the polynomial γ, $\gamma(\zeta) = \zeta^k - (\zeta - 1)^k$, of degree $k - 1$, and denote by γ_j, $j = 0, \ldots, k-1$, the coefficients of ζ^j of this polynomial. Prove that the order of accuracy of the explicit $k-$step method described by the constants $\alpha_0, \ldots, \alpha_k$, $\gamma_0, \ldots, \gamma_{k-1}$ is k. (It can be shown that this is the only choice of constants $\gamma_0, \ldots, \gamma_{k-1}$, for which order of accuracy at least k is achieved; see Exercise 4.22.)

4.25 Assume that the order of accuracy of an implicit $k-$step method described by the constants $\alpha_0, \ldots, \alpha_k$ and $\beta_0, \ldots, \beta_k$ is at least k. Prove that the order of accuracy of an explicit $k-$step method described by the constants $\alpha_0, \ldots, \alpha_k$ and $\tilde{\beta}_0, \ldots, \tilde{\beta}_{k-1}$ is at least k, if and only if $\tilde{\beta}_i$ are the coefficients of ζ^i of the polynomial $\tilde{\sigma}(\zeta) = \sigma(\zeta) - \beta_k(\zeta - 1)^k$, with σ the polynomial with coefficients $\beta_0, \ldots, \beta_k$. See also Exercise 4.24.
[*Hint*: Check that relations (4.51) yield in the present case $\sigma(\mathrm{e}^x) - \tilde{\sigma}(\mathrm{e}^x) = O(x^k)$, and infer that $\sigma(y) - \tilde{\sigma}(y) = O((y-1)^k)$ for $y \to 1$; see also Exercise 4.22.]

4.26 Let $k \in \mathbb{N}$ and given constants $\alpha_0, \ldots, \alpha_k \in \mathbb{R}$ be such that $\alpha_0 + \cdots + \alpha_k = 1$. As we know, see Exercise 4.22, for exactly one choice of constants $\tilde{\beta}_0, \ldots, \tilde{\beta}_{k-1} \in \mathbb{R}$ has the corresponding explicit $k-$step method described by the constants $\alpha_0, \ldots, \alpha_k$ and $\tilde{\beta}_0, \ldots, \tilde{\beta}_{k-1}$ order of accuracy at least k. Prove that an implicit $k-$step method described by the constants $\alpha_0, \ldots, \alpha_k$ and $\beta_0, \ldots, \beta_k$ has order of accuracy k, if and only if $\beta_k \neq 0$ and $\beta_0, \ldots, \beta_{k-1}$ are the coefficients of ζ^i of the polynomial $\sigma(\zeta) = \beta_k(\zeta - 1)^k + \tilde{\sigma}(\zeta)$, with $\tilde{\sigma}$ the polynomial with coefficients $\tilde{\beta}_0, \ldots, \tilde{\beta}_{k-1}$.
[*Hint*: Use the previous Exercise.]

Bibliography

1. G. D. Akrivis, V. A. Dougalis: *Numerical Methods for Ordinary Differential Equations*. Crete University Press, Heraklion, 2006 (in Greek).
2. K. Atkinson, W. Han, D. E. Stewart: *Numerical Solution of Ordinary Differential Equations*. Wiley, New York, 2009.
3. G. Birkhoff, G.–C. Rota: *Ordinary Differential Equations*. 3rd ed., Wiley, New York, 1978.
4. J. Butcher: *Implicit Runge–Kutta processes*. Math. Comp. **18** (1964), pp. 50–64.
5. J. Butcher: *Numerical Methods for Ordinary Differential Equations*. 2nd ed., Wiley, 2008.
6. E. A. Coddington, N. Levinson: *Theory of Ordinary Differential Equations*. McGraw–Hill, New York, 1955.
7. M. Crouzeix: *Sur l'approximation des équations différentielles operationelles linéaires par des méthodes de Runge–Kutta*. Thèse d'état, Univ. Paris VI, 1975.
8. M. Crouzeix, A. L. Mignot: *Analyse Numérique des Équations Différentielles*. 2nd ed., Masson, Paris, 1989.
9. G. Dahlquist: *Convergence and stability in the numerical integration of ordinary differential equations*. Math. Scand. **4** (1956), pp. 33–53.
10. K. Dekker, J. G. Verwer: *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations*. North Holland, Amsterdam, 1984.
11. J. R. Dormand: *Numerical Methods for Differential Equations*: *a Computational Approach*. CRC Press, Boca Raton, New York, 1996.

12. C. W. Gear: *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice–Hall, Englewood Cliffs, N. J., 1971.
13. D. F. Griffiths, D. J. Higham: *Numerical Methods for Ordinary Differential Equations. Initial Value Problems*. Springer–Verlag, Berlin, 2010.
14. R. D. Grigorieff: *Numerik Gewöhnlicher Differentialgleichungen*. Bd. I, II, B. G. Teubner, Stuttgart, Bd. I: 1972, Bd. II: 1977.
15. E. Hairer, Ch. Lubich, G. Wanner: *Geometric Numerical Integration. Structure–Preserving Algorithms for Ordinary Differential Equations*. 2nd ed., Springer–Verlag, Berlin, 2006.
16. E. Hairer, S. P. Nørsett, G. Wanner: *Solving Ordinary Differential Equations I – Nonstiff Problems*. Springer–Verlag, Berlin, 2nd revised ed., 1993, corr. 2nd printing, 2000.
17. E. Hairer, G. Wanner: *Solving Ordinary Differential Equations II – Stiff and Differential–Algebraic Problems.* 2nd revised ed., Springer–Verlag, Berlin, 2010.
18. P. Henrici: *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York, 1962.
19. A. Iserles: *A First Course in the Numerical Analysis of Differential Equations*. 2nd ed., Cambridge University Press, Cambridge, 2008.
20. J. D. Lambert: *Numerical Methods for Ordinary Differential Systems*. Wiley, Chichester, 1991.
21. L. F. Shampine, M. G. Gordon: *Computer Solution of Ordinary Differential Equations*: *The Initial Value Problem*. W. H. Freeman, San Francisco, 1975.
22. V. Thomée: *Galerkin Finite Element Methods for Parabolic Problems*. 2nd ed., Springer–Verlag, Berlin, 2006.
23. A. Tveito, H. P. Langtangen, B. F. Nielsen, X. Cai: *Elements of Scientific Computing.* Springer–Verlag, Berlin, 2011.
24. W. Walter: *Ordinary Differential Equations.* Springer–Verlag, New York, 1998.

Subject Index

Name Index