Numerical Analysis II

Peter Philip∗

Lecture Notes

Originally created for the class of spring semester 2015 at LMU Munich,
revised for the classes of spring semester 2017 and spring semester 2018

July 27, 2018

Contents

1 Discrete Fourier Transform (DFT) 3
  1.1 Definition and First Properties . . . . . . . . . . . . . . . . . . . . . . . . 3
  1.2 Approximation of Periodic Functions . . . . . . . . . . . . . . . . . . . . 6
  1.3 Trigonometric Interpolation . . . . . . . . . . . . . . . . . . . . . . . . . 10
  1.4 Fast Fourier Transform (FFT) . . . . . . . . . . . . . . . . . . . . . . . . 14

2 Numerical Solution of ODE 20
  2.1 Setting and Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20
  2.2 Explicit Single-Step Methods . . . . . . . . . . . . . . . . . . . . . . . . 24
     2.2.1 Explicit Euler Method . . . . . . . . . . . . . . . . . . . . . . . . 24
     2.2.2 General Error Estimates . . . . . . . . . . . . . . . . . . . . . . . 25
     2.2.3 Classical Explicit Runge-Kutta Method . . . . . . . . . . . . . . . 30
  2.3 General Runge-Kutta Methods . . . . . . . . . . . . . . . . . . . . . . . 32
     2.3.1 Definition, Butcher Tableaus, First Properties . . . . . . . . . . . . 32
     2.3.2 Existence and Uniqueness . . . . . . . . . . . . . . . . . . . . . . . 37
     2.3.3 Autonomous ODE . . . . . . . . . . . . . . . . . . . . . . . . . . . 43
     2.3.4 Stability Functions . . . . . . . . . . . . . . . . . . . . . . . . . . 47

∗E-Mail: philip@math.lmu.de
## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.5 Higher Orders of Consistency and Convergence</td>
<td>51</td>
</tr>
<tr>
<td>2.4 Extrapolation</td>
<td>72</td>
</tr>
<tr>
<td>2.5 Stepsize Control</td>
<td>79</td>
</tr>
<tr>
<td>2.6 Stiff Equations</td>
<td>82</td>
</tr>
<tr>
<td>2.7 Collocation Methods</td>
<td>90</td>
</tr>
<tr>
<td><strong>A Linear Algebra</strong></td>
<td>98</td>
</tr>
<tr>
<td>A.1 Inner Products and Orthogonality</td>
<td>98</td>
</tr>
<tr>
<td>A.2 Unitary Matrices</td>
<td>99</td>
</tr>
<tr>
<td><strong>B Lipschitz Continuity</strong></td>
<td>101</td>
</tr>
<tr>
<td><strong>C Ordinary Differential Equations (ODE)</strong></td>
<td>104</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td>104</td>
</tr>
</tbody>
</table>
1 Discrete Fourier Transform (DFT)

1.1 Definition and First Properties

Definition 1.1. Let \( n \in \mathbb{N} = \{1, 2, 3, \ldots \} \). The discrete Fourier transform (DFT) is the map

\[
\mathcal{F} : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad (f_0, \ldots, f_{n-1}) \mapsto (d_0, \ldots, d_{n-1}),
\]

where

\[
d_k := \frac{1}{n} \sum_{j=0}^{n-1} f_j e^{-\frac{ijk2\pi}{n}},
\]

\( i \in \mathbb{C} \) denoting the imaginary unit.

Important applications of DFT include the approximation and interpolation of periodic functions (Sec. 1.2 and Sec. 1.3), practical applications include digital data transmission (cf. Rem. 1.9). First, we need to study some properties of DFT and its inverse map.

Remark 1.2. Introducing the vectors

\[
f := \begin{pmatrix} f_0 \\ \vdots \\ f_{n-1} \end{pmatrix} \in \mathbb{C}^n, \quad d := \begin{pmatrix} d_0 \\ \vdots \\ d_{n-1} \end{pmatrix} \in \mathbb{C}^n,
\]

we can write (1.1) in matrix form as

\[
\mathcal{F} : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad d := \mathcal{F}(f) = \frac{1}{n} A f,
\]

where

\[
A = \begin{pmatrix}
1 & 1 & 1 & \ldots & 1 \\
1 & \omega & \omega^2 & \ldots & \omega^{n-1} \\
1 & \omega^2 & \omega^4 & \ldots & \omega^{2(n-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \omega^{n-1} & \omega^{2(n-1)} & \ldots & \omega^{(n-1)^2}
\end{pmatrix},
\]

that means \( A = (a_{kj}) \in \mathcal{M}(n, \mathbb{C}) \) is a symmetric complex matrix with

\[
\forall k,j \in \{0, \ldots, n-1\} \quad a_{kj} = \omega^{kj} = e^{rac{ijk2\pi}{n}}, \quad \omega := e^{\frac{i2\pi}{n}}.
\]

In (1.2), \( \overline{A} \) denotes the complex conjugate matrix of \( A \) (cf. (A.5a) in the Appendix).

Proposition 1.3. Let \( A \in \mathcal{M}(n, \mathbb{C}) \), \( n \in \mathbb{N} \), be the matrix defined in (1.3a), and set \( B := \frac{1}{\sqrt{n}} \overline{A} \).

(a) The columns of \( B \) form an orthonormal basis (cf. Def. A.3(a)) with respect to the standard inner product on \( \mathbb{C}^n \).
The matrix $B$ is unitary, i.e. invertible with $B^{-1} = B^*$. 

DFT, as defined in (1.1), is invertible and its inverse map, called inverse DFT, is given by

$$F^{-1} : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad (d_0, \ldots, d_{n-1}) \mapsto (f_0, \ldots, f_{n-1}), \quad (1.4a)$$

where

$$f_k := \sum_{j=0}^{n-1} d_j e^{\frac{ijk2\pi}{n}}. \quad (1.4b)$$

If $d, f \in \mathbb{C}^n$ are such that $d = F(f)$, then

$$\sum_{j=0}^{n-1} |d_j|^2 = \frac{1}{n} \sum_{j=0}^{n-1} |f_j|^2, \quad (1.5a)$$

i.e.

$$\|F(f)\|_2 = \frac{1}{\sqrt{n}} \|f\|_2. \quad (1.5b)$$

Proof. (a): For each $k = 0, \ldots, n-1$, we let $b_k$ denote the $k$th column of $B$. Then, noting $|\omega| = 1$,

$$\langle b_k, b_k \rangle = \sum_{j=0}^{n-1} b_{kj} \bar{b}_{kj} = \frac{1}{n} \sum_{j=0}^{n-1} |\omega|^{2jk} = \frac{n}{n} = 1.$$ 

Moreover, for each $k, l \in \{0, \ldots, n-1\}$ and $k \neq l$, we note $\omega^{j-k} \neq 1$, but $(\omega^{j-k})^n = 1$, and we use the geometric sum to obtain

$$\langle b_k, b_l \rangle = \sum_{j=0}^{n-1} b_{kj} \bar{b}_{jl} = \frac{1}{n} \sum_{j=0}^{n-1} \omega^{-jk} \omega^{jl} = \frac{n}{n} \sum_{j=0}^{n-1} (\omega^{j-k})^j = \frac{1}{n} \cdot \frac{1 - (\omega^{j-k})^n}{1 - \omega^{j-k}} = 0,$$

proving (a).

(b) is immediate due to (a) and the equivalences (i) and (iii) of Th. A.8 in the Appendix. (c) follows from (b), as

$$\forall d \in \mathbb{C}^n \quad f := F^{-1}(d) = \left(\frac{1}{\sqrt{n}} B\right)^{-1} d = \sqrt{n}B^*d \quad (1.6)$$

and

$$\forall k, j \in \{0, \ldots, n-1\} \quad \sqrt{n} b_{kj}^* = \frac{\sqrt{n}}{n} a_{kj} = e^{\frac{i jk2\pi}{n}}.$$

(d): According to Th. A.8(viii) in the Appendix, $B$ is isometric with respect to $\| \cdot \|_2$. Thus,

$$\|F(f)\|_2 = \left\| \frac{1}{\sqrt{n}} Bf \right\|_2 = \frac{1}{\sqrt{n}} \|f\|_2,$$

proving (1.5b). Then (1.5a) also follows, as it is merely (1.5b) squared. ■
As it turns out, the inverse DFT of Prop. 1.3(c) has the useful property that there are several simple ways to express it in terms of (forward) DFT:

**Proposition 1.4.** Let \( n \in \mathbb{N} \) and let \( \mathcal{F} \) denote DFT as defined in (1.1).

(a) One has
\[
\forall \quad d = (d_0, \ldots, d_{n-1}) \in \mathbb{C}^n \quad \mathcal{F}^{-1}(d) = n \mathcal{F}(d_0, d_{n-1}, d_{n-2}, \ldots, d_1).
\]

(b) Applying complex conjugation, one has
\[
\forall \quad d \in \mathbb{C}^n \quad \mathcal{F}^{-1}(d) = \overline{n \mathcal{F}(d)}.
\]

(c) Let \( \sigma \) be the map that swaps the real and imaginary parts of a complex number, i.e.
\[
\sigma : \mathbb{C} \rightarrow \mathbb{C}, \quad \sigma(a + bi) := b + ai = i(a + bi).
\]

Using componentwise application, we can also consider \( \sigma \) to be defined on \( \mathbb{C}^n \):
\[
\sigma : \mathbb{C}^n \rightarrow \mathbb{C}^n, \quad \sigma(z_0, \ldots, z_{n-1}) := (\sigma(z_0), \ldots, \sigma(z_{n-1})) = i z.
\]

Then one has
\[
\forall \quad d \in \mathbb{C}^n \quad \mathcal{F}^{-1}(d) = n \sigma(\mathcal{F}(\sigma d)).
\]

*Proof.* (a): We let \( f = (f_0, \ldots, f_{n-1}) := n \mathcal{F}(d_0, d_{n-1}, d_{n-2}, \ldots, d_1) \) and compute for each \( k = 0, \ldots, n - 1 \):
\[
f_k \overset{(1.1b)}{=} \sum_{j=0}^{n-1} d_j e^{-\frac{ijk2\pi}{n}} = \sum_{j=1}^{n-1} d_j e^{-\frac{i(n-j)k2\pi}{n}} = \sum_{j=1}^{n-1} d_j e^{-\frac{jk2\pi}{n}} e^{-\frac{1}{n}k2\pi} = n \sum_{j=0}^{n-1} d_j e^{\frac{jk2\pi}{n}},
\]
which, according to (1.4), proves (a).

(b): Letting \( f := n \mathcal{F}(\overline{d}) \), we obtain for each \( k = 0, \ldots, n - 1 \):
\[
f_k \overset{(1.1b)}{=} \sum_{j=0}^{n-1} \overline{d_j} e^{-\frac{ijk2\pi}{n}} = \sum_{j=0}^{n-1} d_j e^{\frac{ijk2\pi}{n}}.
\]

Taking the complex conjugate of \( f_k \) and comparing with (1.4) proves (b).

(c): Using the \( \mathbb{C} \)-linearity of DFT and (b), we compute
\[
n \sigma(\mathcal{F}(\sigma d)) = n \sigma(\mathcal{F}(i \overline{d})) = n \overline{i \mathcal{F}(\overline{d})} = n \overline{\mathcal{F}(\overline{d})} \overset{(b)}{=} \overline{n \mathcal{F}(\overline{d})} = \overline{\mathcal{F}^{-1}(d)},
\]
which proves (c).\[\blacksquare\]
1 DISCRETE FOURIER TRANSFORM (DFT)

1.2 Approximation of Periodic Functions

As a first application of the discrete Fourier transform, we consider the approximation of periodic functions; how this is used in the field of data transmission is indicated in Rem. 1.9 below. The approximation is based on the Fourier series representation of (sufficiently regular) periodic functions, which we will proceed to, briefly, consider as a preparation. We will make use of the following well-known relations between the exponential function and the trigonometric functions sine and cosine (see [Phi16a, (8.46)]):

\[ e^{iz} = \cos z + i \sin z \quad \text{(Euler formula),} \]
\[ \forall z \in \mathbb{C} \quad \cos z = \frac{e^{iz} + e^{-iz}}{2}, \]
\[ \forall z \in \mathbb{C} \quad \sin z = \frac{e^{iz} - e^{-iz}}{2i}. \]

**Definition 1.5.** Let \( L \in \mathbb{R}^+ \) and let \( f : [0, L] \rightarrow \mathbb{R} \) be integrable.

(a) For each \( j \in \mathbb{Z} \), we define numbers
\[
\begin{align*}
a_j & := a_j(f) := \frac{1}{L} \int_0^L f(t) \cos \left( -\frac{j2\pi t}{L} \right) \, dt \in \mathbb{R}, \\
b_j & := b_j(f) := \frac{1}{L} \int_0^L f(t) \sin \left( -\frac{j2\pi t}{L} \right) \, dt \in \mathbb{R}, \\
c_j & := c_j(f) := \frac{1}{L} \int_0^L f(t) e^{-ij2\pi t/L} \, dt \in \mathbb{C}. \end{align*}
\]

We call the \( a_j, b_j \) the **real Fourier coefficients** of \( f \) and the \( c_j \) the **complex Fourier coefficients** of \( f \). As a consequence of (1.7a), one has
\[ \forall j \in \mathbb{Z} \quad c_j = a_j + i b_j. \]

(b) For each \( n \in \mathbb{N} \), we define the functions
\[ S_n := S_n(f) : [0, L] \rightarrow \mathbb{R}, \quad S_n(x) := \sum_{j=-n}^n c_j e^{ij2\pi x/L}. \]

and consider the series
\[ (S_n)_{n \in \mathbb{N}}, \text{ i.e. } \forall x \in [0, L] \quad (S_n(x))_{n \in \mathbb{N}} =: \sum_{j=-\infty}^{\infty} c_j e^{ij2\pi x/L}. \]

We call the \( S_n \) the **Fourier partial sums** of \( f \) and the series of (1.11) the **Fourier series** of \( f \).
Remark 1.6. (a) We verify that the Fourier partial sums $S_n$ of (1.10) are, indeed, real-valued. The key observation is an identity that we will prove in a form so that we can use it both here and, later, in the proof of Th. 1.13(a). To this end, we fix $n \in \mathbb{N}$ and assume, for each $j \in \{-n, \ldots, n\}$, we have numbers $a_j, b_j \in \mathbb{R}$ and $c_j \in \mathbb{C}$ (not necessarily given by (1.8)), satisfying

$$\forall j \in \{-n, \ldots, n\} \quad c_j = a_j + ib_j \quad \text{and} \quad \forall j \in \{-n+1, \ldots, n-1\} \setminus \{0\} \quad (a_j = a_{-j}, \quad b_j = -b_{-j}). \quad (1.12)$$

Then, for each $x \in \mathbb{R}$, we compute

$$\sum_{j=-n}^{n} c_j e^{ij2\pi x/L} = \sum_{j=-n}^{n} (a_j + ib_j) e^{ij2\pi x/L} \quad (1.12)$$

$$= (a_{-n} + i b_{-n}) e^{-in2\pi x/L} + a_0 + ib_0 + (a_n + i b_n) e^{in2\pi x/L} + \sum_{j=1}^{n-1} a_j \left( e^{ij2\pi x/L} + e^{-ij2\pi x/L} \right) + i \sum_{j=1}^{n-1} b_j \left( e^{ij2\pi x/L} - e^{-ij2\pi x/L} \right) \quad (1.7b), (1.7c)$$

$$= (a_{-n} + i b_{-n}) e^{-in2\pi x/L} + a_0 + ib_0 + (a_n + i b_n) e^{in2\pi x/L} + 2 \sum_{j=1}^{n-1} a_j \cos \left( \frac{j2\pi x}{L} \right) - 2 \sum_{j=1}^{n-1} b_j \sin \left( \frac{j2\pi x}{L} \right). \quad (1.13)$$

We now come back to the case, where $a_j, b_j, c_j$ are defined by (1.8). Then (1.12) holds due to (1.9) together with the symmetry of cosine (i.e. $\cos(z) = \cos(-z)$) and the antisymmetry of sine (i.e. $\sin(z) = -\sin(-z)$), where we have $a_j = a_{-j}$ and $b_j = -b_{-j}$ even for $j = n$. Thus, for each $x \in [0, L]$, we use (1.13) to obtain

$$S_n(x) = \sum_{j=-n}^{n} c_j e^{ij2\pi x/L} \quad (1.13), b_0 = 0 \quad \Rightarrow \quad a_0 + 2 \sum_{j=1}^{n} a_j \cos \left( \frac{j2\pi x}{L} \right) - 2 \sum_{j=1}^{n} b_j \sin \left( \frac{j2\pi x}{L} \right),$$

showing $S_n(x)$ to be real-valued.

(b) In general, the question, whether (and in which sense) the Fourier series of $f$, as defined in (1.11), converges to $f$ does not have an easy answer. We will provide some sufficient conditions in the following Th. 1.7. For convergence results under weaker assumptions, see, e.g., [Rud87, Sec. 4.26, Sec. 5.11] or [Heu08, Th. 136.2, Th. 137.1, Th. 137.2].

Theorem 1.7. Let $L \in \mathbb{R}^+$ and let $f : [0, L] \rightarrow \mathbb{R}$ be integrable and periodic in the sense that $f(0) = f(L)$. Moreover, assume that $f$ is continuously differentiable (i.e. $f \in C^1([0, L])$) with $f'(0) = f'(L)$. Then the following convergence results for the Fourier series $(S_n)_{n \in \mathbb{N}}$ of $f$, with $S_n$ according to (1.10), hold:
1 DISCRETE FOURIER TRANSFORM (DFT)

(a) The Fourier series of \( f \) converges uniformly to \( f \). In particular, one has the pointwise convergence
\[
\forall \ x \in [0, L] \quad f(x) = \sum_{j=-\infty}^{\infty} c_j e^{ij2\pi x/L} = \lim_{n \to \infty} S_n(x).
\] (1.14)

(b) The Fourier series of \( f \) converges to \( f \) in \( L^p[0, L] \) for each \( p \in [1, \infty] \).

Proof. (a): For the main work of the proof, we refer to [Wer11, Th. IV.2.9], where (a) is proved for the case \( L = 2\pi \). Here, we only verify that the case \( L = 2\pi \) then also implies the general case of \( L \in \mathbb{R}^+ \): Assume (a) holds for \( L = 2\pi \), let \( L \in \mathbb{R}^+ \) be arbitrary, and \( f \in C^1([0, L]) \) with \( f(0) = f(L) \) as well as \( f'(0) = f'(L) \). Define
\[
g : [0, 2\pi] \rightarrow \mathbb{R}, \quad g(t) := f \left( \frac{tL}{2\pi} \right).
\]
Then \( g \) is continuously differentiable with
\[
g' : [0, 2\pi] \rightarrow \mathbb{R}, \quad g'(t) := \frac{L}{2\pi} f' \left( \frac{tL}{2\pi} \right).
\]
Also \( g(0) = f(0) = f(L) = g(2\pi) \) and \( g'(0) = \frac{L}{2\pi} f'(0) = \frac{L}{2\pi} f'(L) = g'(2\pi) \), i.e. we know (a) to hold for \( g \). Next, we note \( f(t) = g \left( \frac{t 2\pi}{L} \right) \) for each \( t \in [0, L] \) and compute, for each \( j \in \mathbb{Z} \),
\[
c_j(f) = \frac{1}{L} \int_0^L f(t) e^{-ij2\pi t/L} \, dt = \frac{1}{L} \int_0^L g \left( \frac{t 2\pi}{L} \right) e^{-ij2\pi t/L} \, dt
\]
\[
= \frac{1}{2\pi} \int_0^{2\pi} g(s) e^{-ij s} \, ds = c_j(g),
\]
where we used the change of variables \( s := \frac{t 2\pi}{L} \). Thus,
\[
\lim_{n \to \infty} \left\| S_n(f) - f \right\|_{\infty} = \lim_{n \to \infty} \sup \left\{ \left| f(x) - \sum_{j=-n}^{n} c_j(f) e^{ij2\pi x/L} \right| : x \in [0, L] \right\}
\]
\[
= \lim_{n \to \infty} \sup \left\{ \left| g \left( \frac{x 2\pi}{L} \right) - \sum_{j=-n}^{n} c_j(f) e^{ij2\pi x/L} \right| : x \in [0, L] \right\}
\]
\[
= \lim_{n \to \infty} \sup \left\{ \left| g(t) - \sum_{j=-n}^{n} cj(g) e^{ij t} \right| : t \in [0, 2\pi] \right\}
\]
\[
= \lim_{n \to \infty} \left\| S_n(g) - g \right\|_{\infty} \quad \text{for} \quad g = 0,
\]
which completes the proof of (a).
(b): According to (a), we have convergence in $L^\infty[0, L]$ (note that the continuity of $f$ and the compactness of $[0, L]$ imply $f \in L^\infty[0, L]$). As a consequence of the Hölder inequality (cf. [Phi17a, Th. 2.42]), one then also obtains, for each $p \in [1, \infty[$,

$$\lim_{n \to \infty} \|S_n(f) - f\|_p = L^{\frac{1}{p}} \lim_{n \to \infty} \|S_n(f) - f\|_\infty = 0,$$

thereby establishing the case. □

For a periodic function $f$, applying the inverse discrete Fourier transform can yield a useful approximation of the values of $f$ at equidistant points:

**Example 1.8.** Let $L \in \mathbb{R}^+$ and let $f : [0, L] \to \mathbb{R}$ be integrable and such that $f(0) = f(L)$. Given $n \in \mathbb{N}$, we now define the equidistant points

$$\forall k \in \{0, \ldots, 2n+1\} \quad x_k := kh, \quad h := \frac{L}{2n+1}. \quad (1.15)$$

Plugging $x := x_k$ into (1.10) then yields the following approximations $f_k$ of $f(x_k)$,

$$f_k := S_n(f)(x_k) = \sum_{j=-n}^{n} c_j e^{\frac{ik\pi}{2n+1}} = \sum_{j=0}^{2n} c_{j-n} e^{\frac{i(k-n)\pi}{2n+1}}$$

$$= e^{-\frac{ink\pi}{2n+1}} \sum_{j=0}^{2n} c_{j-n} e^{\frac{ik\pi}{2n+1}}, \quad (1.16)$$

where we have omitted $f(x_{2n+1})$ in (1.16), since we know $f(x_{2n+1}) = f(L) = f(0) = f(x_0)$, due to the assumed periodicity of $f$. Comparing (1.16) with (1.4), we see that

$$\left( f_0, e^{\frac{in\pi}{2n+1}} f_1, \ldots, e^{\frac{i2n\pi}{2n+1}} f_{2n} \right) = \mathcal{F}^{-1}(c_{-n}, \ldots, c_n). \quad (1.17)$$

If Th. 1.7(a) holds (e.g. under the hypotheses of Th. 1.7), then, letting

$$F_n := (f(x_0), \ldots, f(x_{2n})), \quad G_n := (f_0, \ldots, f_{2n}),$$

we conclude

$$\lim_{n \to \infty} \|F_n - G_n\|_\infty = \lim_{n \to \infty} \sup \left\{ |f(x_k) - S_n(f)(x_k)| : k \in \{0, \ldots, 2n\} \right\} \overset{\text{Th. 1.7(a)}}{=} 0. \quad (1.18)$$

**Remark 1.9.** A physical application of the above-described approximation of periodic functions in connection with DFT is digital data transmission. Given an analog signal, represented by a function $f$, the signal is compressed into finitely many $c_j$ according to (1.16) and (1.8c) (physically, this can be accomplished using high-pass and low-pass filters). The digital signal consisting of the $c_j$ is then transmitted and, at the remote location, one wants to reconstruct the original $f$, at least approximately. According to (1.17) and (1.18), an approximation of $f$ can be obtained by applying inverse DFT to the $c_j$ (combined with a suitable interpolation method).
1.3 Trigonometric Interpolation

Another useful application of DFT is the interpolation of periodic functions by means of trigonometric functions.

Definition 1.10. Let \( L \in \mathbb{R}^+ \), \( n \in \mathbb{N} \), and \( d := (d_0, \ldots, d_{n-1}) \in \mathbb{C}^n \). Then the function

\[
p := p(d) : \mathbb{R} \rightarrow \mathbb{C}, \quad p(x) := \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi x}{L}}, \tag{1.19}
\]

is called a trigonometric polynomial (which is justified due to Euler’s formula). As it turns out, the following modification will be more useful for our purposes here: We let

\[
r := r(d) : \mathbb{R} \rightarrow \mathbb{C},
\]

\[
r(x) := e^{-\frac{ix\pi}{L}} p(d)(x) = e^{-\frac{x \pi}{L}} \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi x}{L}} = \sum_{k=0}^{n-1} d_k e^{\frac{ijk2\pi}{nL}}. \tag{1.20}
\]

Theorem 1.11. In the situation of Def. 1.10, let \( x_0, \ldots, x_{n-1} \) be defined by \( x_j := jL/n \) for each \( j \in \{0, \ldots, n-1\} \). Moreover, let \( z := (z_0, \ldots, z_{n-1}) \in \mathbb{C}^n \).

(a) The function \( r : \mathbb{R} \rightarrow \mathbb{C} \) of Def. 1.10 satisfies

\[
\forall j \in \{0, \ldots, n-1\} \quad r(x_j) = z_j \tag{1.21}
\]

if, and only if,

\[
\mathcal{F}\left((-1)^0z_0, \ldots, (-1)^{n-1}z_{n-1}\right) = d. \tag{1.22}
\]

(b) Let \( m \in \mathbb{N} \). If \( f : [0, L] \rightarrow \mathbb{C} \) is an \( m \) times continuously differentiable function such that \( f(0) = f(L) \), and the function \( r : \mathbb{R} \rightarrow \mathbb{C} \) of Def. 1.10 satisfies (1.21) with \( z_j := f(x_j) \), then there exists \( c_m \in \mathbb{R}^+ \) such that one has the following error estimate with respect to the \( L^2 \)-norm:

\[
\| r - f \|_2 \leq \frac{c_m \left( \| f \|_2 + \| f^{(m)} \|_2 \right)}{n^m} \tag{1.23}
\]

– recall that, for a square-integrable function \( g : [0, L] \rightarrow \mathbb{C} \), the \( L^2 \)-norm is defined by \( \|g\|_2 := \left( \int_0^L |g|^2 \right)^{\frac{1}{2}} \).

Proof. (a): According to (1.20) and using \( x_j = jL/n \), (1.21) is equivalent to

\[
\forall j \in \{0, \ldots, n-1\} \quad z_j = e^{-\frac{j \pi x_j}{L}} \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi}{nL}} = e^{-ij\pi} \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi}{nL}} \]

\[
= (-1)^j \sum_{k=0}^{n-1} d_k e^{\frac{ik2\pi}{nL}},
\]
which, according to (1.4), is equivalent to
\[ F^{-1}(d) = ((-1)^0 z_0, \ldots, (-1)^{n-1} z_{n-1}) . \]

Applying \( F \) to the previous equation completes the proof of (a).

(b) is stated as [Pla06, Th. 3.10], where it is referred to [SV01] for the proof. It actually turns out to be a special case of the more general result [HB09, Th. 52.6]. \( \blacksquare \)

**Example 1.12.** Consider
\[ f : [0, 1] \to \mathbb{R}, \quad f(x) := \begin{cases} x & \text{for } 0 \leq x \leq \frac{1}{2}, \\ 1-x & \text{for } \frac{1}{2} \leq x \leq 1. \end{cases} \]

It is an exercise to use different values of \( n \) (e.g. \( n = 4 \) and \( n = 16 \)) to compute and plot the corresponding \( r \) satisfying the conditions of Th. 1.11(a) with \( z_j = f(x_j) \), \( x_j = j/n \), for instance by using MATLAB.

When approximating real-valued periodic functions \( f : [0, L] \to \mathbb{R} \), it makes sense to use real-valued trigonometric functions. Indeed, we can choose real-valued trigonometric functions such that we can make use of the above results:

**Theorem 1.13.** Let \( L \in \mathbb{R}^+ \), let \( n \in \mathbb{N} \) be even, and \((a, b) := (a_0, \ldots, a_{\frac{n}{2}}, b_1, \ldots, b_{\frac{n}{2}-1}) \in \mathbb{R}^n\). Define the trigonometric function
\[ t := t(a, b) : \mathbb{R} \to \mathbb{R}, \quad t(x) := a_0 + 2 \sum_{k=1}^{\frac{n}{2}-1} \left( a_k \cos \left( \frac{k2\pi x}{L} \right) + b_k \sin \left( \frac{k2\pi x}{L} \right) \right) + a_{\frac{n}{2}} \cos \left( \frac{n\pi x}{L} \right) . \] (1.24)

Furthermore, let \( d = (d_0, \ldots, d_{n-1}) \in \mathbb{C}^n \), let the function \( r = r(d) \) be defined according to (1.20), and assume the relations
\[ \forall \quad \begin{array}{ll} d_0 = a_0, & d_{\frac{n}{2}} = a_{\frac{n}{2}}, \\ d_{\frac{n}{2}+k} = a_k - ib_k, & d_{\frac{n}{2}-k} = a_k + ib_k. \end{array} \] (1.25)

(a) The above conditions imply \( t = \text{Re } r \).

(b) If \( x_0, \ldots, x_{n-1} \in [0, L] \) are defined by \( x_j := jL/n \) for each \( j \in \{0, \ldots, n-1\} \), \( z := (z_0, \ldots, z_{n-1}) \in \mathbb{R}^n \), and \( r \) satisfies the conditions of Th. 1.11(a), then
\[ \forall \quad \begin{array}{l} t(x_j) = z_j \quad (1.26) \end{array} \]

and, moreover,
\[ \forall \quad \begin{array}{l} a_k = \frac{1}{n} \sum_{j=0}^{n-1} z_j \cos \left( \frac{jk2\pi}{n} \right), \quad b_k = \frac{1}{n} \sum_{j=0}^{n-1} z_j \sin \left( \frac{jk2\pi}{n} \right). \end{array} \] (1.27)
(c) Let $f : [0, L] \rightarrow \mathbb{R}$. If $f$ and $r$ satisfy the conditions of Th. 1.11(b), then there exists $c_m \in \mathbb{R}^+$ such that the error estimate (1.23) holds with $r$ replaced by $t$.

Proof. (a): To apply (1.13), we let

$$c_\frac{n}{2} := 0, \quad c_k := d_k + \frac{n}{2}. \quad (1.28a)$$

Then, for each $k \in \{1, \ldots, n/2 - 1\}$,

$$\text{Re}(c_k) = \text{Re}(d_{\frac{n}{2} + k}) \quad (1.25) \quad a_k = \text{Re}(d_{\frac{n}{2} - k}) = \text{Re}(c_{-k}) \quad (1.25)$$

as well as

$$\text{Im}(c_k) = \text{Im}(d_{\frac{n}{2} + k}) \quad (1.25) \quad -b_k = -\text{Im}(d_{\frac{n}{2} - k}) = -\text{Im}(c_{-k}), \quad (1.25)$$

i.e. (1.12) is satisfied and, hence, (1.13) applies. We also note

$$c_{-\frac{n}{2}} = d_0 \quad (1.25), \quad c_0 = d_{\frac{n}{2}} = a_0, \quad c_{\frac{n}{2}} = 0. \quad (1.25)$$

We use the above in (1.13) (with $n$ replaced by $\frac{n}{2}$ and $j$ replaced by $k$) to obtain, for each $x \in \mathbb{R}$,

$$r(x) = \sum_{k=0}^{n-1} d_k e^{\frac{(k-n/2)2\pi x}{L}} = \sum_{k=-\frac{n}{2}}^{n-1} d_{k+\frac{n}{2}} e^{\frac{ik2\pi x}{L}} = \sum_{k=-\frac{n}{2}}^{\frac{n}{2}} c_k e^{\frac{ik2\pi x}{L}}$$

$$(1.13) \quad = \frac{n}{2} e^{-in\pi x/L} + c_0 + c_{\frac{n}{2}} e^{in\pi x/L}$$

$$+ 2 \sum_{k=1}^{\frac{n}{2}-1} \text{Re}(c_k) \cos \left( \frac{k2\pi x}{L} \right) - 2 \sum_{k=1}^{\frac{n}{2}-1} \text{Im}(c_k) \sin \left( \frac{k2\pi x}{L} \right)$$

$$(1.28) \quad = a_0 + 2 \sum_{k=1}^{\frac{n}{2}-1} \left( a_k \cos \left( \frac{k2\pi x}{L} \right) + b_k \sin \left( \frac{k2\pi x}{L} \right) \right)$$

$$+ a_{\frac{n}{2}} \left( \cos \left( \frac{n\pi x}{L} \right) - i \sin \left( \frac{n\pi x}{L} \right) \right),$$

proving $t = \text{Re} r$.

(b): (1.26) is an immediate consequence of (a). Furthermore, as a consequence of (1.22) and (1.1b), we have, for each $k \in \{0, \ldots, n\}$,

$$d_k = \frac{1}{n} \sum_{j=0}^{n-1} z_j (-1)^j e^{-\frac{ij2\pi k}{n}}. \quad (1.29)$$
Thus, for each $k \in \{1, \ldots, \frac{n}{2} - 1\}$,

$$a_k \stackrel{(1.25)}{=} \frac{1}{2}(d_{\frac{n}{2}-k} + d_{\frac{n}{2}+k})$$

$$= \frac{1}{2n} \sum_{j=0}^{n-1} z_j (-1)^j \left( e^{-ijk\pi\frac{n}{n}} + e^{-ijk2\pi\frac{n}{n}} \right)$$

$$\stackrel{(1.29)}{=} \frac{1}{n} \sum_{j=0}^{n-1} z_j \cos \left( \frac{jk2\pi}{n} \right),$$

thereby establishing the case. Analogously, for each $k \in \{1, \ldots, \frac{n}{2} - 1\}$,

$$b_k \stackrel{(1.25)}{=} \frac{1}{2i}(d_{\frac{n}{2}-k} - d_{\frac{n}{2}+k})$$

$$= \frac{1}{2in} \sum_{j=0}^{n-1} z_j (-1)^j \left( e^{-ijk2\pi\frac{n}{n}} - e^{-ijk\pi\frac{n}{n}} \right)$$

$$\stackrel{(1.7c)}{=} \frac{1}{n} \sum_{j=0}^{n-1} z_j \sin \left( \frac{jk2\pi}{n} \right),$$

thereby completing the proof of (1.27).

(c): According to (a), we have $t = \text{Re} r$. Since $f$ and $f^{(m)}$ are real-valued, we obtain from Th. 1.11(b):

$$\|t - f\|_2^2 = \int_0^L |\text{Re} r - f|^2 \leq \int_0^L (|\text{Re} r - f|^2 + |\text{Im} r|^2) = \int_0^L |r - f|^2$$

$$\leq \left( c_m \left( \|f\|_2 + \|f^{(m)}\|_2 \right) \right)^2,$$

which establishes the case.

Remark 1.14. To make use of Th. 1.13 to interpolate a real-valued function $f$ on $[0, L]$, one would proceed as follows:

(i) Choose $n \in \mathbb{N}$ even and compute the equidistant $x_j := jL/n \in [0, L]$.

(ii) Compute the corresponding values $z_j := f(x_j)$.

(iii) Compute $d := F((-1)^0 z_0, \ldots, (-1)^{n-1} z_{n-1})$ according to (1.22).

(iv) Compute the $a_k$ and $b_k$ from (1.25), i.e. from $a_k = \frac{1}{2}(d_{\frac{n}{2}-k} + d_{\frac{n}{2}+k})$ and $b_k = \frac{1}{2i}(d_{\frac{n}{2}-k} - d_{\frac{n}{2}+k})$.

(v) Obtain the real-valued interpolating function $t$ from (1.24).
1 Discrete Fourier Transform (DFT)

1.4 Fast Fourier Transform (FFT)

If one computes DFT using matrix multiplication as described in Rem. 1.2, then one needs $O(n^2)$ complex multiplications to apply $F$ to a vector with $n$ components. However, if one is able to choose $n$ to be a power of 2, then one can apply a method called fast Fourier transform (FFT) to reduce the complexity to just $O(n \log_2 n)$ complex multiplications (cf. Rem. 1.23 below). We will study the FFT method in the present section.

Remark 1.15. (a) FFT can also be used to apply inverse DFT by making use of Prop. 1.4.

(b) There are variants of FFT that make use of the prime factorization of $n$ and which can be applied if $n$ is not a power of 2 (see [SK11, p. 161] and the references given there). However, these variants fail to be fast if $n$ is (a large) prime (or, more generally, if $n$ has few large prime factors). A different approach that is also fast for large prime numbers $n$ is to extend the vectors to be transformed to larger vectors of size $m$, where $m$ is a power of 2. One can simply do the extension by so-called zero padding, i.e. by adding components with value zero to the original vector. However, as $\omega$ in (1.3) depends on $n$, the details of this approach are somewhat tricky. In the literature, this is known as Bluestein’s algorithm and it involves convolutions and the so-called chirp $z$-transform (see [Blu68, RSR69]). We will not study the details of these approaches in this class, i.e. FFT as presented below applies only to the case where $n$ is a power of 2.

FFT is based on the following Th. 1.17, which allows to compute the DFT of a vector of length $2n$ from the DFT of two vectors of length $n$. In preparation for Th. 1.17, we introduce some notation.

Notation 1.16. If we want to emphasize the dimension on which a DFT operator operates, we write the dimension as a superscript of the operator, denoting DFT from $\mathbb{C}^n$ to $\mathbb{C}^n$, $n \in \mathbb{N}$, by $F^{(n)} : \mathbb{C}^n \rightarrow \mathbb{C}^n$. For each $k \in \{0, \ldots, n-1\}$, we let $F_k^{(n)} := \pi_k^{(n)} \circ F^{(n)} : \mathbb{C}^n \rightarrow \mathbb{C}$, where $\pi_k^{(n)} : \mathbb{C}^n \rightarrow \mathbb{C}$ is the projection $(z_0, \ldots, z_{n-1}) \mapsto z_k$, denote the $k$th coordinate function of $F^{(n)}$.

Theorem 1.17. Let $n \in \mathbb{N}$. Then, for each $k \in \{0, \ldots, n-1\}$, it holds that

$$\forall (f_0, \ldots, f_{2n-1}) \in \mathbb{C}^{2n} \quad F_k^{(n)}(f_0, f_1, \ldots, f_{n-1}) + e^{-i\pi n \frac{k}{n}} F_k^{(n)}(f_n, f_{n+1}, \ldots, f_{2n-1}) = F_k^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1})$$

(1.30a)

and

$$\forall (f_0, \ldots, f_{2n-1}) \in \mathbb{C}^{2n} \quad F_k^{(n)}(f_0, f_1, \ldots, f_{n-1}) - e^{-i\pi n \frac{k}{n}} F_k^{(n)}(f_n, f_{n+1}, \ldots, f_{2n-1}) = F_{n+k}^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1}).$$

(1.30b)
Proof. Given \((f_0, \ldots, f_{2n-1}) \in \mathbb{C}^{2n}\), we compute
\[
\mathcal{F}_k^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1})
\]
\[
= \frac{1}{2n} \left( \sum_{j=0}^{n-1} f_j e^{-\frac{i(2j+k)2\pi}{2n}} + \sum_{j=0}^{n-1} f_{n+j} e^{-\frac{i(2j+k+1)2\pi}{2n}} \right)
\]
proving (1.30a), and
\[
\mathcal{F}_k^{(2n)}(f_0, f_n, f_1, f_{n+1}, \ldots, f_{n-1}, f_{2n-1})
\]
\[
= \frac{1}{2n} \left( \sum_{j=0}^{n-1} f_j e^{-\frac{i(j+k)2\pi}{2n}} + e^{-\frac{ik\pi}{n}} \sum_{j=0}^{n-1} f_{n+j} e^{-\frac{i(j+k)2\pi}{n}} \right)
\]
proving (1.30b).

Example 1.18. (a) If \(n = 2^q, q \in \mathbb{N}\), then one can apply Th. 1.17 \(q\) times to reduce an application of \(\mathcal{F}^{(n)}\) to \(n\) applications of \(\mathcal{F}^{(1)}\). Moreover, we observe that \(\mathcal{F}^{(1)}\) is merely the identity on \(\mathbb{C}\): According to (1.1):
\[
\forall f_0 \in \mathbb{C} \quad \mathcal{F}^{(1)}(f_0) = \frac{1}{4} f_0 e^0 = f_0.
\]
(b) For \(n = 8 = 2^3\), one can build the result of an application of \(\mathcal{F}^{(n)}\) from trivial
one-dimensional transforms in 3 steps as illustrated in the scheme in (1.31) below.

\[
\begin{array}{cccccccc}
  f_0 & f_4 & f_2 & f_6 & f_1 & f_5 & f_3 & f_7 \\
  \mapsto & \mapsto & \mapsto & \mapsto & \mapsto & \mapsto & \mapsto & \mapsto \\
  \mathcal{F}^{(1)}(f_0) & \mathcal{F}^{(1)}(f_4) & \mathcal{F}^{(1)}(f_2) & \mathcal{F}^{(1)}(f_6) & \mathcal{F}^{(1)}(f_1) & \mathcal{F}^{(1)}(f_5) & \mathcal{F}^{(1)}(f_3) & \mathcal{F}^{(1)}(f_7) \\
  \mapsto & \mapsto & \mapsto & \mapsto & \mapsto & \mapsto & \mapsto & \mapsto \\
  \mathcal{F}^{(2)}(f_0, f_4) & \mathcal{F}^{(2)}(f_2, f_6) & \mathcal{F}^{(2)}(f_1, f_5) & \mathcal{F}^{(2)}(f_3, f_7) \\
  \mapsto & \mapsto & \mapsto & \mapsto \\
  \mathcal{F}^{(4)}(f_0, f_2, f_4, f_6) & \mathcal{F}^{(4)}(f_1, f_3, f_5, f_7) \\
  \mapsto & \mapsto \\
  \mathcal{F}^{(8)}(f_0, f_1, f_2, f_3, f_4, f_5, f_6, f_7) & \\
\end{array}
\] (1.31)

The scheme in (1.31) also illustrates that getting all the indices right needs some careful accounting. It turns out that this can be done quite elegantly and efficiently by a trick known as bit reversal. Before studying bit reversal and its application to FFT systematically, let us see how it works in the present example: In (1.31), we want to compute \( \mathcal{F}^{(8)}(f_0, \ldots, f_7) \) and need to figure out the order of indices we need in the first row of (1.31). Using the bit reversal algorithm we can do that directly, without having to compute (and store) the intermediate lines of the scheme: We start by writing the numbers 0, ..., 7 in their binary representation instead of in their usual decimal representation:

\[
\begin{align*}
  (0)_{10} &= (000)_2, & (1)_{10} &= (001)_2, & (2)_{10} &= (010)_2, & (3)_{10} &= (011)_2, \\
  (4)_{10} &= (100)_2, & (5)_{10} &= (101)_2, & (6)_{10} &= (110)_2, & (7)_{10} &= (111)_2. \\
\end{align*}
\] (1.32)

Then we reverse the bits of each number in (1.32) (and we also provide the resulting decimal representation):

\[
\begin{align*}
  (0)_{10} &= (000)_2, & (4)_{10} &= (100)_2, & (2)_{10} &= (010)_2, & (6)_{10} &= (110)_2, \\
  (1)_{10} &= (001)_2, & (5)_{10} &= (101)_2, & (3)_{10} &= (011)_2, & (7)_{10} &= (111)_2. \\
\end{align*}
\] (1.33)

We note that, as if by magic, we have obtained the correct order needed in the first row of (1.31). In the following, we will prove that this procedure works in general.

**Definition 1.19.** For each \( q \in \mathbb{N}_0 \), we denote \( \mathcal{M}_q := \mathbb{Z}_{2^q} = \{0, \ldots, 2^q - 1\} \). As every \( n \in \mathcal{M}_q \) has a (unique) binary representation \( n = \sum_{j=0}^{q-1} b_j 2^j, b_j \in \{0,1\} \), we can define bit reversal (in \( q \) dimensions) as the map

\[
\rho_q : \mathcal{M}_q \rightarrow \mathcal{M}_q, \quad \rho_q \left( \sum_{j=0}^{q-1} b_j 2^j \right) := \sum_{j=0}^{q-1} b_{q-1-j} 2^j. \] (1.34)

**Lemma 1.20.** Let \( q \in \mathbb{N}_0 \).

(a) The bit reversal map \( \rho_q : \mathcal{M}_q \rightarrow \mathcal{M}_q \) is bijective with \( \rho_q^{-1} = \rho_q \).
(b) One has
\[ \forall n \in M_q \quad \rho_q(n) = \rho_{q+1}(2n). \]

(c) One has
\[ \forall n \in M_q \quad 2^q + \rho_q(n) = \rho_{q+1}(2n + 1). \]

Proof. (a): Since, for each \( j \in \{0, \ldots, q-1\} \), one has \( q - 1 - (q - 1 - j) = j \), this is immediate from (1.34).

(b): If \( n \in M_q \), then \( 0 \leq n \leq 2^q - 1 \), i.e. \( 0 \leq 2n \leq 2^{q+1} - 2 < 2^{q+1} - 1 \) and \( 2n \in M_{q+1} \). Moreover, for \( n = \sum_{j=0}^{q-1} b_j 2^j \in M_q \), \( b_j \in \{0, 1\} \), we compute
\[
\rho_{q+1}(2n) = \rho_{q+1} \left( 2 \sum_{j=0}^{q-1} b_j 2^j \right) = \rho_{q+1} \left( 0 \cdot 2^0 + \sum_{j=1}^{q-1} b_{j-1} 2^j + b_{q-1} 2^q \right) = \sum_{j=0}^{q-1} b_{q-1-j} 2^j = \rho_q(n),
\]
proving (b).

(c): If \( n \in M_q \), then \( 0 \leq n \leq 2^q - 1 \), i.e. \( 0 \leq 2n + 1 \leq 2^{q+1} - 1 \) and \( 2n + 1 \in M_{q+1} \). Moreover, for \( n = \sum_{j=0}^{q-1} b_j 2^j \in M_q \), \( b_j \in \{0, 1\} \), we compute
\[
\rho_{q+1}(2n + 1) = \rho_{q+1} \left( 1 + 2 \sum_{j=0}^{q-1} b_j 2^j \right) = \rho_{q+1} \left( 1 \cdot 2^0 + \sum_{j=1}^{q-1} b_{j-1} 2^j + b_{q-1} 2^q \right) = \sum_{j=0}^{q-1} b_{q-1-j} 2^j + 1 \cdot 2^q = 2^q + \rho_q(n),
\]
proving (c). \( \square \)

Definition 1.21. We define the fast Fourier transform (FFT) algorithm for the computation of the \( n \)-dimensional discrete Fourier transform \( F^{(n)} \) in the case \( n = 2^q \), \( q \in \mathbb{N}_0 \), by the following recursion: We start with the \( n = 2^q \) complex numbers \( z_0, \ldots, z_{n-1} \in \mathbb{C} \) (we will see below (cf. Th. 1.22) that, to compute \( d := F^{(n)}(f) \), \( f \in \mathbb{C}^n \), one has to choose \( z_0 := f_{\rho_0(0)}, z_1 := f_{\rho_0(1)}, \ldots, z_{n-1} := f_{\rho_0(n-1)} \):\)
\[
d_{0,0} := z_0, \quad \ldots, \quad d_{0,2^q-1} := z_{2^q-1} = z_{n-1}. \tag{1.35a}
\]

Then, for each \( 1 \leq r \leq q \), one defines recursively the following \( 2^{q-r} \) vectors \( d^{(r,0)}, \ldots, d^{(r,2^{q-r}-1)} \in \mathbb{C}^{2^r} \) by
\[
\forall k \in \{0, \ldots, 2^{r-1}-1\} \quad \forall j \in \{0, \ldots, 2^{q-r}-1\}
\begin{align*}
d_k^{(r,j)} & := d_k^{(r-1,2j)} + e^{-\frac{ik\pi}{2^{r-1}}} d_k^{(r-1,2j+1)} \\
d_k^{(r,2^{r-1}+j)} & := d_k^{(r-1,2j)} - e^{-\frac{ik\pi}{2^{r-1}}} d_k^{(r-1,2j+1)} \tag{1.35b}
\end{align*}
\]
(note that the recursion ends after \( q \) steps with the single vector \( d^{(q,0)} \in \mathbb{C}^n \)).
Theorem 1.22. Let \( n = 2^q, q \in \mathbb{N}_0, \) and \( z = (z_0, \ldots, z_{n-1}) \in \mathbb{C}^n. \) If, for \( r \in \{0, \ldots, q\}, j \in \{0, \ldots, 2^{q-r} - 1\}, \) the \( d^{(r,j)} \) are defined by the FFT algorithm of Def. 1.21, then
\[
\forall r \in \{0, \ldots, q\}, \quad \forall j \in \{0, \ldots, 2^{q-r} - 1\}, \quad d^{(r,j)} = \mathcal{F}^{(2^r)}(z_{j2^r + \rho_r(0)}, z_{j2^r + \rho_r(1)}, \ldots, z_{j2^r + \rho_r(2^{q-r}-1)}). \tag{1.36}
\]
In particular,
\[
d^{(q,0)} = \mathcal{F}^{(n)}(z_{\rho_q(0)}, z_{\rho_q(1)}, \ldots, z_{\rho_q(2^q-1)}), \tag{1.37}
\]
i.e., to compute \( \mathcal{F}^{(n)}(f_0, \ldots, f_{n-1}) \) for a given \( f \in \mathbb{C}^n, \) one has to set
\[
z := (f_{\rho_q(0)}, \ldots, f_{\rho_q(n-1)}) \tag{1.38}
\]
to obtain
\[
d^{(q,0)} = \mathcal{F}^{(n)}(f_0, \ldots, f_{n-1}). \tag{1.39}
\]
Proof. The proof of (1.36) is conducted via induction with respect to \( r \in \{0, \ldots, q\}. \) For \( r = 0, \) (1.36) reads
\[
\forall j \in \{0, \ldots, 2^{q-r} - 1\}, \quad d^{(0,j)} = \mathcal{F}^{(1)}(z_j), \tag{1.40}
\]
which is correct due to (1.35a) and Example 1.18(a). If \( r \in \{1, \ldots, q\} \) and \( j \in \{0, \ldots, 2^{q-r} - 1\}, \) then, for each \( k \in \{0, \ldots, 2^{q-r} - 1\}, \) we calculate
\[
d^{(r,j)} = \frac{1}{2} \left( d^{(r-1,j)} - e^{-\frac{2\pi i}{2^{q-r}} d^{(r-1,j+1)}} \right)
\]
\[
\text{ind. hyp.} \quad \frac{1}{2} \mathcal{F}_k^{(2^{r-1})}(z_{j2^{r-1} + \rho_{r-1}(0)}, z_{j2^{r-1} + \rho_{r-1}(1)}, \ldots, z_{j2^{r-1} + \rho_{r-1}(2^{q-r}-1)})
\]
\[
+ \frac{1}{2} e^{-\frac{2\pi i}{2^{q-r}} \mathcal{F}_k^{(2^{r-1})}}(z_{(j+1)2^{r-1} + \rho_{r-1}(0)}, \ldots, z_{(j+1)2^{r-1} + \rho_{r-1}(2^{q-r}-1)})
\]
\[
= \frac{1}{2} \mathcal{F}_k^{(2^{r-1})}(z_{j2^r + \rho_r(0)}, z_{j2^r + \rho_r(1)}, \ldots, z_{j2^r + \rho_r(2^{q-r}-1)})
\]
\[
+ \frac{1}{2} e^{-\frac{2\pi i}{2^{q-r}} \mathcal{F}_k^{(2^{r-1})}}(z_{j2^r + 2^{r-1} + \rho_r(0)}, \ldots, z_{j2^r + 2^{r-1} + \rho_r(2^{q-r}-1)})
\]
\[
\tag{1.41}
\]
where the equality at (\*) holds due to (1.30a) and Lem. 1.20(b),(c), since, for each \( \alpha \in \{0, \ldots, 2^{q-r} - 1\}, \)
\[
j2^r + \rho_{r-1}(\alpha) = j2^r + \rho_r(2\alpha) \quad \text{and} \quad j2^r + 2^{r-1} + \rho_{r-1}(\alpha) = j2^r + \rho_r(2\alpha + 1). \tag{1.42}
\]
Similarly,
\[
d^{(r,j)} = \frac{1}{2} \left( d^{(r-1,j)} - e^{-\frac{2\pi i}{2^{q-r}} d^{(r-1,j+1)}} \right)
\]
\[
\text{ind. hyp.} \quad \frac{1}{2} \mathcal{F}_k^{(2^{r-1})}(z_{j2^r + \rho_r(0)}, z_{j2^r + \rho_r(1)}, \ldots, z_{j2^r + \rho_r(2^{q-r}-1)})
\]
\[
- \frac{1}{2} e^{-\frac{2\pi i}{2^{q-r}} \mathcal{F}_k^{(2^{r-1})}}(z_{j2^r + 2^{r-1} + \rho_r(0)}, \ldots, z_{j2^r + 2^{r-1} + \rho_r(2^{q-r}-1)})
\]
\[
\tag{1.30a),(1.42)
\]
\[
\mathcal{F}_k^{(2^{r-1})}(z_{j2^r + \rho_r(0)}, z_{j2^r + \rho_r(1)}, \ldots, z_{j2^r + \rho_r(2^{q-r}-1)}), \tag{1.43}
\]

1 DISCRETE FOURIER TRANSFORM (DFT)
completing the induction proof of (1.36).

Cleary, (1.36) turns into (1.37) for \( r = q, \ j = 0 \); and (1.39) follows from (1.37) and (1.38) as \( \rho_q \circ \rho_q = \text{Id} \). \[ \blacksquare \]

**Remark 1.23.** As mentioned at the beginning of the section, FFT reduces the number of complex multiplications needed to compute an \( n \)-dimensional DFT from \( \mathcal{O}(n^2) \) (when done by matrix multiplication) to \( \mathcal{O}(n \log_2 n) \) (an actually quite significant improvement – even for a moderate \( n \) like \( n = 8 = 2^3 \), one has \( n^2 = 64 \) versus \( n \log_2 n = 24 \), i.e. one gains almost a factor 3; the larger \( n \) gets the more essential it becomes to use FFT).

Let us check the \( \mathcal{O}(n \log_2 n) \) claim for FFT as computed by (1.35) for \( n = 2^q, \ q \in \mathbb{N} \), (where we need to estimate the number of multiplications involved in the executions of (1.35b)):

(i) In a preparatory step, one computes the \( q - 1 \) factors \( e^{-i\pi \frac{r}{2^r}}, \ldots, e^{-i\pi \frac{q-1}{2^{q-1}}} \) by starting with \( e^{-i\pi \frac{q-1}{2^{q-1}}} \) and using

\[
\forall r \in \{0, \ldots, q-2\} \quad e^{-i\pi \frac{r}{2^r}} = \left( e^{-i\pi \frac{r}{2^{r+1}}} \right)^2.
\]

Thus, the number of multiplications used in this step is

\[
N_1 := q - 2 < q.
\]

It is used that \( e^{-i\pi q} = -1 \) and \( e^{-i\pi \frac{q-1}{2^{q-1}}} \) is assumed to be given (it has to be precomputed once, which adds a constant number of multiplications, depending on the desired accuracy).

Then (1.35b) has to be executed \( q \) times, namely once for each \( r = 1, \ldots, q \). We compute the number of multiplications needed to obtain the \( d^{(r,j)} \):

(ii) Starting from the precomputed \( \theta_r := e^{-i\pi \frac{r}{2^r}} \), one uses successive multiplications with \( \theta_r \) to obtain the factors \( \theta_r^k = e^{-i\pi \frac{rk}{2^r}}, \ k = 1, \ldots, 2^r - 1 \). Thus, the number of multiplications for this step is

\[
N_{2,r} := 2^{r-1} - 2 < 2^{r-1}.
\]

(iii) The computation of each \( d_k^{(r,j)} \) needs two multiplications (one by the \( e \)-factor and one by \( \frac{1}{2^r} \)). As the range for \( k \) is \( \{0, \ldots, 2^r - 1\} \) and the range for \( j \) is \( \{0, \ldots, 2^{q-r} - 1\} \), the total number of multiplications for this step is

\[
N_{3,r} := 2 \cdot 2^r \cdot 2^{q-r} = 2^{q+1}.
\]

Finally, summing all multiplications from steps (i) – (iii), one obtains

\[
N = N_1 + \sum_{r=1}^{q} (N_{2,r} + N_{3,r}) < q + \sum_{r=1}^{q} (2^{r-1} + 2^{q+1}) = q + 2^q - 1 + q 2^{q+1} < \log_2 n + n + 2n \log_2 n,
\]

which is \( \mathcal{O}(n \log_2 n) \) as claimed.
2 Numerical Solution of Ordinary Differential Equations (ODE)

2.1 Setting and Motivation

We write $\mathbb{K}$ if a statement is meant to be valid for both $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$. Even if $\mathbb{K} = \mathbb{C}$, differentiability will always mean $\mathbb{R}$-differentiability, identifying $\mathbb{C}^n \cong \mathbb{R}^{2n}$.

**Definition 2.1.** Let $n \in \mathbb{N}$. Given $G \subseteq \mathbb{R} \times \mathbb{K}^n$ and $f : G \to \mathbb{K}^n$, we call

\[ y' = f(x, y) \quad (2.1) \]

an (explicit) ODE of first order. A solution to this ODE is a differentiable function $\phi : I \to \mathbb{K}^n$, defined on a nontrivial (bounded or unbounded, open or closed or half-open) interval $I \subseteq \mathbb{R}$, satisfying the two conditions

(i) $\{(x, \phi(x)) \in I \times \mathbb{K}^n : x \in I\} \subseteq G$,

(ii) $\phi'(x) = f(x, \phi(x))$ for each $x \in I$,

where condition (i) is necessary to even formulate condition (ii).

To single out a specific solution to an ODE, we need further requirements, e.g., so-called initial conditions:

**Definition 2.2.** Let $n \in \mathbb{N}$. An initial value problem for (2.1) consists of the ODE (2.1) plus the initial condition

\[ y(\xi) = \eta \quad (2.2) \]

with given $\xi \in \mathbb{R}$ and $\eta \in \mathbb{K}^n$. A solution $\phi$ to the initial value problem is a differentiable function $\phi$ as in Def. 2.1 that is a solution to the ODE and that also satisfies (2.2) (with $y$ replaced by $\phi$) – in particular, this requires $\xi \in I$.

It can also be of interest to consider higher-order ODE, i.e. ODE that involve higher derivatives of the unknown function $y$. However, as it turns out, one can write every higher-order ODE as an equivalent first-order ODE (see, e.g., [Phi16c, Sec. 3.1]) with a corresponding equivalence of initial value problems. Thus, we will restrict ourselves to first-order ODE in this class.

**Remark 2.3.** It is often useful to write an initial value problem as in Def. 2.2 as an equivalent integral equation: The following result is a simple consequence of the fundamental theorem of calculus: If $G \subseteq \mathbb{R} \times \mathbb{K}^n$, $n \in \mathbb{N}$, and $f : G \to \mathbb{K}^n$ is
continuous, then, for each \((\xi, \eta) \in G\), the initial value problem consisting of (2.1) and (2.2) is equivalent to the integral equation

\[
y(x) = \eta + \int_{\xi}^{x} f(t, y(t)) \, dt, \tag{2.3}
\]

in the sense that a differentiable function \(\phi : I \rightarrow \mathbb{K}^n\), with \(\xi \in I \subseteq \mathbb{R}\) being a nontrivial interval, and \(\phi\) satisfying Def. 2.1(i) is a solution to (2.1) and (2.2) if, and only if,

\[
\forall x \in I \quad \phi(x) = \eta + \int_{\xi}^{x} f(t, \phi(t)) \, dt,
\]

i.e. if, and only if, \(\phi\) is a solution to the integral equation (2.3).

Under suitable hypotheses, initial value problems have a unique (maximal) solution:

**Theorem 2.4.** If \(G \subseteq \mathbb{R} \times \mathbb{K}^n\) is open, \(n \in \mathbb{N}\), and \(f : G \rightarrow \mathbb{K}^n\) is continuous and locally Lipschitz with respect to \(y\) (cf. Def. B.1(b)), then, for each \((\xi, \eta) \in G\), the initial value problem consisting of (2.1) and (2.2) has a unique maximal solution (where a solution is called maximal provided it is defined on an open interval and cannot be extended to a solution on some larger open interval).

**Proof.** The theorem follows by combining [Phi16c, Th. 3.15] with [Phi16c, Th. 3.22]. ■

For most ODE, no explicit solution formulas are available, and numerical methods are required to approximate solutions. Stable numerical methods virtually always require that small changes of the input data result in small changes of the output data. The following result Th. 2.6 shows that initial value problems for ODE are not hopeless in this regard. To prove Th. 2.6, we will make use of the following version of Gronwall’s inequality:

**Proposition 2.5** (Gronwall’s Inequality). Let \(a, b \in \mathbb{R}, \ a < b\), \(\alpha, \phi \in C[a, b], \ \alpha \geq 0, \ C \in \mathbb{R}\). If \(\phi\) satisfies the integral inequality

\[
\forall x \in [a, b] \quad \phi(x) \leq C + \int_{a}^{x} \alpha(t) \phi(t) \, dt, \tag{2.4}
\]

then \(\phi\) can be estimated in the following way:

\[
\forall x \in [a, b] \quad \phi(x) \leq C \exp \left( \int_{a}^{x} \alpha(t) \, dt \right). \tag{2.5}
\]

**Proof.** Fix \(\epsilon > 0\) and define the auxiliary function

\[
g : [a, b] \rightarrow \mathbb{R}, \quad g(x) := (C + \epsilon) \exp \left( \int_{a}^{x} \alpha(t) \, dt \right).
\]
We differentiate $g$ to note it satisfies the ODE
\[ \forall \quad x \in [a,b] \quad g'(x) = (C + \epsilon) \alpha(x) \exp \left( \int_a^x \alpha(t) \, dt \right) = \alpha(x)g(x), \]
and, equivalently,
\[ \forall \quad x \in [a,b] \quad g(x) = C + \epsilon + \int_a^x \alpha(x)g(x). \]
We claim that
\[ \forall \quad x \in [a,b] \quad \phi(x) < g(x). \tag{2.6} \]
Due to (2.4) and $\epsilon > 0$, we know $\phi(a) < g(a)$. Seeking a contradiction, assume
\[ s := \text{sup} \left\{ x \in [a,b] : \phi(t) < g(t) \text{ for each } a \leq t < x \right\} < b. \]
Then the continuity of $\phi$ and $g$ implies $s > a$ and $\phi(s) = g(s)$. On the other hand,
\[ \phi(s) \leq C + \int_a^s \alpha(t)\phi(t) \, dt < C + \epsilon + \int_a^s \alpha(t)g(t) \, dt = g(s) \]
in contradiction to $\phi(s) = g(s)$. Thus, the assumption $s < b$ was wrong, which means (2.6) is true. Since $\epsilon > 0$ was arbitrary, (2.6) implies (2.5). \[ \blacksquare \]

**Theorem 2.6** (Continuity in Initial Conditions). If $G \subseteq \mathbb{R} \times \mathbb{K}^n$ is open, $n \in \mathbb{N}$, and $f : G \longrightarrow \mathbb{K}^n$ is continuous and globally $L$-Lipschitz with respect to $y$, i.e.
\[ \exists \quad L \geq 0 \quad \forall \quad (x,y),(x,y') \in G \quad \| f(x,y) - f(x,y') \| \leq L \| y - y' \|, \tag{2.7} \]
then the solutions to (2.1) depend continuously on the initial condition: Let $\phi, \psi : I \longrightarrow \mathbb{K}^n$ both be solutions to (2.1) defined on the same interval $I \subseteq \mathbb{R}$ with $\xi \in I$, then,
\[ \forall \quad x \in I, \quad x \geq \xi \quad \| \phi(x) - \psi(x) \| \leq e^{L(x-\xi)} \| \phi(\xi) - \psi(\xi) \|. \tag{2.8} \]

**Proof.** As both $\phi$ and $\psi$ satisfy an integral equation corresponding to (2.3), we obtain, for each $x \in I$ with $x \geq \xi$:
\[ \| \phi(x) - \psi(x) \| = \left\| \phi(\xi) - \psi(\xi) + \int_{\xi}^x \left( f(t, \phi(t)) - f(t, \psi(t)) \right) \, dt \right\| \]
\[ \leq \| \phi(\xi) - \psi(\xi) \| + L \int_{\xi}^x \| \phi(t) - \psi(t) \| \, dt. \tag{2.7} \]
We now note that we can apply Prop. 2.5 with $a := \xi$, $C := \| \phi(\xi) - \psi(\xi) \|$, $\alpha \equiv L$, and $\phi$ replaced by $\| \phi - \psi \|$ to obtain (2.8). \[ \blacksquare \]
The general idea of most numerical approximation methods for initial value problems for ODE is to start at $x_0 := \xi$ and to proceed by discrete steps $h_0, h_1, \ldots$ to $x_1 := x_0 + h_0 > x_0, x_2 := x_1 + h_1 > x_1, \ldots$, while, simultaneously, starting with $y_0 := \eta$ and proceeding to approximations $y_1$ of $y(x_1), y_2$ of $y(x_2), \ldots$. This gives rise to the following definition:

**Definition 2.7.** Let $n \in \mathbb{N}, G \subseteq \mathbb{R} \times \mathbb{K}^n, f : G \rightarrow \mathbb{K}^n, (\xi, \eta) \in G$. A (discrete) numerical approximation for the initial value problem

\begin{align}
  y' &= f(x, y), \\
  y(\xi) &= \eta,
\end{align}

is a (finite or infinite) sequence $((x_k, y_k))_{k \in \{0,1,\ldots\}}$ in the domain $G$ of $f$ such that $(x_0, y_0) = (\xi, \eta)$ and $x_0 < x_1 < \ldots$. Thus, each such numerical approximation comes with a sequence of stepsizes $(h_k)_{k \in \{0,1,\ldots\}}$, where

$$
\forall \ k \in \{0,1,\ldots\} \quad h_k := x_{k+1} - x_k. \quad (2.10)
$$

Let $m \in \mathbb{N}$. We speak of an *m-step method* for the numerical solution of (2.9) if, and only if, the sequence $(y_k)_{k \in \{0,1,\ldots\}}$ is given by a recursion of the form

$$
\forall \ k \in \{0,1,\ldots\} \quad y_{k+m} = \sum_{j=0}^{m-1} \alpha_j y_{k+j} + h_{k+m-1} \varphi(x_k, \ldots, x_{k+m-1}, x_{k+m}, y_k, y_{k+1}, \ldots, y_{k+m-1}, y_{k+m}, h_{k+m-1}),
$$

with $\alpha_0, \ldots, \alpha_{m-1} \in \mathbb{R}$ and a defining function (sometimes also called increment function)

$$
\varphi : D_{\varphi} \rightarrow \mathbb{K}^n, \quad \text{where} \quad D_{\varphi} \subseteq \mathbb{R}^{m+1} \times (\mathbb{K}^n)^{m+1} \times \mathbb{R}^+
$$

is a suitable domain. If $m = 1$, then one speaks of a *single-step* or *one-step* method, for $m > 1$ of a *multistep* method. The method is called *explicit* if the increment function $\varphi$ does not, actually, depend on $y_{k+m}$ and *implicit* otherwise.

**Remark 2.8.** (a) Note that, in (2.11), the sum involving the $\alpha_j$ and the factor $h_{k+m}$ in front of $\varphi$ could have been incorporated into the increment function. However, the form given in (2.11) seems to be the one commonly used in the literature; indeed, using this form is convenient for a large number of commonly considered methods (we will see examples below).

(b) We note that the distinction between explicit and implicit methods as defined in Def. 2.7 is somewhat arbitrary and the property of being explicit or implicit is more a property of the *presentation* of the method rather than of the method itself: Clearly, every explicit method can be written as an implicit method (e.g. by adding $y_{k+m}$ on both sides of (2.11), followed by a division by 2). On the other hand, if the method is implicit and it can not be written, equivalently, as an explicit method, then (2.11) can not be uniquely solved for $y_{k+m}$ (i.e. either there is no solution or more than one solution) and the method is, actually, not well-defined. We will revisit the explicit/implicit issue in Sec. 2.3, when discussing so-called
(general) Runge-Kutta (RK) methods. For RK methods, one also defines explicit and implicit versions, however, the meaning is not the same as in Def. 2.7 (for an RK method, there is a clearer (parametric) distinction between explicit and implicit – but every RK method, at least when it is well-defined and written in its standard from, is explicit in the sense of Def. 2.7, cf. Rem. 2.27(b)). Still, the distinction between explicit and implicit methods made in Def. 2.7 is extremely common in the literature, and not without usefulness. For example, explicit methods tend to be computationally much simpler to execute, as obtaining \( y_{k+m} \) basically means one evaluation of \( \varphi \). For implicit methods, \( y_{k+m} \) is obtained as a solution to (2.11), which, in general, constitutes a coupled system of \( n \) nonlinear equations, and might be very difficult to solve. The system (2.11) is typically again solved by a suitable numerical method, e.g. Newton’s method. In particular, obtaining \( y_{k+m} \) usually requires several evaluations of \( \varphi \). On the other hand, depending on the ODE to be solved, the additional computational cost in each step of such an implicit method might be more than compensated by better convergence properties. The type of ODE, where this occurs is known as stiff ODE (cf. Sec. 2.6).

(c) A priori, the methods defined in Def. 2.7 yield discrete approximations \( y_0, y_1, \ldots \) to \( y(x_0), y(x_1), \ldots \), but they do not yield an approximating function \( u : I \rightarrow \mathbb{K}^n \) of \( y : I \rightarrow \mathbb{K}^n \) on a suitable interval with \( x_0 \in I \subseteq \mathbb{R} \). To pass from the \( y_k \) to a function \( u \) requires interpolation. Different interpolation methods are possible, where interpolation by splines is often used, piecewise linear (i.e. affine) splines being the most simple possible choice.

2.2 Explicit Single-Step Methods

The most simple methods considered in Def. 2.7 are explicit methods for \( m = 1 \), i.e. explicit single-step methods, which are the subject of the present section.

2.2.1 Explicit Euler Method

The explicit Euler method is the most simple of the explicit single-step methods. It is not really sufficiently accurate for practical use, but it is still instructive to study to familiarize ourselves with some of the relevant ideas and issues.

Definition 2.9. In the situation of Def. 2.7, we call the numerical method explicit Euler method if, and only if, \( m = 1, \alpha_0 = 1 \), and the defining function is

\[
\varphi : G \times \mathbb{R}^+ \rightarrow \mathbb{K}^n, \quad \varphi(x, y, h) := f(x, y),
\]

i.e., given \( \xi = x_0 < x_1 < \ldots \) and \( h_k \) according to (2.10), the explicit Euler method consists of the recursion

\[
\begin{align*}
\forall k \in \{0, 1, \ldots \} & \quad y_0 = \eta, \\
y_{k+1} = y_k + h_k f(x_k, y_k).
\end{align*}
\]
Remark 2.10. (a) In general, there is no guarantee that $y_{k+1}$ is well-defined by (2.12). More precisely, $y_{k+1}$ is well-defined by (2.12) if, and only if, $(x_k, y_k) \in G$. Thus, (2.12) is supposed to mean that the recursion continues as long as $(x_k, y_k) \in G$, and it terminates with $y_k$ if $(x_k, y_k) \notin G$.

(b) For $n = 1$ and $\mathbb{K} = \mathbb{R}$, the explicit Euler method can easily be visualized: If $y$ is a solution to (2.9), then, at $x_k$, its slope is $y'(x_k) = f(x_k, y(x_k))$. Thus, in step $k + 1$, the Euler method approximates $y$ by the line through $(x_k, y_k)$ with slope $f(x_k, y_k)$ (using that $y_k$ is supposed to approximate $y(x_k)$). This line is described by

$$l(x) = y_k + (x - x_k) f(x_k, y_k),$$

i.e. $y_{k+1} = l(x_{k+1}) = y_k + h_k f(x_k, y_k)$ as in (2.12).

2.2.2 General Error Estimates

We will now introduce notions and results that help to gauge the accuracy (and the expected error) of explicit single-step methods.

Notation 2.11. Given a real interval $[a, b)$, $a < b$, the finite sequence $\Delta = (x_0, \ldots, x_N)$ $\in \mathbb{R}^{N+1}$, $N \in \mathbb{N}$, is called a partition of $[a, b)$ if, and only if, $a = x_0 < x_1 < \cdots < x_N = b$. Given such a partition $\Delta$ of $[a, b]$, define the numbers

$$h_{\text{max}}(\Delta) := \max \{h_k : k \in \{0, \ldots, N - 1\}\}, \quad h_k := x_{k+1} - x_k,$$

$$h_{\text{min}}(\Delta) := \min \{h_k : k \in \{0, \ldots, N - 1\}\},$$

(2.13a) (2.13b)

The number $h_{\text{max}}(\Delta)$ is called the mesh size of $\Delta$. Moreover, let $\Pi([a, b])$ denote the set of all partitions of $[a, b]$.

Definition 2.12. Consider the initial value problem (2.9) and assume $f$ is such that (2.9) has a unique solution $\phi : I \to \mathbb{K}^n$ on some open interval $I \subseteq \mathbb{R}$. Let $b > \xi$ be such that $[\xi, b] \subseteq I$. Moreover, consider an explicit single-step method according to Def. 2.7, with $a_0 = 1$ and given by a defining function

$$\varphi : \mathcal{D}_\varphi \to \mathbb{K}^n, \quad \mathcal{D}_\varphi \subseteq \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+.$$

Then, for each partition $\Delta = (x_0, \ldots, x_N)$ of $[\xi, b]$, $\varphi$ defines a recursion

$$\forall k \in \{0, 1, \ldots\} \quad y_{k+1} = y_k + h_k \varphi(x_k, y_k, h_k), \quad h_k := x_{k+1} - x_k.$$ (2.14)

(a) We call the method well-defined if, and only if, there exists $h(\varphi) \in \mathbb{R}^+$ such that, for each partition $\Delta = (x_0, \ldots, x_N)$ of $[\xi, b]$ with $h_{\text{max}}(\Delta) < h(\varphi)$, the recursion (2.14) provides all approximations $y_0, \ldots, y_N$ (cf. Rem. 2.10), i.e. if, and only if,

$$\forall \Delta \in \Pi([\xi, b]), \quad h_{\text{max}}(\Delta) < h(\varphi) \quad \forall k \in \{0, 1, \ldots, N - 1\} \quad (x_k, y_k, h_k) \in \mathcal{D}_\varphi.$$ (2.15)
(b) The method is said to have order of convergence \( p \in \mathbb{N} \) if, and only if, it is well-defined and satisfies
\[
\exists C \geq 0 \quad \forall \Delta \in \Pi([\xi, b]), \quad h_{\text{max}}(\Delta) < h(\varphi) \quad \max \left\{ \| y_k - \phi(x_k) \| : k \in \{0, \ldots, N\} \right\} \leq C h_{\text{max}}^p(\Delta), \quad (2.16)
\]
where \( h(\varphi) \in \mathbb{R}^+ \) is as in (a) and the quantity on the left-hand side of the inequality in (2.16) is known as the method’s global truncation error (it depends on both \( \varphi \) and \( \Delta \)).

(c) For each \( x \in [\xi, b] \) and each \( h \in ]0, b - x] \) such that \( (x, \phi(x), h) \in \mathcal{D}_\varphi \), we call
\[
\lambda(x, h) := \phi(x) + h \varphi(x, \phi(x), h) - \phi(x + h)
\]
the method’s local truncation error at point \( (x + h, \phi(x + h)) \) with respect to the stepsize \( h \). We call \( \lambda \) well-defined if, and only if, there exists \( h_\lambda(\varphi) \in \mathbb{R}^+ \) such that \( \lambda(x, h) \) is defined for each \( h \in I_x := ]0, \min\{h_\lambda(\varphi), b - x\} \). If \( \lambda \) is well-defined, then the method is said to be consistent if, and only if,
\[
\forall x \in [\xi, b] \quad \lim_{h \downarrow 0} \frac{\lambda(x, h)}{h} = 0; \quad (2.18)
\]
and consistent of order \( p \in \mathbb{N} \) if, and only if,
\[
\exists C \geq 0 \quad \forall x \in [\xi, b] \quad \forall h \in I_x \quad \| \lambda(x, h) \| \leq C h^{p+1}. \quad (2.19)
\]

**Lemma 2.13.** Consider the situation of Def. 2.12.

(a) Assume the local truncation error \( \lambda \) to be well-defined. Then the method given by \( \varphi \) is consistent if, and only if,
\[
\forall x \in [\xi, b] \quad \lim_{h \downarrow 0} \varphi(x, \phi(x), h) = f(x, \phi(x)). \quad (2.20)
\]

(b) If the method given by \( \varphi \) is consistent of order \( p \in \mathbb{N} \), then it is consistent.

**Proof.** (a): As a consequence of (2.17),
\[
\lim_{h \downarrow 0} \frac{\lambda(x, h)}{h} \quad \text{exists} \quad \iff \quad \lim_{h \downarrow 0} \varphi(x, \phi(x), h) \quad \text{exists}
\]
Moreover, if the limits exist, then
\[
\lim_{h \downarrow 0} \frac{\lambda(x, h)}{h} = -\phi'(x) + \lim_{h \downarrow 0} \varphi(x, \phi(x), h) = -f(x, \phi(x)) + \lim_{h \downarrow 0} \varphi(x, \phi(x), h),
\]
implying the equivalence between (2.18) and (2.20).

(b): If (2.19) holds, then
\[
\forall x \in [\xi, b] \quad \lim_{h \downarrow 0} \frac{\| \lambda(x, h) \|}{h} = \lim_{h \downarrow 0} (Ch^p) = 0,
\]
proving the method to be consistent. ■
We will see in Cor. 2.17 below that, if \( \varphi \) is sufficiently regular, then consistence of order \( p \) implies order of convergence \( p \) as well. In preparation, we prove the following lemma:

**Lemma 2.14.** Let \( N \in \mathbb{N} \) and consider the finite sequence \((a_0, \ldots, a_N)\) in \( \mathbb{R}_0^+ \). If there are numbers \( \beta, h_0, \ldots, h_{N-1} \in \mathbb{R}_0^+ \) and \( L \in \mathbb{R}^+ \) such that

\[
\forall k \in \{0, \ldots, N-1\} \quad a_{k+1} \leq (1 + h_k L) a_k + h_k \beta, \tag{2.21a}
\]

then

\[
\forall k \in \{0, \ldots, N\} \quad a_k \leq e^{Lx_k - \frac{1}{L}} \beta + a_0 e^{Lx_k}, \quad \text{where} \quad x_k := \sum_{j=0}^{k-1} h_j. \tag{2.21b}
\]

**Proof.** We prove (2.21b) via induction on \( k \): For \( k = 0 \), we have the true statement \( a_0 \leq a_0 \). Thus, let \( k \in \{0, \ldots, N-1\} \). Then one estimates

\[
\begin{align*}
    a_{k+1} &\leq (1 + h_k L) a_k + h_k \beta \quad \text{ind.hyp.} \leq e^{h_k L} (1 + h_k L) \left( e^{Lx_k} - \frac{1}{L} \beta + a_0 e^{Lx_k} \right) + h_k \beta \\
    &\leq \left( e^{L(x_k+h_k)} - \frac{1}{L} - 1 + h_k L \right) \beta + a_0 e^{L(x_k+h_k)} \\
    &= \frac{e^{Lx_{k+1}} - 1}{L} \beta + a_0 e^{Lx_{k+1}},
\end{align*}
\]

completing the proof. \( \blacksquare \)

**Definition 2.15.** The function \( \varphi : D_\varphi \to K^n \), \( n \in \mathbb{N} \), with \( D_\varphi \subseteq \mathbb{R} \times K^n \times \mathbb{R}^+ \) is called (globally) \( L \)-Lipschitz with respect to \( y \) if, and only if,

\[
\exists \ L \geq 0 \quad \forall (x,y,h),(x,\bar{y},h) \in D_\varphi \quad \|\varphi(x,y,h) - \varphi(x,\bar{y},h)\| \leq L \|y - \bar{y}\|. \tag{2.22}
\]

The following Th. 2.16 does not only account for the truncation error, but also for a possible error in the initial value as well as possible rounding errors in each step.

**Theorem 2.16.** Consider the setting of Def. 2.12. However, instead of the recursion (2.14), we consider the modification

\[
\begin{align*}
    v_0 &= \eta + e_0, \\
    \forall k \in \{0,1,\ldots\} \quad v_{k+1} &= v_k + h_k \varphi(x_k, v_k, h_k) + r_k, \quad h_k := x_{k+1} - x_k, \tag{2.23}
\end{align*}
\]

where \( e_0 \in K^n \) represents a possible error in the initial value and the \( r_k \in K^n \) represent possible rounding errors in the \( k \)th step. We assume

\[
\forall k \in \{0,1,\ldots\} \quad \|r_k\| \leq \delta \in \mathbb{R}_0^+. \tag{2.24}
\]
If the recursion (2.23) is well-defined in the sense of Def. 2.12(a), with \( y_0, \ldots, y_N \) replaced by \( v_0, \ldots, v_N \), the underlying explicit single-step method is consistent of order \( p \in \mathbb{N} \), and \( \varphi \) is globally \( L \)-Lipschitz with respect to \( y \), \( L > 0 \), then

\[
\forall \Delta \in \Pi([\xi, \delta]), h_{\text{max}}(\Delta) < \min\{h(\varphi), h(\lambda(\varphi))\} \leq \frac{e^{L(b-\xi)} - 1}{L} \left( \frac{\delta}{h_{\min}(\Delta)} \right) + \|e_0\| e^{L(b-\xi)},
\]

(2.25)

\( h(\varphi) \) and \( h(\lambda(\varphi)) \) being as in Def. 2.12(a), (c), respectively, and \( C \) being the constant given by (2.19).

**Proof.** To simplify notation, we write \( h_{\text{max}} := h_{\text{max}}(\Delta) \) and \( h_{\text{min}} := h_{\min}(\Delta) \). Assume \( h_{\text{max}} < \min\{h(\varphi), h(\lambda(\varphi))\} \). Introducing, for \( k \in \{0, \ldots, N\} \), the abbreviations

\[
\phi_k := \phi(x_k), \quad e_k := v_k - \phi_k,
\]

and, for \( k \in \{0, \ldots, N-1\} \),

\[
\lambda_k := \lambda(x_k, h_k) = \phi_k + h_k \varphi(x_k, \phi_k, h_k) - \phi_{k+1},
\]

(2.26)

the idea is to apply Lem. 2.14 with \( a_k := \|e_k\| \). Using (2.26) together with the recursion (2.23) for \( v_k \), we obtain, for each \( k \in \{0, \ldots, N-1\} \),

\[
e_{k+1} = v_{k+1} - \phi_{k+1} = e_k + h_k \varphi(x_k, v_k, h_k) + r_k + \lambda_k - h_k \varphi(x_k, \phi_k, h_k).
\]

Applying the norm and using the assumed Lipschitz continuity of \( \varphi \) with respect to \( y \) together with (2.19) then yields

\[
\|e_{k+1}\| \leq (1 + h_k L) \|e_k\| + h_k \left( C h_{\text{max}}^p + \frac{\delta}{h_{\min}} \right),
\]

which is (2.21a) with \( a_k = \|e_k\| \) and \( \beta = C h_{\text{max}}^p + \frac{\delta}{h_{\min}} \). Thus, (2.21b) yields

\[
\forall k \in \{0, \ldots, N\} \quad \|e_k\| \leq \frac{e^{L(x_k-\xi)} - 1}{L} \left( C h_{\text{max}}^p + \frac{\delta}{h_{\min}} \right) + \|e_0\| e^{L(x_k-\xi)},
\]

proving (2.25). \( \blacksquare \)

**Corollary 2.17.** Consider the setting of Def. 2.12. If the explicit single-step method under consideration is well-defined, consistent of order \( p \in \mathbb{N} \), and \( \varphi \) is globally \( L \)-Lipschitz with respect to \( y \), \( L > 0 \), then the method has order of convergence \( p \). More precisely,

\[
\forall \Delta \in \Pi([\xi, \delta]), h_{\text{max}}(\Delta) < \min\{h(\varphi), h(\lambda(\varphi))\} \leq \frac{C}{L} \left( e^{L(b-\xi)} - 1 \right),
\]

(2.27b)

\( h(\varphi) \) and \( h(\lambda(\varphi)) \) being as in Def. 2.12(a), (c), respectively, and \( C \) being the constant given by (2.19).
Proof. One merely has to set $e_0 := 0$ and $\delta := 0$ in Th. 2.16. ■

Corollary 2.18. If one considers the situation of Th. 2.16 with $e_0 = 0$ and equidistant stepizes (i.e. $h := h_{\min}(\Delta) = h_{\max}(\Delta) < \min\{h(\varphi), h_\lambda(\varphi)\}$), then the bound for the total error given by (2.25) is minimized for

$$h := h_{\text{opt}} := \left(\frac{\delta}{p}\right)^\frac{1}{p+1},$$

leading to

$$\max \left\{ \|v_k - \phi(x_k)\| : k \in \{0, \ldots, N\} \right\} \leq (1 + p) K \left(\frac{\delta}{p}\right)^\frac{p}{p+1},$$

where

$$K := \max \{C, 1\} \left(\frac{e^{L(b-\xi)} - 1}{L} \right).$$

Proof. For $e_0 = 0$ and equidistant stepizes, the estimate in (2.25) becomes

$$\max \left\{ \|v_k - \phi(x_k)\| : k \in \{0, \ldots, N\} \right\} \leq \frac{e^{L(b-\xi)} - 1}{L} \left( Ch^p + \delta \right)$$

$$\leq \frac{e^{L(b-\xi)} - 1}{L} \left( Ch^p + \frac{\delta}{h} \right) =: \alpha(h).$$

Clearly, $\alpha$ is differentiable on $\mathbb{R}^+$ and $\alpha'(h) = K(\delta h^{p-1} - \frac{\delta}{h})$. Thus, $\alpha'$ has its only zero at $h_{\text{opt}}$, where the sign changes from negative to positive, proving $\alpha$ to be minimal at $h_{\text{opt}}$. Moreover,

$$\alpha(h_{\text{opt}}) = K \left(\frac{\delta}{p}\right)^\frac{p}{p+1} + \delta \left(\frac{p}{\delta}\right)^\frac{1}{p+1} = K \left(\frac{\delta}{p}\right)^\frac{p}{p+1} \left(1 + \delta \left(\frac{p}{\delta}\right)^\frac{1}{p+1} \left(\frac{p}{\delta}\right)^\frac{1}{p+1} \right)$$

$$= (1 + p) K \left(\frac{\delta}{p}\right)^\frac{p}{p+1},$$

proving (2.28b). ■

Remark 2.19. We conclude from Cor. 2.18 that it is actually counterproductive to reduce the stepsize to values below $h_{\text{opt}}$ if rounding errors are known to be of order $\delta$.

Theorem 2.20. In the setting of Def. 2.12, consider the explicit Euler method of Def. 2.9 and assume the local truncation error $\lambda$ is well-defined (for example, $G = \mathbb{R} \times \mathbb{K}^n$ is sufficient).

(a) If $f \in C^1(G, \mathbb{K}^n)$, $G \subseteq \mathbb{R} \times \mathbb{K}^n$ open, then the method is consistent of order 1.

(b) If, in addition to $f \in C^1(G, \mathbb{K}^n)$, the method is well-defined (again, $G = \mathbb{R} \times \mathbb{K}^n$ being sufficient) and $f$ is globally Lipschitz with respect to $y$, then the method has order of convergence 1.
Proof. (a): If \( f \in C^1(G, \mathbb{K}^n) \), then the solution \( \phi : I \rightarrow \mathbb{K}^n \) to (2.9) is \( C^2 \) according to Prop. C.1 in the Appendix. Thus, we can apply Taylor’s theorem with the remainder term in integral form to obtain, for each \( x \in [\xi, b[ \) and each \( h \in ]0, b - x[ \),

\[
\phi(x + h) = \phi(x) + \phi'(x)h + \int_x^{x+h} (x + h - t) \phi''(t) dt.
\] (2.29)

Applying (2.29) in the local truncation error yields, for each \( x \in [\xi, b[ \) and each sufficiently small \( h > 0 \),

\[
\lambda(x, h) = \phi(x) + h \varphi(x, \phi(x), h) - \phi(x + h) = - \int_x^{x+h} (x + h - t) \phi''(t) dt,
\]

and, thus,

\[
\|\lambda(x, h)\| \leq C h^2,
\]

where

\[
C = \frac{1}{2} \max\{\|\phi''(t)\| : t \in [\xi, b]\},
\]

showing the method to be consistent of order 1 according to Def. 2.12(c).

(b): Clearly, if \( f \) is globally Lipschitz with respect to \( y \), then so is \( \varphi \). Thus, the explicit Euler method has order of convergence 1 according to (a) and Cor. 2.17. ■

2.2.3 Classical Explicit Runge-Kutta Method

While the explicit Euler method discussed above is still relatively simple, it is also too inaccurate for most practical applications. The classical explicit Runge-Kutta (RK) method introduced below is slightly more involved, yielding a fourth-order method, often sufficiently good for serious use.

**Definition 2.21.** In the situation of Def. 2.7, we call the numerical method *classical explicit Runge-Kutta (RK) method* if, and only if, \( m = 1 \), \( \alpha_0 = 1 \), and the defining function is

\[
\varphi : \mathcal{D}_\varphi \rightarrow \mathbb{K}^n, \quad \varphi(x, y, h) := \frac{1}{6} \left( k_1(x, y) + 2k_2(x, y, h) + 2k_3(x, y, h) + k_4(x, y, h) \right),
\]

\[
k_1(x, y) := f(x, y),
\]

\[
k_2(x, y, h) := f \left( x + \frac{h}{2}, y + \frac{h}{2}k_1(x, y) \right),
\]

\[
k_3(x, y, h) := f \left( x + \frac{h}{2}, y + \frac{h}{2}k_2(x, y, h) \right),
\]

\[
k_4(x, y, h) := f \left( x + h, y + hk_3(x, y, h) \right),
\]

\[
\varphi(x, y, h) := \frac{1}{6} \left( k_1(x, y) + 2k_2(x, y, h) + 2k_3(x, y, h) + k_4(x, y, h) \right),
\]

\[
k_1(x, y) := f(x, y),
\]

\[
k_2(x, y, h) := f \left( x + \frac{h}{2}, y + \frac{h}{2}k_1(x, y) \right),
\]

\[
k_3(x, y, h) := f \left( x + \frac{h}{2}, y + \frac{h}{2}k_2(x, y, h) \right),
\]

\[
k_4(x, y, h) := f \left( x + h, y + hk_3(x, y, h) \right),
\]
i.e., given \( \xi = x_0 < x_1 < \ldots \) and \( h_k = x_{k+1} - x_k \), the classical explicit RK method consists of the recursion

\[
y_0 = \eta, \quad \forall k \in \{0, 1, \ldots\} \quad y_{k+1} = y_k + \frac{h_k}{6} \left( k_1(x_k, y_k) + 2k_2(x_k, y_k, h_k) + 2k_3(x_k, y_k, h_k) + k_4(x_k, y_k, h_k) \right).
\]

(2.30)

Remark 2.22. As for the explicit Euler method in Rem. 2.10(a), we have the issue that, in general, there is no guarantee \( y_{k+1} \) is well-defined by (2.30). More precisely, \( y_{k+1} \) is well-defined by (2.30) if, and only if,

\[
(x_k, y_k) \in G, \quad \left( x_k + h_k/2, y_k + h_kk_1(x_k, y_k)/2 \right) \in G,
\]

\[
\left( x_k + h_k/2, y_k + h_kk_2(x_k, y_k, h_k)/2 \right) \in G,
\]

and (2.30) is supposed to mean that the recursion continues as long as (2.31) holds, and it terminates with \( y_k \), if (2.31) fails.

Theorem 2.23. In the setting of Def. 2.12, consider the classical explicit RK method of Def. 2.21 and assume the local truncation error \( \lambda \) is well-defined (for example, \( G = \mathbb{R} \times \mathbb{K}^n \) is sufficient).

(a) If \( f \in C^4(G, \mathbb{K}^n), G \subseteq \mathbb{R} \times \mathbb{K}^n \) open, then the method is consistent of order 4.

(b) If, in addition to \( f \in C^4(G, \mathbb{K}^n) \), the method is well-defined (again, \( G = \mathbb{R} \times \mathbb{K}^n \) being sufficient) and \( f \) is globally Lipschitz with respect to \( y \), then the method has order of convergence 4.

Proof. (a): The proof that the classical explicit RK method is consistent of order 4 is already somewhat tedious. It can be carried out using Taylor’s theorem, using the same idea as in the proof of Th. 2.20, see [DB08, Sec. 4.2.2]. A more systematic approach that is also suitable for higher-order methods uses the graph-theoretic ideas of Sec. 2.3.5 below, where we will complete the proof in Ex. 2.62(b).

(b): If \( f \) is globally Lipschitz with respect to \( y \), then so is \( \varphi \) on \( D_{\varphi} \cap (G \times [0, \xi - b]) \): While this holds more generally (cf. Rem. 2.29(b) below), we will directly verify that \( \varphi \) is globally Lipschitz with respect to \( y \) in the present situation: If \( f \) is \( L \)-Lipschitz with
respect to \( y \), \( L \geq 0 \), and \((x, y, h), (x, \bar{y}, h) \in D_\varphi \), then

\[
\| k_1(x, y) - k_1(x, \bar{y}) \| \leq L \| y - \bar{y} \|,
\]
\[
\| k_2(x, y, h) - k_2(x, \bar{y}, h) \| \leq \left( L + \frac{hL^2}{2} \right) \| y - \bar{y} \|,
\]
\[
\| k_3(x, y, h) - k_3(x, \bar{y}, h) \| \leq \left( L + \frac{hL^2}{2} + \frac{h^2L^3}{4} \right) \| y - \bar{y} \|,
\]
\[
\| k_4(x, y, h) - k_4(x, \bar{y}, h) \| \leq \left( L + hL^2 + \frac{h^2L^3}{2} + \frac{h^3L^4}{4} \right) \| y - \bar{y} \|,
\]
\[
\| \varphi(x, y, h) - \varphi(x, \bar{y}, h) \| \leq \frac{1}{6} \left( 6L + 3hL^2 + h^2L^3 + \frac{h^3L^4}{4} \right) \| y - \bar{y} \|.
\]

For \( h \leq b - \xi \), we see \( \varphi \) to be Lipschitz with respect to \( y \). Thus, the classical explicit RK method has order of convergence 4 according to (a) and Cor. 2.17.

2.3 General Runge-Kutta Methods

2.3.1 Definition, Butcher Tableaus, First Properties

In generalization of the methods defined in Def. 2.9 and Def. 2.21, respectively, we define general RK methods:

**Definition 2.24.** Let \( s \in \mathbb{N} \) and consider the situation of Def. 2.7 (where, in particular, we had \( n \in \mathbb{N}, G \subseteq \mathbb{R} \times \mathbb{K}^n, f : G \rightarrow \mathbb{K}^n \)).

(a) We call a numerical method as defined in Def. 2.7 an \( s \)-stage Runge-Kutta (RK) method if, and only if, \( m = 1 \), \( \alpha_0 = 1 \), and the defining function has the form

\[
\varphi : D_\varphi \rightarrow \mathbb{K}^n, \quad \varphi(x, y, h) = \sum_{j=1}^{s} b_j k_j(x, y, h),
\]

where, for each \((x, y, h) \in D_\varphi\), the auxiliary vectors \( k_1(x, y, h), \ldots, k_s(x, y, h) \in \mathbb{K}^n \) satisfy the system

\[
\forall j \in \{1, \ldots, s\} \quad k_j(x, y, h) = f \left( x + c_j h, y + h \sum_{l=1}^{s} a_{jl} k_l(x, y, h) \right),
\]

the coefficients \( b_1, \ldots, b_s \in \mathbb{R} \) are called weights, \( c_1, \ldots, c_s \in \mathbb{R} \) are called nodes, and the matrix \( A := (a_{jl}) \in \mathcal{M}(s, \mathbb{K}) \) is called RK matrix. The coefficients are commonly compiled in a so-called Butcher tableau in the form

\[
\begin{array}{c|cccc}
 c_1 & a_{11} & \cdots & a_{1s} \\
 \vdots & \vdots & \ddots & \vdots \\
 c_s & a_{s1} & \cdots & a_{ss} \\
 \hline
 b_1 & \cdots & b_s
\end{array}
\]

or

\[
\begin{array}{c|c}
 c & A \\
 \hline
 b
\end{array}, \quad c := \begin{pmatrix} c_1 \\ \vdots \\ c_s \end{pmatrix}, \quad b := \begin{pmatrix} b_1 \\ \vdots \\ b_s \end{pmatrix}.
\]

(a) We call a numerical method as defined in Def. 2.7 an \( s \)-stage Runge-Kutta (RK) method if, and only if, \( m = 1 \), \( \alpha_0 = 1 \), and the defining function has the form

\[
\varphi : D_\varphi \rightarrow \mathbb{K}^n, \quad \varphi(x, y, h) = \sum_{j=1}^{s} b_j k_j(x, y, h),
\]

where, for each \((x, y, h) \in D_\varphi\), the auxiliary vectors \( k_1(x, y, h), \ldots, k_s(x, y, h) \in \mathbb{K}^n \) satisfy the system

\[
\forall j \in \{1, \ldots, s\} \quad k_j(x, y, h) = f \left( x + c_j h, y + h \sum_{l=1}^{s} a_{jl} k_l(x, y, h) \right),
\]

the coefficients \( b_1, \ldots, b_s \in \mathbb{R} \) are called weights, \( c_1, \ldots, c_s \in \mathbb{R} \) are called nodes, and the matrix \( A := (a_{jl}) \in \mathcal{M}(s, \mathbb{K}) \) is called RK matrix. The coefficients are commonly compiled in a so-called Butcher tableau in the form

\[
\begin{array}{c|cccc}
 c_1 & a_{11} & \cdots & a_{1s} \\
 \vdots & \vdots & \ddots & \vdots \\
 c_s & a_{s1} & \cdots & a_{ss} \\
 \hline
 b_1 & \cdots & b_s
\end{array}
\]

or

\[
\begin{array}{c|c}
 c & A \\
 \hline
 b
\end{array}, \quad c := \begin{pmatrix} c_1 \\ \vdots \\ c_s \end{pmatrix}, \quad b := \begin{pmatrix} b_1 \\ \vdots \\ b_s \end{pmatrix}.
\]
(b) An RK method is called \textit{explicit} if, and only if, the RK matrix is strictly lower triangular (i.e., lower triangular with all $a_{jj} = 0$); \textit{implicit} otherwise.

(c) We say that an RK method satisfies the \textit{consistency condition} (cf. Rem. 2.29(a) and Th. 2.31(c) below) if, and only if, the sum of the weights equals 1, i.e., if, and only if,

$$
\sum_{j=1}^{s} b_j = 1.
$$

(2.34)

(d) We say that an RK method satisfies the \textit{node condition} if, and only if,

$$
\forall j \in \{1, \ldots, s\}, \quad c_j = \sum_{l=1}^{s} a_{jl}.
$$

(2.35)

It is sometimes useful to rewrite an RK method by replacing the vectors $k_j$ by vectors $u_j$, making use of the following lemma:

\textbf{Lemma 2.25.} Let $s, n \in \mathbb{N}$, $G \subseteq \mathbb{R} \times \mathbb{K}^n$, $f : G \rightarrow \mathbb{K}^n$, $b_j, c_j \in \mathbb{R}$, $a_{jl} \in \mathbb{K}$ for each $j, l \in \{1, \ldots, s\}$.

(a) Let $(x, y, h) \in \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+$ and consider vectors $k_1, \ldots, k_s, u_1, \ldots, u_s \in \mathbb{K}^n$ as well as the systems

$$
\forall j \in \{1, \ldots, s\}, \quad k_j = f \left( x + c_j h, y + h \sum_{l=1}^{s} a_{jl} k_l \right),
$$

(2.36a)

$$
\forall j \in \{1, \ldots, s\}, \quad u_j = y + h \sum_{l=1}^{s} a_{jl} f \left( x + c_l h, u_l \right).
$$

(2.36b)

If $k_1, \ldots, k_s$ satisfy (2.36a) (which, in particular, is supposed to mean, for each $j \in \{1, \ldots, s\}$, that the used argument of $f$ is in $G$) and

$$
\forall j \in \{1, \ldots, s\}, \quad u_j := y + h \sum_{l=1}^{s} a_{jl} k_l,
$$

(2.37a)

then $u_1, \ldots, u_s$ satisfy (2.36b). Conversely, if $u_1, \ldots, u_s$ satisfy (2.36b) (which, in particular, is supposed to mean, for each $j \in \{1, \ldots, s\}$, that the used argument of $f$ is in $G$) and

$$
\forall j \in \{1, \ldots, s\}, \quad k_j := f \left( x + c_j h, u_j \right),
$$

(2.37b)

then $k_1, \ldots, k_s$ satisfy (2.36a).
(b) A function \( \varphi : \mathcal{D}_\varphi \rightarrow \mathbb{K}^n \) satisfies (2.32) for some \((x, y, h) \in \mathcal{D}_\varphi\) if, and only if, it satisfies
\[
\varphi(x, y, h) = \sum_{j=1}^{s} b_j f(x + c_j h, u_j(x, y, h)),
\]
where the auxiliary vectors \(u_1(x, y, h), \ldots, u_s(x, y, h) \in \mathbb{K}^n\) satisfy the system
\[
\forall j \in \{1, \ldots, s\} \quad u_j(x, y, h) = y + h \sum_{l=1}^{s} a_{jl} f(x + c_l h, u_l(x, y, h)).
\]

Proof. (a): If \(k_1, \ldots, k_s\) satisfy (2.36a) and \(u_1, \ldots, u_s\) are given by (2.37a), then (2.36a) can be written as
\[
\forall j \in \{1, \ldots, s\} \quad k_j = f(x + c_j h, u_j).
\]
Plugging these expressions for the \(k_j\) back into (2.37a) then yields (2.36b). Now, if \(u_1, \ldots, u_s\) satisfy (2.36b) and \(k_1, \ldots, k_s\) are given by (2.37b), then (2.36b) can be written as
\[
\forall j \in \{1, \ldots, s\} \quad u_j = y + h \sum_{l=1}^{s} a_{jl} k_l.
\]
Plugging these expressions for the \(u_j\) back into (2.37b) then yields (2.36a).
(b) is now immediate from (a).
\[\blacksquare\]

Definition 2.26. Consider the situation of Def. 2.24.

(a) If the defining function \(\varphi\) of an RK method, as defined in Def. 2.24(a), is written in the form (2.32), then we say it is written in \(k\)-form. If \(\varphi\) is written in the (equivalent) form (2.38), then we say it is written in \(u\)-form.

(b) An RK method is said to have unique local solutions if, and only if, for each \((x, y) \in G\), there exist \(\eta(x, y) \in [0, \infty]\) and \(r(x, y) \in \mathbb{R}^+\) such that, if \(h \in [0, \eta(x, y)]\), then the system (2.38b) has a unique (local) solution \(u(x, y, h) \in B_{r(x,y)}(y, \ldots, y) \subseteq (\mathbb{K}^n)^s \cong \mathbb{K}^{ns}\) (as all norms on \(\mathbb{K}^{ns}\) are equivalent, the existence of \(\eta(x, y)\) and \(r(x, y)\) does not depend on the chosen norm – however, the size of these numbers will, in general, depend on the chosen norm). If the RK method has unique local solutions, then we say it is in standard form if, and only if, for each \((x, y) \in G\) and \(h \in [0, \eta(x, y)]\), we have \((x, y, h) \in \mathcal{D}_\varphi\) and \(u(x, y, h) \in B_{r(x,y)}(y, \ldots, y)\).

Remark 2.27. (a) Given \(\xi = x_0 < x_1 < \ldots\) and \(h_k := x_{k+1} - x_k\), an RK method given according to Def. 2.24(a) consists of the recursion \(y_0 = \eta \in \mathbb{K}^n\),
\[
y_{k+1} = y_k + h_k \sum_{j=1}^{s} b_j k_j(x_k, y_k, h_k) \tag{2.32a}
\]
\[
y_{k+1} = y_k + h_k \sum_{j=1}^{s} b_j f(x_k + c_j h_k, u_j(x_k, y_k, h_k)) \tag{2.38a}
\]
\[
y_{k+1} = y_k + h_k \sum_{j=1}^{s} b_j f(x_k + c_j h_k, u_j(x_k, y_k, h_k)) \tag{2.39}
\]
for $k \in \{0, 1 \ldots \}$, where the $k_j$ and $u_j$ satisfy systems given by (2.32b) and (2.38b), respectively.

(b) From (2.39), we note that every RK method given by a defining function as in (2.32) (or as in (2.38)) is an explicit single-step method according to Def. 2.7, even if the $a_{jl}$ yield an implicit RK method (as the right-hand side of (2.39) does not depend on $y_{k+1}$). However, if the RK method is explicit, then $k_1, \ldots, k_s$ can be computed recursively from (2.32b), whereas, for an implicit RK method, the (in general, nonlinear) system of equations for the $k_j$ will, in general, only have a unique solution under additional hypotheses (and, likewise, for the $u_1, \ldots, u_s$). As mentioned before in Rem. 2.8(b), the explicit/implicit distinction of Def. 2.7 is more a property of the presentation of the method rather than of the method itself. To illustrate this further, we can, e.g., rewrite (2.39) in the implicit form

$$y_{k+1} = y_k + h_k \sum_{j=1}^s b_j f(x_k + c_j h_k, \tilde{u}_j(x_k, y_{k+1}, h_k));$$

where (replacing $y_k$ in the expression for $u_j(x_k, y_k, h_k)$ via (2.39))

$$\tilde{u}_j(x_k, y_{k+1}, h_k) \overset{(2.39)}{=} u_j(x_k, y_k, h_k)$$

$$\tilde{u}_j(x_k, y_{k+1}, h_k) \overset{(2.39)}{=} y_k + h_k \sum_{l=1}^s (a_{jl} - b_j) f(x_k + c_l h_k, \tilde{u}_l(x_k, y_{k+1}, h_k)).$$

While this representation is, in most cases, not of practical use, it shows, e.g., the equivalence of an implicit RK method with the so-called implicit Euler method (see Ex. 2.28(c) below).

**Example 2.28.** (a) Comparing Def. 2.9 with Def. 2.24, we see that the explicit Euler method is a 1-step RK method with Butcher tableau

| 0 | 0 | 1 |

(b) Comparing Def. 2.21 with Def. 2.24, we see that the classical explicit RK method is a 4-step RK method with Butcher tableau

| 0 | \frac{1}{2} | \frac{1}{2} | \frac{1}{2} |
| \frac{1}{2} | 0 | \frac{1}{2} | \frac{1}{2} |
| 1 | 0 | 0 | 1 |

| \frac{1}{6} | \frac{1}{3} | \frac{1}{3} | \frac{1}{6} |

(c) Consider the implicit 1-step RK method with Butcher tableau

| 1 | 1 | 1 | 1 |
Using the representation (2.40), we obtain, since \(a_{11} = b_1 = 1\),
\[
\tilde{u}_1(x_k, y_{k+1}, h_k)^{(2.40b)} = y_{k+1} + h_k (a_{11} - b_1) f(x_k + c_1 h_k, \tilde{u}_1(x_k, y_{k+1}, h_k)) = y_{k+1}.
\]

Thus,
\[
y_{k+1} = y_k + h_k f(x_k + c_1 h_k, \tilde{u}_1(x_k, y_{k+1}, h_k)) = y_k + h_k f(x_{k+1}, y_{k+1}),
\]
showing that this RK method is precisely what is known in the literature as the implicit Euler method.

**Remark 2.29.** Let \(s \in \mathbb{N}\) and assume \(\varphi : \mathcal{D}_\varphi \rightarrow \mathbb{K}^n\) defines an explicit \(s\)-stage RK method according to Def. 2.24(a) (cf. Rem. 2.27). Recall \(f : G \rightarrow \mathbb{K}^n, G \subseteq \mathbb{R} \times \mathbb{K}^n\), \(\mathcal{D}_\varphi \subseteq \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+\).

(a) Since the RK matrix \(A\) is strictly lower triangular, the \(k_j, j \in \{1, \ldots, s\}\), are recursively well-defined by (2.32b) as functions \(k_j : \mathcal{D}_\varphi \rightarrow \mathbb{K}^n\) (likewise, the \(u_j : \mathcal{D}_\varphi \rightarrow \mathbb{K}^n\) are recursively well-defined by (2.38b)). If \(f\) is continuous, then so is \(\varphi\): Indeed, if \(f\) is continuous, then an induction on \(j \in \{1, \ldots, s\}\) shows each \(k_j\) (and each \(u_j\)) to be continuous, implying \(\varphi\) to be continuous as well. Moreover, if \(f\) is continuous and \(G\) is open, then another induction shows each \(k_j(x, y, h)\) (and each \(u_j(x, y, h)\)) to be defined if \((x, y) \in G\) and \(h\) is sufficiently small (in particular, the RK method then has unique local solutions in the sense of Def. 2.26(b)), and we obtain
\[
\forall j \in \{1, \ldots, s\} \lim_{h_i \downarrow 0} k_j(x, y, h) = f(x, y), \quad \lim_{h_i \downarrow 0} u_j(x, y, h) = y,
\]
implying
\[
\lim_{h_i \downarrow 0} \varphi(x, y, h) = f(x, y) \sum_{j=1}^s b_j.
\]

Thus, if \(f\) is continuous on \(G\) open in the situation of Def. 2.12 with the local truncation error \(\lambda\) being well-defined (\(G = \mathbb{R} \times \mathbb{K}^n\) being sufficient), then combining the above with (2.20) shows the method to be consistent in the sense of Def. 2.12(c) if it satisfies the consistency condition (2.34). The consistency condition is also necessary for the method to be consistent if there exists \(x\) in the domain of the solution \(\phi\) of \(y' = f(x, y), y(\xi) = \eta\), with \(f(x, \phi(x)) \neq 0\).

(b) If \(f\) is globally Lipschitz with respect to \(y\), then so is \(\varphi\) on \(D := \mathcal{D}_\varphi \cap (G \times [0, T])\), \(T > 0\): Indeed, if \(f\) is Lipschitz with respect to \(y\), then an induction on \(j \in \{1, \ldots, s\}\) shows each \(k_j\) (and each \(u_j\)) to be Lipschitz with respect to \(y\) on \(D\), implying \(\varphi\) to be Lipschitz with respect to \(y\) on \(D\) as well.

(c) If \(f \in C^k(G, \mathbb{K}^n)\) and \(\mathcal{D}_\varphi\) is open, then \(\varphi \in C^k(\mathcal{D}_\varphi, \mathbb{K}^n)\): One can show (again using straightforward inductions together with the chain rule) that each partial of order \(m\) of \(k_j\) (and of \(u_j\), \(m \in \{1, \ldots, k\}\), \(j \in \{1, \ldots, s\}\) is a polynomial in partials of order \(\leq m\) of \(f\), implying the same for the partials of \(\varphi\).
2.3.2 Existence and Uniqueness

We will now turn to the solvability question regarding the systems (2.32b) and (2.38b) for implicit RK methods. In Th. 2.31 below, we will prove existence and (local) uniqueness of solutions to (2.32b) and (2.38b), provided $f$ is continuous and locally Lipschitz with respect to $y$. The proof will be based on the following Prop. 2.30, which can be viewed as a parametrized version of the Banach fixed point theorem (cf. [Phi16b, Th. 2.29]).

**Proposition 2.30.** Let $(X, d)$, $(Y, d)$ be metric spaces, where the metric space $(Y, d)$ is complete. Let $c \in Y$, $r \in \mathbb{R}^+$, consider the ball $B := B_r(c) \subseteq Y$ and a continuous map $F : X \times B \rightarrow Y$, that is Lipschitz with respect to $y$, satisfying

$$\forall (x,y_1), (x,y_2) \in X \times B \quad d(F(x,y_1), F(x,y_2)) \leq L \, d(y_1, y_2)$$

with some Lipschitz constant $0 \leq L < 1$. If $F$ also satisfies the condition

$$\forall x \in X \quad d(F(x,c), c) < r \, (1 - L),$$

then, for each $x \in X$, there exists a unique $g(x) \in B$ such that

$$g(x) = F(x, g(x)).$$

Moreover, the map $g : X \rightarrow B$, $x \mapsto g(x)$, is continuous and, for each $x \in X$, the sequence $(y_k(x))_{k \in \mathbb{N}_0}$, recursively defined by $y_0(x) := c$ and

$$y_{k+1}(x) := F(x, y_k(x)) \quad \text{for each } k \in \mathbb{N}_0,$$

converges to $g(x)$:

$$\lim_{k \rightarrow \infty} y_k(x) = g(x).$$

**Proof.** For $L = 0$, $F$ is constant and the assertion trivially holds, with $g$ being constant as well. Thus, for the rest of the proof, let $0 < L < 1$. Fix $x \in X$ and let the sequence $(y_k(x))_{k \in \mathbb{N}_0}$ be defined by $y_0(x) := c$ and (2.44). To verify that the sequence is well-defined, we prove, using an induction, it is a sequence in $B$: $y_0(x) = c \in B$ and

$$d(y_1(x), c) = d(F(x,c), c) < r \, (1 - L) \leq r,$$

implies $y_1(x) \in B$. Now let $k, l \in \mathbb{N}_0$, $k > l$. Then, by induction, $y_0, \ldots, y_k \in B$, and

$$d(y_{k+1}(x), y_k(x)) \leq L \, d(y_k(x), y_{k-1}(x)) \leq L^{k-l} \, d(y_{l+1}(x), y_l(x)) \quad (2.46)$$

Thus, for each $l, j \in \mathbb{N}_0$ with $l + j \leq k + 1$,

$$d(y_{l+j}(x), y_l(x)) \leq \sum_{m=l}^{l+j-1} d(y_{m+1}(x), y_m(x)) \leq \sum_{m=l}^{l+j-1} L^{m-l} \, d(y_{l+1}(x), y_l(x))$$

$$\leq \frac{1}{1 - L} \, d(y_{l+1}(x), y_l(x)) \leq \frac{L^l}{1 - L} \, d(y_1(x), y_0(x))$$

$$= \frac{L^l}{1 - L} \, d(y_1(x), c) \quad (2.42), 0 < L \leq L^l \, r \leq r. \quad (2.47)$$
In particular, using \( l := 0 \) and \( j := k + 1 \) in (2.47) shows \( y_{k+1}(x) \in B \), completing the induction proof of \( \{ y_k(x) \}_{k \in \mathbb{N}_0} \) being a well-defined sequence in \( B \). The sequences now give rise to the functions

\[
 f_k : X \rightarrow B, \quad f_k(x) := y_k(x),
\]
defined for each \( k \in \mathbb{N}_0 \). Using the assumed continuity of \( F \) together with an induction on \( k \in \mathbb{N}_0 \), we verify the \( f_k \) to be continuous: \( f_0 \equiv c \) is constant and, hence, continuous. Now let \( k \in \mathbb{N}_0 \). Since \( f_{k+1}(x) = F(x, f_k(x)) \) and \( f_k \) is continuous by induction, \( f_{k+1} \) is continuous as well. To proceed, we, once again, fix \( x \in X \). As \( \lim_{l \rightarrow \infty} L^l = 0 \), (2.47) proves \( \{ f_k(x) \}_{k \in \mathbb{N}_0} \) to be a Cauchy sequence in \( Y \). As \( Y \) is complete, we may define

\[
 g : X \rightarrow Y, \quad g(x) := \lim_{k \rightarrow \infty} f_k(x).
\]

Taking \( l = 0 \) in (2.47) shows

\[
 \forall \quad x \in X \quad d(g(x), c) = \lim_{j \rightarrow \infty} d(y_j(x), c) \overset{(2.47)}{\leq} \frac{1}{1-L} d(y_1(x), c) < r,
\]
i.e. \( g \) maps into \( B \), as desired. The continuity of \( F \) allows one to take limits in (2.44), showing, for each \( x \in X \),

\[
 g(x) = \lim_{k \rightarrow \infty} y_{k+1}(x) = \lim_{k \rightarrow \infty} F(x, y_k(x)) = F(x, g(x)),
\]
i.e. (2.43) is satisfied. We need to show \( g \) is continuous. To this end, we fix \( x \in X \) and \( l \in \mathbb{N}_0 \) in (2.47). For \( j \rightarrow \infty \), we then obtain

\[
 d(g(x), f_l(x)) \leq \frac{L^l}{1-L} d(y_1(x), c) < L^l r,
\]
showing \( f_l \) to converge uniformly to \( g \). As a uniform limit of the continuous functions \( f_l \), the function \( g \) is itself continuous (see, e.g., [Phi16c, Th. 3.5]).

Given \( x \in X \), it remains to show the uniqueness of the \( g(x) \), satisfying (2.43). Suppose, \( y \in B \) also satisfies \( y = F(x, y) \). Then

\[
 d(g(x), y) = d(F(x, g(x)), F(x, y)) \overset{(2.41)}{\leq} L d(g(x), y),
\]
which implies \( 1 \leq L \) for \( d(g(x), y) > 0 \). Thus, \( L < 1 \) implies \( d(g(x), y) = 0 \) and \( g(x) = y \). \( \blacksquare \)

**Theorem 2.31.** Let \( s, n \in \mathbb{N} \), let \( G \subseteq \mathbb{R} \times \mathbb{K}^n \) be open and assume \( f : G \rightarrow \mathbb{K}^n \) to be continuous and locally Lipschitz with respect to \( y \); let \( b_j, c_j \in \mathbb{R} \), \( a_{jl} \in \mathbb{K} \) for each \( j, l \in \{1, \ldots, s\} \); and let \( \|A\|_\infty \) denote the operator norm of the RK matrix \( A = (a_{jl}) \) with respect to \( \| \cdot \|_\infty \) on \( \mathbb{K}^s \).

(a) For each \( (x, y) \in G \), there exist \( \eta(x, y) \in ]0, \infty] \) and \( r(x, y) \in \mathbb{R}^+ \) such that, if \( h \in ]-\eta(x, y), \eta(x, y)[ \), the system (2.36b) has a unique (local) solution

\[
 u(x, y, h) \in B_{r(x,y)}(y, \ldots, y) \subseteq (\mathbb{K}^n)^s \cong \mathbb{K}^{ns}
\]
(in particular, the resulting RK method has unique local solutions in the sense of Def. 2.26(b)). Moreover, the function
\[ g_{x,y} : [0, \eta(x,y)] \to B_{r(x,y)}(y, \ldots, y), \quad g_{x,y}(h) := u(x,y,h), \]
is continuous with
\[ g_{x,y}(0) = u(x,y,0) = (y, \ldots, y). \] (2.48)

If \( f \in C^p(G, \mathbb{K}^n) \), \( p \in \mathbb{N} \), then one can choose \( \eta(x,y) \) such that \( g_{x,y} \) is \( p \) times continuously differentiable.

(b) If, additionally, \( G = \mathbb{R} \times \mathbb{K}^n \) and \( f \) is globally \( L \)-Lipschitz with respect to \( y \) (\( L \in \mathbb{R}^+ \)), then, for each \( h \in [-\eta, \eta[ \) with
\[ \eta := \frac{1}{L \| A \|_{\infty}} \quad (1/0 := \infty), \]
the system (2.36b) has a unique (global) solution \( u(x,y,h) \in (\mathbb{K}^n)^s \cong \mathbb{K}^{ns} \).

(c) If the corresponding RK method is in standard form according to Def. 2.26(b) and \((x,y) \in G\), then
\[ \forall j \in \{1,\ldots,s\} \lim_{h \downarrow 0} u_j(x,y,h) = y, \quad \lim_{h \downarrow 0} k_j(x,y,h) = f(x,y), \]
implying
\[ \lim_{h \downarrow 0} \varphi(x,y,h) = f(x,y) \sum_{j=1}^s b_j. \]

If the local truncation error \( \lambda \) is well-defined (\( G = \mathbb{R} \times \mathbb{K}^n \) being sufficient), then the RK method is consistent in the sense of Def. 2.12(c) if it satisfies the consistency condition (2.34). The consistency condition is also necessary for the method to be consistent if there exists \( x \) in the domain of the solution \( \phi \) of \( y' = f(x,y), \ y(\xi) = \eta \), with \( f(x,\phi(x)) \neq 0 \).

Proof. (a): Fix an arbitrary norm \( \| \cdot \| \) on \( \mathbb{K}^n \) and fix \((x,y) \in G\). Set \( \bar{y} := (y, \ldots, y) \in \mathbb{K}^{ns} \). The idea is to find \( \eta := \eta(x,y) \in \mathbb{R}^+ \) and \( r := r(x,y) \in \mathbb{R}^+ \) such that we can write (2.36b) in the form
\[ u = F(h,u) \]
(corresponding to (2.43)) with
\[ F : X \times B \to \mathbb{K}^{ns}, \quad F_j(h,u) := y + h \sum_{l=1}^{s} a_{jl} f(x + c_l h, u_l) \in \mathbb{K}^n \quad (j \in \{1,\ldots,s\}), \] (2.49)
where \( X := [-\eta,\eta[ \) and \( B := B_r(\bar{y}) \subseteq \mathbb{K}^{ns} \), and to solve \( u = F(h,u) \) by applying Prop. 2.30. If \( A = 0 \), then \( F \equiv \bar{y} \) is constant and the assertions are trivially true. Thus, for the rest of the proof, we assume \( A \neq 0 \), i.e. \( \| A \|_{\infty} > 0 \). As \( G \) is open and \( f \) is locally
Lipschitz with respect to $y$, there exist numbers $\alpha := \alpha(x, y) \in \mathbb{R}^+, r := r(x, y) \in \mathbb{R}^+$, and $L := L(x, y) \in \mathbb{R}_{>0}^+$ such that $K := [x - \alpha, x + \alpha] \times \overline{B}_r(y) \subseteq G$ and

$$\forall (t, z_1), (t, z_2) \in K \quad \|f(t, z_1) - f(t, z_2)\| \leq L\|z_1 - z_2\|.$$ 

As the continuous map $f$ is bounded on compact sets, we may choose $S \in \mathbb{R}^+$ with

$$S > M := \max \{\|f(t, y)\| : t \in [x - \alpha, x + \alpha]\}.$$ 

We now fix some $\theta \in ]0, 1[$ (the rest of the proof will work for each such $\theta \in ]0, 1[$) and define

$$\eta := \min \left\{ \frac{\alpha}{\|c\|_{\infty}}, \frac{r(1 - \theta)}{S\|A\|_\infty}, \frac{\theta}{L\|A\|_{\infty}} \right\}, \quad c := (c_1, \ldots, c_s), \quad 1/0 := \infty \quad (2.50)$$

(note $0 < \eta < \infty$, as $\|A\|_{\infty} > 0$ implies $r(1 - \theta)/S\|A\|_{\infty} < \infty$). For the norm on $\mathbb{K}^n$, we choose the one defined by

$$\forall u \in \mathbb{K}^n \quad \|u\| := \max \{\|u_l\| : l \in \{1, \ldots, s\}\}.$$ 

Then $B = B_r(\tilde{y}) = B_r(y) \times \cdots \times B_r(y)$. The next step is the verification that $F$, as defined in (2.49), is well-defined and satisfies the hypotheses of Prop. 2.30. If $h \in X = ] - \eta, \eta[$, then, for $0 < \|c\|_{\infty}$,

$$\forall l \in \{1, \ldots, s\} \quad |x + c_l h - x| \leq \|c\|_{\infty} |h| < \|c\|_{\infty} \eta \leq \alpha.$$ 

For $\|c\|_{\infty} = 0$, $|x + c_l h - x| = 0 < \alpha$ also holds. Thus, in each case, if $(h, u) \in X \times B$, then, for each $l \in \{1, \ldots, s\}$, $(x + c_l h, u_l) \in [x - \alpha, x + \alpha] \times \overline{B}_r(y) \subseteq G$, i.e. $F$ is well-defined. The continuity of $F$ is then clear from the assumed continuity of $f$. Next, for each $j \in \{1, \ldots, s\}$, each $h \in X$, and each $u, \bar{u} \in B$, one estimates

$$\|F_j(h, u) - F_j(h, \bar{u})\| \leq \eta \sum_{l=1}^{s} |a_{jl}| \|f(x + c_l h, u_l) - f(x + c_l h, \bar{u}_l)\| \leq \eta L \sum_{l=1}^{s} |a_{jl}| \|u_l - \bar{u}_l\|.$$ 

In consequence, we obtain

$$\|F(h, u) - F(h, \bar{u})\| \leq \eta L\|A\|_{\infty}\|u - \bar{u}\| \leq \theta\|u - \bar{u}\|,$$

proving $F$ to satisfy (2.41) (where our $0 < \theta < 1$ plays the role of $L$ in (2.41)). Moreover, for each $h \in X$,

$$\|F(h, \tilde{y}) - \tilde{y}\| = \max \{\|F_j(h, \tilde{y}) - y\| : j \in \{1, \ldots, s\}\} \leq \eta \sum_{l=1}^{s} |a_{jl}| \|f(x + c_l h, y)\| \leq \eta\|A\|_{\infty} M < r(1 - \theta)$$

where $\eta := \min \{\alpha/(\|c\|_{\infty}), \theta/L\|A\|_{\infty}\}$. For the norm on $\mathbb{K}$, we choose the one defined by

$$\|x\| := \sum_{l=1}^{s} \|x_l\|.$$
(recalling $\|A\|_{\infty}$ to be the row sum norm of $A$). Thus, $F$ satisfies (2.42) as well (where our $\bar{y}$ plays the role of $c$ in (2.42)). We now employ Prop. 2.30 to obtain, for each $h \in X$, a unique $g_{x,y}(h) \in B$, satisfying

$$g_{x,y}(h) = F(h, g_{x,y}(h)),$$

as desired. In addition, Prop. 2.30 yields the continuity of the map $g_{x,y} : X \rightarrow B$, $h \mapsto g_{x,y}(h)$. In particular, (2.48) is satisfied due to $F(0, \cdot) \equiv \bar{y}$. If $f \in C^p(G, \mathbb{K}^n)$, $p \in \mathbb{N}$, then the map

$$\Gamma : X \times B \rightarrow \mathbb{K}^{ns}, \quad \Gamma(h, u) := u - F(h, u),$$

is $C^p$ as well, and we can apply the implicit function theorem [Phi16b, Th. 4.49] to the equation

$$\forall (h, u) \in U \times V \quad \left( \Gamma(h, u) = 0 \Leftrightarrow u = g(h) \right).$$

By possibly decreasing the numbers $\eta$ and $r$ from above, we may assume $U = X$ and $V = B$, in which case $g = g_{x,y}$.

(b): As in (a), the case $\|A\|_{\infty} = 0$ is clear, as $F$ is then constant. Thus, we proceed, assuming $\|A\|_{\infty} > 0$. If $G = \mathbb{R} \times \mathbb{K}^n$ and $f$ is globally $L$-Lipschitz with respect to $y$, then, in the proof of (a), for each $\alpha > 0$ and each $r > 0$, the map $F$ is well-defined and $f$ is globally $L$-Lipschitz with respect to $y$ on $K$. Seeking a contradiction, let $\eta_0 := \frac{1}{\|A\|_{\infty}}$ and assume there exists $h \in [-\eta_0, \eta_0]$ and $v, w \in \mathbb{K}^{ns}$ with $v \neq w$, $v = F(h, v)$, and $w = F(h, w)$. Choose

$$\alpha := \begin{cases} \frac{\|c\|_{\infty}}{L\|A\|_{\infty}} & \text{for } \|c\|_{\infty} > 0, \\ \frac{1}{L\|A\|_{\infty}} & \text{for } \|c\|_{\infty} = 0, \end{cases}$$

and choose $\theta \in ]0, 1[$ sufficiently close to 1 such that $\frac{\theta}{L\|A\|_{\infty}} > |h|$. Then choose

$$r > \max \left\{ \|v - \bar{y}\|, \|w - \bar{y}\|, \frac{S\|A\|_{\infty}|h|}{1 - \theta} \right\}.$$

The number $\eta \in \mathbb{R}^+$, defined according to (2.50), then satisfies $\eta > |h|$ and (a) implies the contradiction $v = g_{x,y}(h) = w$.

(c): If the RK method is in standard form, then, with $g_{x,y}$ according to (a), $u(x, y, h) = g_{x,y}(h)$ must hold for each sufficiently small $h$. Then the continuity of $g_{x,y}$ implies

$$\lim_{h \downarrow 0} u(x, y, h) = \lim_{h \downarrow 0} g_{x,y}(h) = g_{x,y}(0) = (2.48), (y, \ldots, y)$$
and, for each \( j \in \{1, \ldots, s\} \), also using the continuity of \( f \),
\[
\lim_{h \to 0} k_j(x, y, h) = \lim_{h \to 0} f(x + c_j h, u_j(x, y, h)) = f(x, y).
\]
Thus,
\[
\lim_{h \to 0} \varphi(x, y, h) = \lim_{h \to 0} \sum_{j=1}^{s} b_j f(x + c_j h, u_j(x, y, h)) = f(x, y) \sum_{j=1}^{s} b_j,
\]
as claimed. If \( x \) is in the domain of the solution \( \phi \) to \( y' = f(x, y) \), \( y(\xi) = \eta \), then
\[
\lim_{h \to 0} \varphi(x, \phi(x), h) = f(x, \phi(x)) \text{ if, and only if, } \sum_{j=1}^{s} b_j = 1 \text{ or } f(x, \phi(x)) = 0.
\]
Thus, Lem. 2.13(a) establishes the case.

**Remark 2.32.** In the situation Th. 2.31, assume \( f : \mathbb{K}^n \to \mathbb{K}^n \) (i.e. \( f \) does not depend on \( x \) – the ODE is autonomous, cf. Sec. 2.3.3 below). Moreover, assume \( f \) to be locally Lipschitz, but not globally Lipschitz. According to Prop. B.3 of the Appendix, \( f \) is globally Lipschitz on each compact subset \( C \subseteq \mathbb{K}^n \). In consequence, for each \( y \in \mathbb{K}^n \) and each \( r \in \mathbb{R}^+ \), there exists a smallest Lipschitz constant \( L_r(y) \in \mathbb{R}_0^+ \) such that \( f \) is \( L_r(y) \)-Lipschitz on \( B_r(y) \) (with \( \lim_{r \to \infty} L_r(y) \to \infty \), as \( f \) is not globally Lipschitz). As in the proof of Th. 2.31(b), one sees that, for each \( h \in ]-\eta, \eta[ \) with \( \eta := \frac{1}{L_r(y)\|A\|_\infty} \), the system (2.36b) has a unique solution \( u(x, y, h) \in B_r(y) \). Thus, if \( \tilde{u}(h) \in \mathbb{K}^n \) are such that, for each sufficiently small \( |h| \in \mathbb{R}^+ \), \( \tilde{u}(h) \) is a solution to (2.36b) with \( \tilde{u}(h) \neq u(x, y, h) \), then
\[
\lim_{h \to 0} \|\tilde{u}(h)\| = \infty.
\]
One can actually still conduct a similar argument if \( f \) is defined on an arbitrary open \( \Omega \subseteq \mathbb{K}^n \), in which case one obtains the \( \tilde{u}(h) \) to go to the boundary of \( G \) for \( h \to 0 \), in the sense that they escape every compact subset \( C \) of \( G \) forever for \( h \to 0 \) (cf. [Phi16c, Def. 3.25]).

**Example 2.33.** We consider the implicit Euler method of Ex. 2.28(c) for the initial value problem
\[
y' = \mu (1 - y^2), \quad y(\xi) = \eta, \quad (\xi, \eta) \in \mathbb{R}^2, \quad \mu \in \mathbb{R}^+.
\]
Here, we have \( f : \mathbb{R}^2 \to \mathbb{R} \), \( f(x, y) := \mu (1 - y^2) \) (which is locally Lipschitz, but not globally Lipschitz), and \( b_1 = c_1 = a_{11} = 1 \). For \( (x, y, h) \in \mathbb{R}^2 \times (\mathbb{R} \setminus \{0\}) \), (2.36b) reduces to
\[
u := u_1 = y + h f(x + h, u) = y + h \mu (1 - u^2)
\]
\begin{equation}
(2.51)
\end{equation}
or, equivalently,
\[
u^2 + \frac{u}{h \mu} - \frac{y + h \mu}{h \mu} = 0.
\]
This quadratic equation for \( u \) has the solutions
\[
u = u(h) = -\frac{1}{2 h \mu} \pm \sqrt{\frac{1}{4 h^2 \mu^2} + \frac{y + h \mu}{h \mu}} = -\frac{1}{2 h \mu} \pm \frac{1}{2 h \mu} \sqrt{1 + 4 h \mu (y + h \mu)}.
\]
which are real, provided $|h|$ is sufficiently small, guaranteeing $|4h\mu (y + h\mu)| \leq 1$. Consistently with Rem. 2.32, we obtain

$$\lim_{h \to 0} \left| \frac{-1 - \sqrt{1 + 4h\mu (y + h\mu)}}{2h\mu} \right| = \infty$$

and, consistently with (2.48), we obtain

$$\lim_{h \to 0} \frac{-1 + \sqrt{1 + 4h\mu (y + h\mu)}}{2h\mu} = \lim_{h \to 0} \frac{4h\mu (y + h\mu)}{2h\mu (1 + \sqrt{1 + 4h\mu (y + h\mu)})} = y.$$ 

Thus, to obtain the RK method in standard form, we choose $\eta(y) \in \mathbb{R}^+$ such that $|4h\mu (y + h\mu)| \leq 1$ for each $h \in ] - \eta(y), \eta(y)[$, and

$$u : \mathbb{R}^2 \times ]0, \eta(y)[ \to \mathbb{R}, \quad u(x, y, h) := \frac{-1 + \sqrt{1 + 4h\mu (y + h\mu)}}{2h\mu}.$$ 

Then the resulting $\varphi$ according to (2.38a) is $\varphi : \mathbb{R}^2 \times ]0, \eta(y)[ \to \mathbb{R}$,

$$\varphi(x, y, h) := f(x + h, u(x, y, h)) = \mu (1 - u^2(x, y, h)).$$

Given $(\xi, \eta) \in \mathbb{R}^2$ with $\xi = x_0 < x_1 < \ldots$ and $h_k := x_{k+1} - x_k$ sufficiently small, this yields the recursion $y_0 = \eta$,

$$y_{k+1} = y_k + h_k\mu \left(1 - u^2(x_k, y_k, h_k)\right) \overset{(2.51)}{=} u(x_k, y_k, h_k) = \frac{-1 + \sqrt{1 + 4h_k\mu (y_k + h_k\mu)}}{2h_k\mu}.$$ 

### 2.3.3 Autonomous ODE

Determining the parameters of RK methods such that a high order of convergence is obtained is, in general, a difficult task. It is somewhat simplified when one restricts oneself to so-called autonomous ODE:

**Definition 2.34.** If $\Omega \subseteq \mathbb{K}^n$, $n \in \mathbb{N}$, and $f : \Omega \to \mathbb{K}^n$, then the $n$-dimensional first-order ODE

$$y' = f(y) \quad (2.52)$$

is called autonomous.

As it turns out, every nonautonomous ODE can equivalently be written as an autonomous ODE:

**Theorem 2.35.** Let $G \subseteq \mathbb{R} \times \mathbb{K}^n$, $n \in \mathbb{N}$, and $f : G \to \mathbb{K}^n$. Then the nonautonomous ODE

$$y' = f(x, y) \quad (2.53a)$$
is equivalent to the autonomous ODE
\[ y' = g(y), \] (2.53b)
where
\[ g : G \rightarrow \mathbb{K}^{n+1}, \quad g(x, y_1, \ldots, y_n) := (1, f(x, y_1, \ldots, y_n)), \] (2.54)
in the following sense:

(a) If \( \phi : I \rightarrow \mathbb{K}^n \) is a solution to (2.53a), then \( \psi : I \rightarrow \mathbb{K}^{n+1}, \quad \psi(x) := (x, \phi(x)), \) is a solution to (2.53b).

(b) If \( \psi : I \rightarrow \mathbb{K}^{n+1}, \quad x \mapsto (\psi_0(x), \psi_1(x), \ldots, \psi_n(x)), \) is a solution to (2.53b) with the property
\[ \exists x_0 \in I \quad \psi_0(x_0) = x_0, \] (2.55)
then \( \phi : I \rightarrow \mathbb{K}^n, \quad \phi(x) := (\psi_1(x), \ldots, \psi_n(x)), \) is a solution to (2.53a).

Proof. (a): If \( \phi : I \rightarrow \mathbb{K}^n \) is a solution to (2.53a) and \( \psi : I \rightarrow \mathbb{K}^{n+1}, \quad \psi(x) := (x, \phi(x)), \) then
\[ \forall x \in I \quad \psi'(x) = (1, \phi'(x)) = (1, f(x, \phi(x))) = g(x, \phi(x)) = g(\psi(x)), \]
showing \( \psi \) is a solution to (2.53b).

(b): If \( \psi : I \rightarrow \mathbb{K}^{n+1} \) is a solution to (2.53b) with the property (2.55) and \( \phi : I \rightarrow \mathbb{K}^n, \quad \phi(x) := (\psi_1(x), \ldots, \psi_n(x)), \) then (2.55) implies \( \psi_0(x) = x \) for each \( x \in I \) and, thus,
\[ \forall x \in I \quad \phi'(x) = (\psi_1'(x), \ldots, \psi_n'(x)) = f(x, \psi_1(x), \ldots, \psi_n(x)) = f(x, \phi(x)), \]
showing \( \phi \) is a solution to (2.53a). \( \square \)

The following Prop. 2.37 shows that RK methods are consistent with the equivalence of Th. 2.35 (such that designing RK methods for autonomous problems does not really constitute a restriction), provided they satisfy both the consistency and the node condition. The proof of Prop. 2.37 makes use of the following lemma:

Lemma 2.36. Let \( s, n \in \mathbb{N}, \quad G \subseteq \mathbb{R} \times \mathbb{K}^n, \quad f : G \rightarrow \mathbb{K}^n, \quad c_j \in \mathbb{R}, \quad a_{jl} \in \mathbb{K} \) for each \( j, l \in \{1, \ldots, s\}. \) Let \( g : G \rightarrow \mathbb{K}^{n+1} \) be defined as in (2.54). Let \( (x, y, h) \in \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+ \) and consider vectors \( k_{f,1}, \ldots, k_{f,s} \in \mathbb{K}^n, \quad k_{g,1}, \ldots, k_{g,s} \in \mathbb{K}^{n+1}, \quad \tilde{k}_{g,1}, \ldots, \tilde{k}_{g,s} \in \mathbb{K}^n, \) where \( k_{g,j} = (k_{g,j,0}, \tilde{k}_{g,j}) \) for each \( j \in \{1, \ldots, s\}. \) Moreover, consider the system (2.36a), which we rewrite here as
\[ \forall j \in \{1, \ldots, s\} \quad k_{f,j} = f \left( x + c_j h, y + h \sum_{l=1}^{s} a_{jl} k_{f,l} \right) \] (2.56a)
and
\[ k_{g,j} = g \left( x, y + h \sum_{l=1}^{s} a_{jl}k_{g,l} \right) = \left( 1, f \left( x + h \sum_{l=1}^{s} a_{jl}k_{g,l,0}, y + h \sum_{l=1}^{s} a_{jl}\tilde{k}_{g,l} \right) \right). \] (2.56b)

If \( k_{g,1}, \ldots, k_{g,s} \) satisfy (2.56b) (which, in particular, is supposed to mean, for each \( j \in \{1, \ldots, s\} \), that the used argument of \( f \) is in \( G \)), then \( k_{g,1,0} = \cdots = k_{g,s,0} = 1 \). If, in addition, the node condition (2.35) holds (i.e. \( c_{j} = \sum_{l=1}^{s} a_{jl} \) for each \( j \in \{1, \ldots, s\} \)), then \( k_{f,1} := \tilde{k}_{g,1}, \ldots, k_{f,s} := \tilde{k}_{g,s} \) satisfy (2.56a) (i.e. (2.36a)).

**Proof.** The first component of (2.56b) immediately yields \( k_{g,1,0} = \cdots = k_{g,s,0} = 1 \). Using this and the node condition in (2.56b), we see that the remaining components of (2.56b) become (2.56a) with \( k_{f,j} \) replaced by \( \tilde{k}_{g,j} \). \(\square\)

**Proposition 2.37.** Let \( s \in \mathbb{N} \) and consider an \( s \)-stage RK method according to Def. 2.24 that has unique local solutions as defined in Def. 2.26(b) (sufficient conditions are that the RK method is explicit or that \( f \) satisfies the conditions of Th. 2.31). Moreover, let the RK method be in standard form (also defined in Def. 2.26(b)). Let \( \varphi_{f} \) be the defining function corresponding to \( f \) and \( \varphi_{g} \) the defining function corresponding to \( g \) of (2.54) (as (2.53b) is autonomous, \( g \) and, thus, \( \varphi_{g} \), do not depend on \( x \)). Let \( (\xi, \eta) \in G, b > \xi, \) and let \( (x_{0}, \ldots, x_{N}) \) be a partition of \([\xi, b]\). Assume
\[ y_{0} = \eta, \forall k \in \{0, 1, \ldots, N-1\} \quad y_{k+1} = y_{k} + h_{k} \varphi_{f}(x_{k}, y_{k}, h_{k}), \quad h_{k} := x_{k+1} - x_{k}, \] (2.57a)

and
\[ z_{0} = (\xi, \eta), \forall k \in \{0, 1, \ldots, N-1\} \quad z_{k+1} = z_{k} + h_{k} \varphi_{g}(z_{k}, h_{k}). \] (2.57b)

Then
\[ \forall k \in \{0, 1, \ldots, N\} \quad z_{k} = (x_{k}, y_{k}) \] (2.57c)

if the RK method satisfies both the consistency condition (2.34) and the node condition (2.35).

**Proof.** For each \( j \in \{1, \ldots, s\} \), we denote by \( k_{f,j} \in \mathbb{K}^{n} \) (resp. by \( k_{g,j} = (1, \tilde{k}_{g,j}) \in \mathbb{K}^{n+1} \), \( \tilde{k}_{g,j} \in \mathbb{K}^{n} \)) the auxiliary vectors of the RK method corresponding to \( \varphi_{f} \) (resp. to \( \varphi_{g} \)). Suppose, the method satisfies (2.34) and (2.35). We show (2.57c) via induction on \( k \). First, \((x_{0}, y_{0}) = (\xi, \eta) = z_{0}\) holds by the definition of \( y_{0} \) and \( z_{0} \), respectively. Now fix \( k \in \{0, \ldots, N-1\} \). As the \( k_{g,j} \) must satisfy the condition (2.32b) (with \( f \) replaced by
Thus, the \( k_{g,j}(z_k, h_k) \) satisfy (2.56b). As we also assume the node condition (2.35), Lem. 2.36 implies
\[
\forall \ j \in \{1, \ldots, s\} \quad k_{g,j}(z_k, h_k) = \left( 1, f(x + h_k \sum_{l=1}^{s} a_{jl} y_k + h_k \sum_{l=1}^{s} a_{jl} \tilde{k}_{g,l}(z_k, h_k)) \right). 
\] (2.58)

As the \( k_{f,j}(x_k, y_k, h_k) \) also satisfy (2.58) (which is merely condition (2.32b)) and, since the \( k_{f,j}(x_k, y_k, h_k) \) are the unique solution to (2.58) in the considered domain (and also using that the RK method is assumed to be in standard form), we obtain
\[
\forall \ j \in \{1, \ldots, s\} \quad \tilde{k}_{g,l}(z_k, h_k) = f(x + c_j h_k, y_k + h_k \sum_{l=1}^{s} a_{jl} \tilde{k}_{g,l}(z_k, h_k)). 
\] (2.59)

Using (2.59), we are now in a position to complete the induction on \( k \): We compute
\[
\begin{align*}
    z_{k+1} & \overset{\text{ind.hyp.}}{=} (x_k, y_k) + h_k \sum_{j=1}^{s} b_j k_{g,j}(z_k, h_k) \\
    & \overset{(2.59)}{=} (x_k, y_k) + h_k \sum_{j=1}^{s} b_j \left( 1, k_{f,j}(x_k, y_k, h_k) \right) \\
    & = \left( x_k + h_k \sum_{j=1}^{s} b_j y_k + h_k \sum_{j=1}^{s} b_j k_{f,j}(x_k, y_k, h_k) \right) \\
    & \overset{(2.34)}{=} \left( x_{k+1}, y_k + h_k \sum_{j=1}^{s} b_j k_{f,j}(x_k, y_k, h_k) \right) \\
    & = (x_{k+1}, y_{k+1}), 
\end{align*}
\] (2.60)
as desired.

**Remark 2.38.** In Prop. 2.37, the consistency condition (2.34) is clearly also necessary for (2.57c) to hold: Indeed, otherwise, the computation (2.60) shows for \( k = 0 \) that the first component of \( z_1 \) does not equal \( x_1 \) if (2.34) is not satisfied. In general, the node condition (2.35) will also be necessary for (2.57c) to hold. However, for certain \( f \) and/or certain discretizations, (2.58) (and, hence, (2.59) and (2.57c)) can be true, even if the node condition fails.
2.3.4 Stability Functions

Example 2.39. Let \( \lambda \in \mathbb{K} \), \( f : \mathbb{R} \times \mathbb{K} \rightarrow \mathbb{K} \), \( f(x,y) := \lambda y \), and consider the one-dimensional autonomous linear initial value problem

\[
y' = f(x,y) = \lambda y, \quad y(\xi) = \eta, \quad (\xi, \eta) \in \mathbb{R} \times \mathbb{K}.
\]

Now let \( s \in \mathbb{N} \) and consider the \( s \)-stage RK method given by weights \( b_1, \ldots, b_s \in \mathbb{R} \) and RK matrix \( A = (a_{jl}) \in \mathcal{M}(s, \mathbb{K}) \) (as \( f \) does not depend on \( x \), the following considerations are independent of the nodes \( c_j \) of the RK method). In \( u \)-form, we have, for the defining function, \( \varphi : D_\varphi \rightarrow \mathbb{K} \),

\[
\varphi(x,y,h) = \sum_{j=1}^{s} b_j \lambda u_j(x,y,h),
\]

where the \( u_1(x,y,h), \ldots, u_s(x,y,h) \in \mathbb{K} \) satisfy the linear system

\[
\forall j \in \{1, \ldots, s\} \quad u_j(x,y,h) = y + h \sum_{l=1}^{s} a_{jl} \lambda u_l(x,y,h) = y + h\lambda \sum_{l=1}^{s} a_{jl} u_l(x,y,h).
\]

(2.61)

For fixed \( (x,y,h) \in \mathbb{R} \times \mathbb{K} \times \mathbb{R} \), we set \( u := u(x,y,h) \in \mathbb{K}^s \), such that (2.61) now reads

\[
\forall j \in \{1, \ldots, s\} \quad u_j = y + h\lambda \sum_{l=1}^{s} a_{jl} u_l
\]

or, in matrix form,

\[
u = y1 + h\lambda Au \iff (\text{Id} - h\lambda A)u = y1,
\]

where \( 1 := (1, \ldots, 1)^t \in \mathbb{K}^s \). As \( \text{Id} \) is invertible and we know the set of invertible matrices, i.e. the general linear group \( \text{GL}(s, \mathbb{K}) = \text{det}^{-1}(\mathbb{K} \setminus \{0\}) \), to be an open subset of \( \mathcal{M}(s, \mathbb{K}) \), there exists \( \epsilon \in \mathbb{R}^+ \) (depending only on \( \lambda \) and \( A \)) such that \( (\text{Id} - h\lambda A) \) is invertible for each \( h \in ]-\epsilon, \epsilon[ \). Thus, if \( |h| < \epsilon \), then

\[
u = \frac{1}{\text{Id} - h\lambda A}y := (\text{Id} - h\lambda A)^{-1}y.
\]

Thus, letting \( b := (b_1, \ldots, b_s)^t \), we obtain

\[
\varphi : \mathbb{R} \times \mathbb{K} \times [0,\epsilon[ \rightarrow \mathbb{K}, \quad \varphi(x,y,h) = \lambda b^t u = \lambda b^t (\text{Id} - h\lambda A)^{-1}y,
\]

and the resulting recursion is (cf. Rem. 2.27(a)) \( y_0 = \eta \),

\[
y_{k+1} = y_k + h_k \lambda b^t (\text{Id} - h_k \lambda A)^{-1}y_k = R(h_k \lambda) y_k,
\]

where

\[
R : D_R \rightarrow \mathbb{K}, \quad R(z) := 1 + zb^t (\text{Id} - z A)^{-1}1, \quad D_R \subseteq \mathbb{K},
\]

(2.62)

defined on a superset \( D_R \) of \( \{z \in \mathbb{K} : |z| < \epsilon\} \). In the following Def. and Rem. 2.40, we will see that the so-called stability function \( R \) is always a rational function.
Definition and Remark 2.40. Let \( s \in \mathbb{N}, b \in \mathbb{R}^s, A \in \mathcal{M}(s, \mathbb{K}) \).

(a) The function \( R \) defined in (2.62) above, is called the stability function of the RK method with RK matrix \( A \) and weights vector \( b \). Recall from Linear Algebra that the determinant \( \det : \mathcal{M}(s, \mathbb{K}) \rightarrow \mathbb{K}, B \mapsto \det B \), is a polynomial of degree \( s \) with real coefficients in the entries \( b_{jl} \) of \( B \) (the coefficients are actually \( \pm 1 \)). Moreover, recall from Linear Algebra that, for invertible \( B \in \text{GL}(s, \mathbb{K}) \),
\[
B^{-1} = (\det B)^{-1} B^\#,
\]
where, for \( s > 1 \), the entries of \( B^\# \) are the determinants of the \((s - 1) \times (s - 1)\) submatrices of \( B \). Thus, each entry of \( B^{-1} \) is a rational function \( R_{jl} \) of the entries \( b_{jl} \) of \( B \), and \( R_{jl} = P_{jl} / \det \), where, for \( s > 1 \), the entries of \( B^{-1} \) are rational functions \( R_{jl} \) of the entries \( b_{jl} \) of \( B \), and \( R_{jl} = P_{jl} / \det \).

Letting
\[
\forall \, d \in \mathbb{N}_0 \quad \mathcal{P}_d(\mathbb{K}) := \left\{ (P : \mathbb{K} \rightarrow \mathbb{K}) : \sum_{j=0}^d \lambda_j z^j ; \quad \lambda_0, \ldots, \lambda_d \in \mathbb{K} \right\}
\]
and coming back to (2.62), we see that \( R \) is a rational function, where \( R = \hat{P} / \hat{Q} \) with polynomials \( \hat{P}, \hat{Q} \in \mathcal{P}_s(\mathbb{R}) \). As \( R(0) = 1 \), one has \( \hat{P}(0) = \hat{Q}(0) \neq 0 \), and, dividing numerator and denominator of \( R = \hat{P} / \hat{Q} \) by \( \hat{P}(0) \) as well as by any common prime factors, we obtain
\[
R = \frac{P}{Q}, \quad P, Q \in \mathcal{P}_s(\mathbb{R}), \quad P(0) = Q(0) = 1, \quad P, Q \text{ mutually prime}, \quad (2.64)
\]
and \( P, Q \) are uniquely determined by (2.64). In particular, \( \mathcal{D}_R = \{ z \in \mathbb{K} : Q(z) \neq 0 \} \).

(b) Note that, if the RK method is explicit, i.e. if \( A \) is strictly lower triangular, then, in (2.64), \( Q \equiv 1 \) and \( R = P \) is a polynomial with real coefficients, \( 0 \leq \deg P \leq s \): This can be seen directly from (2.61), using an induction or, alternatively, by observing that, for each \( z \in \mathbb{K} \), \( \det(\text{Id} - zA) = \det \text{Id} = 1 \) in this case.

Remark 2.41. Another useful fact to recall from Linear Algebra is that, given a polynomial \( P : \mathbb{K} \rightarrow \mathbb{K} \), we can substitute numbers with matrices, yielding the following matrix mapping (still denoted by \( P \) for the simplicity of notation):
\[
P : \mathcal{M}(s, \mathbb{K}) \rightarrow \mathcal{M}(s, \mathbb{K}), \quad B \mapsto P(B)
\]
(where, as usual, for \( P \equiv \lambda \in \mathbb{K} \), \( P(B) := \lambda \text{Id} \)). Moreover, given polynomials \( P, Q : \mathbb{K} \rightarrow \mathbb{K} \), the rational function \( P/Q \) also yields a matrix mapping, namely
\[
(P/Q) : \{ B \in \mathcal{M}(s, \mathbb{K}) : \det(Q(B)) \neq 0 \} \rightarrow \mathcal{M}(s, \mathbb{K}), \quad B \mapsto P(B) (Q(B))^{-1}.
\]
This mapping is independent of the representation of the rational function in the sense that, if \( P/Q = \tilde{P} / \tilde{Q}, \det(Q(B)) \neq 0 \), and \( \det(\tilde{Q}(B)) \neq 0 \), then \( (P/Q)(B) = (\tilde{P} / \tilde{Q})(B) \).
Example 2.42. Somewhat surprisingly, everything of Ex. 2.39 still works analogously for higher-dimensional autonomous linear initial value problems, if one makes use of the matrix mappings of Rem. 2.41 above. Let \( n \in \mathbb{N}, B \in \mathcal{M}(n, \mathbb{K}) \). Let \( f : \mathbb{R} \times \mathbb{K}^n \to \mathbb{K}^n \), \( f(x, y) := B y \), and consider the linear initial value problem

\[
y' = f(x, y) = B y, \quad y(\xi) = \eta, \quad (\xi, \eta) \in \mathbb{R} \times \mathbb{K}^n.
\]

As \( B \) and \( f \) do not depend on \( x \), the above linear initial value problem has so-called constant coefficients (cf. [Phi16c, Sec. 4.6.2]). As in Ex. 2.39, let \( s \in \mathbb{N} \) and consider the \( s \)-stage RK method given by weights \( b_1, \ldots, b_s \in \mathbb{R} \) and RK matrix \( A = (a_{jl}) \in \mathcal{M}(s, \mathbb{K}) \) (as \( f \) does not depend on \( x \), the following considerations are independent of the nodes \( c_j \) of the RK method). In \( u \)-form, we have, for the defining function,

\[
\varphi : \mathcal{D}_\varphi \to \mathbb{K}^n, \quad \varphi(x, y, h) = \sum_{j=1}^{s} b_j B u_j(x, y, h),
\]

where the \( u_1(x, y, h), \ldots, u_s(x, y, h) \in \mathbb{K}^n \) satisfy the linear system

\[
\forall j \in \{1, \ldots, s\} \quad u_j(x, y, h) = y + h \sum_{l=1}^{s} a_{jl} B u_l(x, y, h) = y + h B \sum_{l=1}^{s} a_{jl} u_l(x, y, h). \tag{2.65}
\]

For fixed \((x, y, h) \in \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}\), if we define the larger vectors \( u := (u_1^t, \ldots, u_s^t)^t \in \mathbb{K}^{ns} \), \( \vec{y} := (y^t, \ldots, y^t)^t \in \mathbb{K}^{ns} \), and the block matrix

\[
h \tilde{A} := h \begin{pmatrix} a_{11} B & \cdots & a_{1s} B \\ \vdots & & \vdots \\ a_{s1} B & \cdots & a_{ss} B \end{pmatrix} = \begin{pmatrix} a_{11} h B & \cdots & a_{1s} h B \\ \vdots & & \vdots \\ a_{s1} h B & \cdots & a_{ss} h B \end{pmatrix} \in \mathcal{M}(ns, \mathbb{K}),
\]

then, using blockwise matrix multiplication, (2.65) can be written in matrix form as

\[
u = \vec{y} + h \tilde{A} u \iff (\text{Id} - h \tilde{A}) u = \vec{y}.
\]

As in Ex. 2.39, there exists \( \epsilon \in \mathbb{R}^+ \) (now depending only on \( B \) and \( A \)) such that \((\text{Id} - h \tilde{A})\) is invertible for each \( h \in ]-\epsilon, \epsilon[\). Thus, if \( |h| < \epsilon \), then

\[
u = \left( \text{Id} - h \tilde{A} \right)^{-1} \vec{y} = \left( \text{Id} - h \tilde{A} \right)^{-1} \begin{pmatrix} \text{Id}_n \\ \vdots \\ \text{Id}_n \end{pmatrix} y,
\]

where the \( j \)-th row block of this equation yields \( u_j \in \mathbb{K}^n \). Moreover, due to blockwise matrix multiplication, one obtains \((\text{Id} - h \tilde{A})^{-1}\) from \((\text{Id} - z A)^{-1} \in \mathcal{M}(s, \mathbb{K})\) by replacing each \( z \in \mathbb{K} \) by \( h B \in \mathcal{M}(n, \mathbb{K}) \) and each 1 in \( \text{Id} \) by \( \text{Id}_n \). Thus, we obtain \( \varphi : \mathbb{R} \times \mathbb{K}^n \times [0, \epsilon[ \to \mathbb{K}^n \),

\[
\varphi(x, y, h) = \sum_{j=1}^{s} b_j B u_j(x, y, h) = B (b_1 \text{Id}_n, \ldots, b_s \text{Id}_n) \left( \text{Id} - h \tilde{A} \right)^{-1} \begin{pmatrix} \text{Id}_n \\ \vdots \\ \text{Id}_n \end{pmatrix} y
\]
and the resulting recursion is $y_0 = \eta,$

$$y_{k+1} = y_k + h_k \varphi(x_k, y_k, h_k) = R(h_k B)(y_k),$$

where $R$ is the stability function as defined in (2.62), but now interpreted in the sense of Rem. 2.41, i.e. as a function, mapping a subset $\mathcal{D}_R$ of $\mathcal{M}(s, \mathbb{K})$ into $\mathcal{M}(s, \mathbb{K}).$ Using the representation $R = P/Q$ of (2.64), we obtain $\mathcal{D}_R = \{M \in \mathcal{M}(s, \mathbb{K}) : \det(Q(M)) \neq 0\},$ where $Q(0) = \text{Id}$ implies $\mathcal{D}_R$ to contain some open neighborhood of 0. According to the choice of $\epsilon > 0$ above, we also have $h_k B \in \mathcal{D}_R$ for each $h_k \in [-\epsilon, \epsilon[.$

**Example 2.43.** We compute the stability function $R$ of Def. and Rem. 2.40 for a few concrete RK methods:

(a) For the explicit Euler method, we have $s = 1,$ $b = (1),$ $A = 0$ (cf. Ex. 2.28(a)). Thus,

$$R : \mathbb{K} \longrightarrow \mathbb{K}, \quad R(z) = 1 + z.$$ 

(b) For the implicit Euler method, we have $s = 1,$ $b = (1),$ $A = \text{Id} = (1)$ (cf. Ex. 2.28(c)). Thus,

$$R : \mathbb{K} \setminus \{1\} \longrightarrow \mathbb{K}, \quad R(z) = 1 + \frac{z}{1 - z} = \frac{1}{1 - z}.$$

(c) For the classical explicit RK method, we have $s = 4,$ $b^t = (\frac{1}{6} \frac{1}{3} \frac{1}{3} \frac{1}{6}),$

$$A = \begin{pmatrix}
0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & 0 & 0 \\
0 & \frac{1}{2} & 0 & 0 \\
0 & 0 & 1 & 0
\end{pmatrix},$$

(cf. Ex. 2.28(b)). Thus, for each $z \in \mathbb{K},$

$$\text{Id} - z A = \begin{pmatrix}
1 & 0 & 0 & 0 \\
-\frac{z}{2} & 1 & 0 & 0 \\
0 & -\frac{z}{2} & 1 & 0 \\
0 & 0 & -z & 1
\end{pmatrix}, \quad (\text{Id} - z A)^{-1} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
\frac{z}{2} & 1 & 0 & 0 \\
\frac{z^2}{4} & \frac{z}{2} & 1 & 0 \\
\frac{z^3}{4} & \frac{z^2}{2} & z & 1
\end{pmatrix}$$

and $R : \mathbb{K} \longrightarrow \mathbb{K},$

$$R(z) = 1 + z \begin{pmatrix}
\frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{pmatrix} (\text{Id} - z A)^{-1} \begin{pmatrix}
1 \\
1 \\
1 \\
1
\end{pmatrix}$$

$$= 1 + \frac{z}{6} + \frac{z^2}{6} + \frac{z}{3} + \frac{z^2}{12} + \frac{z^3}{6} + \frac{z}{3} + \frac{z^4}{24} + \frac{z^3}{12} + \frac{z^2}{6} + \frac{z}{6}$$

$$= 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24}.$$
As an example of a 2-stage implicit RK method, consider the so-called implicit trapezoidal method, which has Butcher tableau

\[
\begin{array}{c|ccc}
0 & 0 & 0 \\
1 & \frac{1}{2} & \frac{1}{2} \\
\hline & \frac{1}{2} & \frac{1}{2}
\end{array}
\]

The name comes from the fact that the recursion can be written as

\[
y_{k+1} = y_k + h_k \sum_{j=1}^{2} b_j f(x_k + c_j h_k, u_j) = y_k + \frac{h_k}{2} \left( f(x_k, u_1) + f(x_k + h_k, u_2) \right),
\]

where

\[
u_1 = y_k + h_k \sum_{l=1}^{2} a_{1l} f(x_k + c_l h_k, u_l) = y_k,
\]

\[
u_2 = y_k + h_k \sum_{l=1}^{2} a_{2l} f(x_k + c_l h_k, u_l) = y_k + \frac{h_k}{2} \left( f(x_k, u_1) + f(x_k + h_k, u_2) \right) = y_{k+1},
\]

i.e.

\[
y_{k+1} = y_k + \frac{h_k}{2} \left( f(x_k, y_k) + f(x_{k+1}, y_{k+1}) \right)
\]

(if \( f \) is \( \mathbb{R} \)-valued, then the term added to \( y_k \) is the area of the trapezoid with vertices \((x_k, 0), (x_{k+1}, 0), (x_k, f(x_k, y_k)), (x_{k+1}, f(x_{k+1}, y_{k+1}))\)). To compute the stability function, we note that, for each \( z \in \mathbb{K} \setminus \{2\}, \)

\[
\text{Id} - zA = \begin{pmatrix} 1 & 0 \\ -\frac{1}{2} & 1 - \frac{1}{2} \end{pmatrix}, \quad (\text{Id} - zA)^{-1} = \begin{pmatrix} 1 & 0 \\ \frac{1}{2z} & -\frac{2}{z} \end{pmatrix}
\]

and \( R : \mathbb{K} \setminus \{2\} \rightarrow \mathbb{K}, \)

\[
R(z) = 1 + z \left( \frac{1}{2} \right) \left( \text{Id} - zA \right)^{-1} \left( \frac{1}{2} \right) = 1 + \frac{z}{2} + \frac{z^2}{2 - z} + \frac{z}{2 - z}
= \frac{4 - 2z + 2z - z^2 + z^2 + 2z}{2(2 - z)} = \frac{2 + z}{2 - z}.
\]

### 2.3.5 Higher Orders of Consistency and Convergence

The following Prop. 2.45 shows that, to obtain consistency of order \( p \), one needs at least \( p \) stages for an explicit RK method, and at least \( \lceil p/2 \rceil \) stages for an implicit RK method. The proof makes use of the following lemma.

**Lemma 2.44.** Let \( p \in \mathbb{N}_0 \), let \( P, Q : \mathbb{K} \rightarrow \mathbb{K} \) be polynomials with real coefficients, neither \( P \) nor \( Q \) the zero polynomial, and consider the rational function \( R := P/Q \). If there exist \( \epsilon > 0 \) and \( C \in \mathbb{R}_0^+ \) such that

\[
\forall h \in [0, \epsilon] \setminus Q^{-1}(0) \quad |R(h) - e^h| \leq C h^{p+1},
\]
If the method is explicit, then
\[ p \leq \deg P + \deg Q \]
(in other words, if the approximation of the exponential function by the rational function \( R \) is “consistent of order \( p \), then \( p \leq \deg P + \deg Q \)).

Proof. Let \( k := \deg P \), \( j := \deg Q \). Seeking a contradiction, assume (2.66) holds with \( p > k + j \). Set \( \mathcal{D} := ]0, \epsilon[ \setminus Q^{-1}\{0\} \). Then
\[
\forall h \in \mathcal{D} \quad \left| \frac{P(h) - Q(h) e^h}{h^{p+1}} \right| \leq |Q(h)| C. \tag{2.67}
\]
We show, by induction on \( k \in \mathbb{N}_0 \), that (2.67) must be false, which is the desired contradiction. More precisely, what we show by induction on \( k \) is
\[
\lim_{x \to 0} \left| \frac{P(x) - Q(x) e^x}{x^{p+1}} \right| = \infty.
\]

For \( k = 0 \), \( P \) is constant. Thus, applying l'Hôpital’s rule [Phil16a, Th. 9.26] yields
\[
\lim_{x \to 0} \left| \frac{P(x) - Q(x) e^x}{x^{p+1}} \right| = \lim_{x \to 0} \left| \frac{(Q'(x) + Q(x)) e^x}{(p + 1) x^p} \right| = 1 \cdot \lim_{x \to 0} \left| \frac{(Q'(x) + Q(x))}{(p + 1) x^p} \right| \equiv Q, \quad p > \deg Q \infty.
\]

Now consider \( k > 0 \). This time, we apply l'Hôpital’s rule together with the induction hypothesis to obtain
\[
\lim_{x \to 0} \left| \frac{P(x) - Q(x) e^x}{x^{p+1}} \right| = \lim_{x \to 0} \left| \frac{P'(x) - (Q'(x) + Q(x)) e^x}{(p + 1) x^p} \right| \overset{\text{ind.hyp.}}{=} \infty,
\]
where the induction hypothesis applies, as \( \deg P' = k - 1 \), \( \deg(Q' + Q) \leq j \) and \( p > k - 1 + j \). Thus, the induction and the proof of the lemma are complete. \( \blacksquare \)

**Proposition 2.45.** Let \( s \in \mathbb{N} \) and consider an \( s \)-stage RK method according to Def. 2.24. Consider the identity \( f : \mathbb{K}^n \to \mathbb{K}^n \), and the corresponding initial value problem
\[
y' = y, \tag{2.68a}
y(0) = \eta \in \mathbb{K}^n \setminus \{0\}. \tag{2.68b}
\]
Assume the RK method to be consistent of order \( p \in \mathbb{N} \) in the sense of Def. 2.12(c).

(a) If the method is implicit, then \( p \leq 2s \).

(b) If the method is explicit, then \( p \leq s \).

Proof. Clearly, (2.68) is solved by \( \phi : \mathbb{R} \to \mathbb{K}^n \), \( \phi(x) = e^x \eta \). According to Ex. 2.42 with \( B := \text{Id} \), we obtain for the local truncation error
\[
\lambda(0, h) = \phi(0) + h \varphi(0, \phi(0), h) - \phi(h) = R(h \text{Id})(\eta) - \eta e^h
\]
\[
= (R(h) - e^h) \eta = \left( \frac{P(h)}{Q(h)} - e^h \right) \eta = \frac{P(h) - Q(h) e^h}{Q(h)} \eta.
\]
where $R$ is the (rational) stability function of (2.62) and $P, Q$ are the polynomials of degree at most $s$ from (2.64). Now Def. 2.12(c) implies

$$\exists \epsilon, C \in \mathbb{R}^+ \quad \forall h \in [0, \epsilon], \quad \|\lambda(0, h)\| \leq C h^{p+1}.$$ 

Thus, for $\eta \neq 0$, Lem. 2.44 implies $p \leq \deg P + \deg Q \leq s + s = 2s$, proving (a). If the method is explicit, then we know $Q \equiv 1$, i.e. $\deg Q = 0$, from Def. and Rem. 2.40. Thus, in this case, for $\eta \neq 0$, Lem. 2.44 implies $p \leq \deg P + \deg Q \leq s$, proving (b). 

**Remark 2.46.** (a) As defined in Def. 2.12(c), the order of consistency of a method will, in general, depend on the right-hand side $f$ of the ODE. So it might happen that an $s$-stage RK method is, by accident, consistent of order $p > 2s$ ($p > s$ in the explicit case) for some particular $f$. However, as a consequence of Prop. 2.45, it can never be consistent of order $p > 2s$ ($p > s$ in the explicit case) for any set of functions containing the identity.

(b) For implicit RK methods, the bound $p \leq 2s$ of Prop. 2.45(a) is optimal: There exist implicit RK methods, where $p = 2s$ (see Ex. 2.94 below). For explicit RK methods, the bound $p \leq s$ of Prop. 2.45(b) is, for $p > 4$ not optimal: In general, finding the optimal bounds in this case is difficult, cf. final paragraphs of [DB08, Sec. 4.2.3].

To obtain higher-order RK methods, one has to determine the parameters of the RK method such that, when using Taylor expansions of the exact solution $\phi$ and of the defining function $\varphi$ to calculate the local truncation error, sufficiently many low-order terms cancel out (cf. proof of Th. 2.20(a)). The difficulty lies in the fact that these Taylor expansions quickly become rather complicated. To help with the accounting of the relevant terms, Butcher developed a graph-theoretic method, which we present in the rest of this section. The key objects that are employed are so-called (unlabeled, finite) rooted trees. Here, we will directly develop the theory of such trees as far as necessary, without requiring any prior knowledge of graph theory. In regular graph theory, a (labeled) tree is a connected graph without any cycles (a graph without cycles, not necessarily connected, is called a forest). One can then show that each of our unlabeled trees represents an entire class of such labeled trees, but this is of no consequence to our considerations.

The central observation will then be that both the evaluation of derivatives and the construction of certain vectors in the Taylor expansion of the defining function can be related to the mentioned unlabeled, finite, rooted trees. We start with the definition of a multiset, followed by a recursive definition of the aforementioned trees and some related notions.

**Definition and Remark 2.47.** Let $M$ be a set.

(a) Each function $\mathcal{M} : M \rightarrow \mathbb{N}_0$ is called an $M$-multiset or an unordered tuple with entries from $M$. The name multiset comes from the fact that each multiset in
\( M \) can be interpreted as a set with elements from \( M \), possibly containing certain elements multiple times. For example, consider the following multiset \( M \) in \( \{1, 2, 3\} \): \( M(1) := 2 \), \( M(2) := 0 \), \( M(3) := 5 \). It can be written in the form \( M = [1, 1, 3, 3, 3, 3, 3, 3] \), where, here and in the following, we use brackets \([ \ ]\) instead of braces \( \{ \} \) to distinguish multisets from sets. Noting that we can also write the same multiset as \( M = [3, 3, 1, 3, 3, 3] \) explains the alternative name unordered tuple for a multiset. Also note that, for each set \( M \), we can identify the empty set \( \emptyset \) with the constant \( M \)-multiset \( M \equiv 0 \) (i.e. \([ ] = \emptyset \)) and the set \( M \) itself with the constant \( M \)-multiset \( M \equiv 1 \).

(b) If \( M \) is an \( M \)-multiset, then we define its order or cardinality \(|M|\) by setting

\[
|M| := \sum_{m \in M} M(m) \in \mathbb{N}_0 \cup \{\infty\}.
\]

We call \( M \) an unordered \( n \)-tuple if, and only if, \(|M| = n \in \mathbb{N} \); we call \( M \) finite if, and only if, \(|M| < \infty \).

(c) According to (a), the set of \( M \)-multisets is the same as \( \mathcal{F}(M, \mathbb{N}_0) \), the set of functions from \( M \) into \( \mathbb{N}_0 \). We now define the combinatorial function

\[
\delta : \left\{ \mathcal{M} \in \mathcal{F}(M, \mathbb{N}_0) : |\mathcal{M}| < \infty \right\} \rightarrow \mathbb{N},
\]

\[
\delta(\mathcal{M}) := \# \left\{ \left( f : \{1, \ldots, |\mathcal{M}|\} \rightarrow M : \forall m \in M \# f^{-1}(\{m\}) = M(m) \right) \right\}.
\]

Thus, if \( \mathcal{M} \) is an unordered \( n \)-tuple, then \( \delta(\mathcal{M}) \) is defined to be the number of different ordered \( n \)-tuples with precisely the same entries as \( \mathcal{M} \). Note that \( \delta(\emptyset) = 1 \), since \(|\emptyset| = 0 \) and the empty function \( f = \emptyset \) is the only function \( f : \emptyset \rightarrow M \). Also note \( \delta(\mathcal{M}) = 1 \) if there exists \( m \in M \) with \( M(m) = |\mathcal{M}| \) (i.e. if \( \mathcal{M} = [m, \ldots, m] \)). If \( N \subseteq M \) is a finite subset of \( M \), \( k := \#N \in \mathbb{N}_0 \), and \( \mathcal{M} = \chi_N \) is the characteristic function of \( N \) (i.e. \(|\mathcal{M}| = k \) and all entries of \( \mathcal{M} \) are distinct), then \( \delta(\mathcal{M}) = k! \).

Some further examples are

\[
\delta[1, 2, 2] = 3, \quad \delta[1, 1, 2, 2] = \binom{5}{2} = \frac{5!}{2! \cdot 3!} = 10, \quad \delta[1, 1, 2, 3] = 2 \cdot \binom{4}{2} = 12.
\]

**Definition and Remark 2.48.** In the following, we define finite unlabeled rooted trees. However, as these are the only kinds of trees we will consider in this class, we will merely call them trees.

(a) We define, recursively, for each \( d \in \mathbb{N} \), the set \( T_d \) of trees of depth \( d \), where we also set \( T_{\leq d} := T_1 \cup \cdots \cup T_d \). We define \( T_1 := \{\emptyset\} \), i.e. the empty set \( \emptyset \) is the only tree of depth 1. Now let \( d \in \mathbb{N} \), and assume \( T_1, \ldots, T_d \) have already been defined. Define

\[
T_{d+1} := \left\{ (\mathcal{M} : T_{\leq d} \rightarrow \mathbb{N}_0) : \exists T \in T_d M(T) \geq 1, |\mathcal{M}| < \infty \right\}.
\]
i.e. the set of trees of depth $d + 1$ consists precisely of all finite $\mathcal{T}_{\leq d}$-multisets, containing at least one tree of depth $d$. Moreover, we define the set of all trees $\mathcal{T} := \bigcup_{d \in \mathbb{N}} \mathcal{T}_d$. Even though it is not necessary for the rigorous logical arguments that we conduct in the following, it can be useful to represent and visualize trees as graphs in the following way: In regular graph theory, a graph is a pair $G = (V, E)$ consisting of a set of vertices $V$ and a set of edges $E \subseteq \binom{V}{2}$, where $\binom{V}{2}$ denotes the set of all subsets of $V$ that have precisely two elements. We represent the tree $\emptyset$ by a single vertex (the root) and no edge:

\[
\emptyset = .
\]

Writing “=” in the previous formula and below in similar situations is somewhat of an abuse of notation, as it does not actually constitute a set-theoretic equality. It is rather meant in the sense that we use the drawn graph as a new notation for the unlabeled tree on the left that it represents (rather than for the graph consisting of vertices and edges that it also represents). If $\mathcal{M} : \mathcal{T}_{\leq d} \to \mathbb{N}_0$ is a tree of depth $d + 1$, then we obtain the graph representing $\mathcal{M}$ by adding a new root vertex and connecting this new root with each of the previous roots via a new edge, where the root is always at the bottom of the following drawings. Thus, for example, we obtain the representations

\[
T_1 := [\emptyset, \emptyset, \emptyset] = \begin{array}{c}
\emptyset
\end{array}, \quad T_2 := [[[\emptyset]]] = \begin{array}{c}
\emptyset
\end{array}, \quad T_3 := [[[\emptyset, \emptyset], \emptyset, [[\emptyset]]]] = \begin{array}{c}
\emptyset
\end{array}.
\]

(b) Recursively, we define a function $\# : \mathcal{T} \to \mathbb{N}$, that assigns each tree $T$ its order $\#T$: Define $\#\emptyset := 1$ and, for $d \in \mathbb{N},$

\[
\forall \mathcal{M} \in \mathcal{T}_{\leq d+1} \to \mathbb{N}_0, \quad \#\mathcal{M} := 1 + \sum_{T \in \mathcal{T}_{\leq d}} \mathcal{M}(T) \cdot \#T,
\]

where we note that the definition of $\mathcal{M} \in \mathcal{T}_{d+1}$ guarantees that only finitely many summands of the above sum are nonzero. In other words, if $\mathcal{M} = [T_1, \ldots, T_N]$, $N \in \mathbb{N}$, then $\#\mathcal{M} = 1 + \#T_1 + \cdots + \#T_N$. The order of a tree $T$ can, thus, be interpreted as its number of vertices in its graph representation described in (a). For example, for the trees $T_1, T_2, T_3$ from (a), we obtain

\[
\#T_1 = 1 + 3\#\emptyset = 4,
\]
\[
\#T_2 = 1 + \#[[\emptyset]] = 1 + 1 + \#\emptyset = 3,
\]
\[
\#T_3 = 1 + \#[[\emptyset, \emptyset]] + \#\emptyset + \#T_2 = 1 + 3 + 1 + 3 = 8.
\]

Similar to the notation introduced in (a), we define, for each $k \in \mathbb{N},$

\[
\mathcal{T}^{\#k} := \{T \in \mathcal{T} : \#T = k\}, \quad \mathcal{T}^{\leq k} := \{T \in \mathcal{T} : \#T \leq k\};
\]

and, for each $d, k \in \mathbb{N},$

\[
\mathcal{T}^{\#k}_d := \mathcal{T}^{\#k} \cap \mathcal{T}_d, \quad \mathcal{T}^{\leq k}_d := \mathcal{T}^{\leq k} \cap \mathcal{T}_{\leq d}, \quad \text{etc.}
\]
(c) Recursively, we define a **factorial function** \( ! : T \rightarrow \mathbb{N}, T \mapsto T! \), as well as a **weight function** \( \alpha : T \rightarrow \mathbb{Q}^+ \): Define \( \emptyset! := \alpha(\emptyset) := 1 \) and, for \( d \in \mathbb{N}, \)

\[
\forall M \in T_{d+1}, \quad M! := \#M \cdot \prod_{T \in T \leq d} (T!)^{M(T)}, \quad \alpha(M) := \frac{\delta(M)}{(|M|)!} \cdot \prod_{T \in T \leq d} (\alpha(T))^{M(T)},
\]

where \( |M| \) and \( \delta(M) \) are as defined in Def. and Rem. 2.47(b),(c), respectively, and where the definition of \( M \in T_{d+1} \) guarantees that only finitely many factors of the above products are \( \neq 1 \). In other words, if \( M = [T_1, \ldots, T_N], \ N \in \mathbb{N}, \) then \( M! = \#M \cdot T_1! \cdots T_N! \) and \( \alpha(M) = \frac{\delta(M)}{N!} \cdot \alpha(T_1) \cdots \alpha(T_N). \) Note that the tree factorial function can be interpreted as an extension of the usual factorial function on \( \mathbb{N} \), if one identifies each of the following trees \( \tau_d \) with the number \( d \in \mathbb{N} \): Recursively, define, for each \( d \in \mathbb{N}, \) \( \tau_1 := \emptyset, \tau_{d+1} := [\tau_d] \). A simple induction then shows that, for each \( d \in \mathbb{N}, \) \( \tau_d \) has both depth \( d \) and order \( d, \) and \( \tau_d! = d! \) holds as well. Moreover, for the trees \( T_1, T_2, T_3 \) from (a), we obtain

\[
T_1! = 4 \cdot 1^3 = 4, \quad T_2! = 3! = 6, \quad T_3! = 8 \cdot [\emptyset, \emptyset]! \cdot 1 \cdot 3! = 8 \cdot 3 \cdot 6 = 144,
\]

\[
\alpha(T_1) = \frac{3!}{3!} \cdot \alpha(\emptyset)^3 = \frac{1}{6}, \quad \alpha(T_2) = \frac{1}{1} \cdot \frac{3!}{1} = 1, \quad \alpha(T_3) = \frac{3!}{3!} \cdot \alpha([\emptyset, \emptyset]) \cdot 1 = \frac{1}{2}! = \frac{1}{2}.
\]

**Proposition 2.49.** (a) For each \( T \in T_d, \ d \in \mathbb{N}, \) one has

\[
d \leq \#T.
\]

For the tree \( \tau_d, \) defined in Def. and Rem. 2.48(c), one has \( \tau_d \in T_d, \ \#\tau_d = d. \)

(b) One has \( T = \bigcup_{k \in \mathbb{N}} T^{\#k} \). For each \( d, k \in \mathbb{N}, \) one has

\[
T_d = \bigcup_{k \in \mathbb{N}} T_d^{\#k}, \quad T^{\#k} = \bigcup_{d \in \{1, \ldots, k\}} T_d^{\#k}.
\]

(c) \( \#T_1 = \#T^{\#1} = 1. \) For each \( k \in \mathbb{N}, \) \( T^{\#k} \) is finite. For each \( d \in \mathbb{N} \) with \( d \geq 2, \) \( T_d \) is infinite and countable.

(d) \( T \) is infinite and countable.

**Proof.** (a): That \( d \leq \#T \) for each \( T \in T_d, \ d \in \mathbb{N}, \) follows by induction on \( d \in \mathbb{N} \) from the definition of \( \# : T \rightarrow \mathbb{N} \) in Def. and Rem. 2.48(b), noting that, for each \( M \in T_{d+1}, \) there must be at least one \( T \in T_d \) with \( M(T) \geq 1. \) As mentioned in Def. and Rem. 2.48(c), \( \tau_d \in T_d, \ \#\tau_d = d \) also follows via a simple induction on \( d \in \mathbb{N} \).

(b): \( T = \bigcup_{k \in \mathbb{N}} T^{\#k} \) follows as it is clear from the definition of \( \# : T \rightarrow \mathbb{N} \) in Def. and Rem. 2.48(b) that the order map is, indeed, \( \mathbb{N} \)-valued. Then, for each \( d \in \mathbb{N}, \)

\[
T_d = T_d \cap T = T_d \cap \bigcup_{k \in \mathbb{N}} T^{\#k} = \bigcup_{k \in \mathbb{N}} T_d^{\#k}.
\]
Analogous, for each $k \in \mathbb{N}$,

$$\mathcal{T}^{#k} = \mathcal{T}^{#k} \cap \mathcal{T} = \mathcal{T}^{#k} \cap \bigcup_{d \in \mathbb{N}} \mathcal{T}_d \overset{(a)}{=} \bigcup_{d \in \{1, \ldots, k\}} \mathcal{T}_d^{#k}.$$ (c): While $\#\mathcal{T}_1 = \#\mathcal{T}_1^{#1} = 1$ is immediate, we show by induction on $d \in \mathbb{N}$ that each $\mathcal{T}_d^{#k}$ is finite: $\mathcal{T}_1^{#1} = \mathcal{T}_1$ and $\mathcal{T}_k^{#k} = \emptyset$ for $k \geq 2$. For $d \in \mathbb{N}$, from the definition of $\# : \mathcal{T} \rightarrow \mathbb{N}$ in Def. and Rem. 2.48(b) and the definition of $\mathcal{T}_{d+1}$ in Def. and Rem. 2.48(a), we obtain

$$\mathcal{T}_d^{#k} \subseteq \mathcal{F} := \mathcal{F}(\mathcal{T}^{#d}_{\leq (k-1)}, \{1, \ldots, k-1\})$$

and $\mathcal{F}$ is finite due to the induction hypothesis. Thus, $\mathcal{T}_d^{#k}$ is finite as well. In consequence, according to (b), for each $k \in \mathbb{N}$, $\mathcal{T}^{#k}$ is a finite union of finite sets and, hence, finite. Clearly, $\mathcal{T}_1$ is already infinite and, thus, each $\mathcal{T}_d$, $d \geq 2$, is infinite. On the other hand, according to (b), each $\mathcal{T}_d$ is a countable union of finite sets and, hence, countable. (d): $\mathcal{T}$ is infinite as, by (c), each $\mathcal{T}_d$, $d \geq 2$, is infinite. On the other hand, by (b) and (c), $\mathcal{T}$ is a countable union of finite sets and, hence, countable.

We will now proceed to explain the relation between derivatives of $f \in C^p(\Omega, \mathbb{K}^m)$, $\Omega \subseteq \mathbb{K}^n$ ($p, n, m \in \mathbb{N}$), and (finite unlabeled rooted) trees: First, we need to recall that we can interpret higher total derivatives $D^\alpha f$ of $f$ as multilinear maps on $(\mathbb{K}^n)^\alpha$ (see, e.g., [Phi16b, Sec. 4.6]). We restate the definition of $D^\alpha f$, where, to simplify notation, we also let $f(0) := D^0 f := f$.

**Definition 2.50.** Let $n, m, p, \alpha \in \mathbb{N}$. Let $\Omega \subseteq \mathbb{K}^n$ be open, $f : \Omega \rightarrow \mathbb{K}^m$, $f \in C^p(\Omega, \mathbb{K}^m)$. For $\alpha \leq p$, define the total derivative of order $\alpha$ of $f$ as follows: For each $y \in \Omega$,

$$f^{(\alpha)}(y) := D^\alpha f(y) : (\mathbb{K}^n)^\alpha \rightarrow \mathbb{K}^m,$$

$$D^\alpha f(y)(h^1, \ldots, h^\alpha)_l = \sum_{j_1, \ldots, j_\alpha=1}^n \partial_{j_1} \cdots \partial_{j_\alpha} f_i(y) h^1_{j_1} \cdots h^\alpha_{j_\alpha}, \quad l \in \{1, \ldots, m\} \quad (2.69)$$

(note that the coordinate functions of $f^{(\alpha)} = D^\alpha f$ are precisely the partials of order $\alpha$ of $f$). As usual, we also let $f^{(0)} := D^0 f := f$.

Next, we observe that, for $m = n$, we can apply total derivatives to (the results of) other total derivates: For example, we can form the following expressions (assuming the highest-order derivative to exist):

$$D_1 := f''(f, f), \quad D_2 := f''(f^4(f, f'), f, f), \quad D_3 := f''(f^4(f', f), f, f''(f, f)),$$

where we omitted the argument $y \in \Omega$ in the notation. As we assume all partials of $f$ to be continuous, the order of differentiation does not matter and the $f^{(\alpha)}$ actually
constitute symmetric multilinear maps, i.e. their results do not depend on the order of the arguments. For example, we can rewrite \( D_2 \) and \( D_3 \) from above as follows:

\[
D_2 := f''(f, f^{(4)}(f', f, f, f)), \quad D_3 := f'''(f''(f, f), f'(f')), f).
\]

The key observation is now that the structure of such nested derivatives is precisely the same as that of the above-defined trees: To obtain the corresponding tree, one merely has to replace each \( f^{(0)} \) with \( \emptyset \), replace parentheses \( () \) with brackets \([\,]\), and omit derivatives \( f^{(\alpha)} \), \( \alpha \geq 1 \). For \( D_1, D_2, D_3 \) from above, we, thus, obtain

\[
\begin{align*}
D_1 &= f''(f, f) \cong [\emptyset, \emptyset] = \emptyset, \\
D_2 &= f''(f, f^{(4)}(f', f, f, f)) \cong [\emptyset, \emptyset, \emptyset, \emptyset], \\
D_3 &= f'''(f''(f, f), f'(f')), f) \cong [\emptyset, \emptyset, [\emptyset]], \emptyset.
\end{align*}
\]

The above considerations lead us to the following definition of derivatives with respect to trees. Bearing in mind \( d \leq \#T \) for each \( T \in \mathcal{T}_d \) according to Prop. 2.49(a), we define \( f^{(T)} \) for each \( f \) being \( p \) times continuously differentiable and for each \( T \in \mathcal{T}^{\# \leq (p+1)} \), recursively over \( d \in \{1, \ldots, p+1\} \):

**Definition 2.51.** Let \( p \in \mathbb{N}_0 \), \( n \in \mathbb{N} \), \( \Omega \subseteq \mathbb{K}^n \) open, \( f \in C^p(\Omega, \mathbb{K}^n) \). Define

\[
f^{(\emptyset)} := f.
\]

For \( d \in \{1, \ldots, p\} \), we recursively define

\[
\forall T = [T_1, \ldots, T_N] \in \mathcal{T}^{\# \leq (p+1)}_{\leq d}, \quad \forall y \in \Omega
f^{(T)}(y) := f^{(N)}(y) \left( f^{(T_1)}(y), \ldots, f^{(T_N)}(y) \right)
\]

(note that \( f^{(T)}(y) \) is then well-defined due to the symmetry of \( f^{(N)}(y) \)).

**Example 2.52.** In this example, we construct all trees up to order 4. For each tree \( T \), we provide its order \( \#T \), its factorial \( T! \), its weight \( \alpha(T) \), and the derivative \( f^{(T)} \), where \( f \in C^3(\Omega, \mathbb{K}^n) \), \( \Omega \subseteq \mathbb{K}^n \) open, \( n \in \mathbb{N} \).

(a) \( T_{11} := \emptyset = \cdot \) is the only tree of order one. We have

\[
T_{11} = \emptyset = \cdot, \quad \#T_{11} = 1, \quad T_{11}! = 1, \quad \alpha(T_{11}) = 1, \quad f^{(T_{11})} = f.
\]

(b) If \( T \in \mathcal{T}^{\#2} \), then \( T = [S] \) with \( \#S = 1 \). Thus \( S = T_{11} \) and \( T_{21} := [\emptyset] \) is the only tree of order 2. We have

\[
T_{21} = [\emptyset] = \hat{1}, \quad \#T_{21} = 2, \quad T_{21}! = 2 \cdot 1 = 2, \quad \alpha(T_{21}) = \frac{1}{1!} = 1, \quad f^{(T_{21})} = f'(f).
\]
(c) We show there are precisely 2 trees of order 3: If \( T \in \mathcal{T}^{#3} \), then \( T = [S] \) with \( \#S = 2 \) or \( T = [S_1, S_2] \) with \( \#S_1 = \#S_2 = 1 \). Thus, we have

\[
\mathcal{T}^{#3} = \{T_{31}, T_{32}\}, \quad T_{31} := [T_{21}] = T, \quad T_{32} := [0, \emptyset] = V,
\]

\[
T_{31} = [\{\emptyset\}] = T, \quad \#T_{31} = 3, \quad T_{31}! = 3 \cdot T_{21}! = 6, \quad \alpha(T_{31}) = \frac{1}{1!} \cdot \alpha(T_{21}) = 1, \quad f^{(T_{31})} = f'(f'(f)),
\]

\[
T_{32} = [0, \emptyset] = V, \quad \#T_{32} = 3, \quad T_{32}! = 3 \cdot 1 \cdot 1 = 3, \quad \alpha(T_{32}) = \frac{1}{2!} \cdot 1 \cdot 1 = \frac{1}{2}, \quad f^{(T_{32})} = f''(f, f).
\]

(d) We show there are precisely 4 trees of order 4: If \( T \in \mathcal{T}^{#4} \), then \( T = [S] \) with \( \#S = 3 \) or \( T = [S_1, S_2] \) with \( \#S_1 + \#S_2 = 3 \) (i.e. \( T = [0, T_{21}] \)) or \( T = [S_1, S_2, S_3] \) with \( \#S_1 = \#S_2 = \#S_3 = 1 \). Thus, we have

\[
\mathcal{T}^{#4} = \{T_{41}, T_{42}, T_{43}, T_{44}\}, \quad T_{41} := [T_{31}] = T, \quad T_{42} := [T_{32}] = V, \quad T_{43} := [0, T_{21}] = V, \quad T_{44} := [0, 0, 0] = V,
\]

\[
T_{41} = [\{\emptyset\}] = T, \quad \#T_{41} = 4, \quad T_{41}! = 4 \cdot T_{31}! = 24, \quad \alpha(T_{41}) = \frac{1}{1!} \cdot \alpha(T_{31}) = 1, \quad f^{(T_{41})} = f'(f'(f'(f'))),
\]

\[
T_{42} = [\{\emptyset\}] = V, \quad \#T_{42} = 4, \quad T_{42}! = 4 \cdot T_{32}! = 12, \quad \alpha(T_{42}) = \frac{1}{2!} \cdot \alpha(T_{32}) = \frac{1}{2}, \quad f^{(T_{42})} = f''(f, f),
\]

\[
T_{43} = [0, \{\emptyset\}] = V, \quad \#T_{43} = 4, \quad T_{43}! = 4 \cdot 1 \cdot 2 = 8, \quad \alpha(T_{43}) = \frac{1}{2!} \cdot 1 \cdot 1 = 1, \quad f^{(T_{43})} = f'''(f, f, f'),
\]

\[
T_{44} = [0, 0, 0] = V, \quad \#T_{44} = 4, \quad T_{44}! = 4 \cdot 1 \cdot 3 = 12, \quad \alpha(T_{44}) = \frac{1}{3!} \cdot 1 \cdot 3 = \frac{1}{6}, \quad f^{(T_{44})} = f''''(f, f, f, f).
\]

Proceeding with the strategy outlined at the beginning of the section, we now want to obtain Taylor expansions of the exact solution \( \phi \) to \( y' = f(y), \ y(\xi) = \eta \), and of the defining function \( \varphi \) of an RK method. We will obtain concise and structured forms of these expansions in Prop. 2.54 and Prop. 2.57 below, making use of the above-defined trees. In preparation, we provide a suitable version of Taylor’s theorem:

**Theorem 2.53** (Taylor). Let \( m, n \in \mathbb{N} \). Let \( \Omega \subseteq \mathbb{K}^n \) be open and \( f \in C^{p+1}(\Omega, \mathbb{K}^m) \) for some \( p \in \mathbb{N}_0 \). Let \( y \in \Omega \) and \( h \in \mathbb{K}^n \) be such that the line segment \( S_{y,y+h} \) between \( y \) and \( y+h \) is a subset of \( \Omega \). Then the following formula, also known as Taylor’s formula, holds:

\[
f(y + h) = \sum_{k=0}^{p} \frac{f^{(k)}(y)(h, \ldots, h)}{k!} \cdot R_p(y, h),
\]

where \( R_p(y, h) \in \mathbb{K}^m \) with \( ||R_p(y, h)|| = O(||h||^{p+1}) \) for \( ||h|| \to 0 \), i.e.

\[
\limsup_{h \to 0} \frac{||R_p(y, h)||}{||h||^{p+1}} =: C(y) \in \mathbb{R}_0^+.
\]

Moreover, the function \( h \mapsto R_p(y, h) \) is continuous in a neighborhood \( N_y \) of 0.
Proof. According to [Phi16b, Th. 4.44], Taylor’s formula holds in the stated form for \( \Omega \subseteq \mathbb{R}^n \) and \( \mathbb{K} \)-valued \( f \) with the remainder term in integral form

\[
R_p(y, h) = \int_0^1 \frac{(1-t)^p}{p!} f^{(p+1)}(y + th)(h, \ldots, h) \, dt.
\]

It then also holds for \( \mathbb{K}^m \)-valued \( f \), since we can apply the \( \mathbb{K} \)-valued version to each coordinate function \( f_i \) of \( f \), \( i = 1, \ldots, m \). It then also extends to \( \Omega \subseteq \mathbb{C}^n \), as we can interpret \( \mathbb{C}^n \) as \( \mathbb{R}^{2n} \) (here, as always in this class, we only consider \( \mathbb{R} \)-differentiability, even if \( \mathbb{K} = \mathbb{C} \)). It remains to verify \( R_p(y, h) \in \mathbb{K}^m \) with \( \| R_p(y, h) \| = O(\| h \|^{p+1}) \) for \( \| h \| \to 0 \): Without loss of generality, we may interpret \( \| \cdot \| \) as the max norm (due to norm equivalence). Moreover, there exists a compact neighborhood \( K \) of \( y \) with \( K \subseteq \Omega \). Each absolute value of the finitely many continuous partials of order \( p+1 \) of \( f \) is bounded on \( K \), say by \( \bar{C}(y) \in \mathbb{R}_0^+ \). Then we can estimate, for \( y + h \in K \),

\[
|R_p(y, h)| \leq \int_0^1 \frac{(1-t)^p}{p!} \sum_{j_1, \ldots, j_{p+1}=1}^n \partial_{j_1} \cdots \partial_{j_{p+1}} f_i(y + th) h_{j_1} \cdots h_{j_{p+1}} \, dt \leq \frac{n^{p+1} \bar{C}(y) \| h \|^{p+1}}{p!}.
\]

Taking the max over \( l \in \{1, \ldots, m \} \) in the above estimate, proves the claimed \( \| R_p(y, h) \| = O(\| h \|^{p+1}) \) for \( \| h \| \to 0 \). Solving Taylor’s formula for \( R_p(y, h) \), the continuity of \( h \mapsto R_p(y, h) \) follows from the continuity of \( f \) and the continuity of the multilinear functions \( f^{(k)}(y) \).

Proposition 2.54. Let \( n \in \mathbb{N} \), \( p \in \mathbb{N}_0 \), \( \Omega \subseteq \mathbb{K}^n \) open, \( f \in C^p(\Omega, \mathbb{K}^n) \). If \( I \subseteq \mathbb{R} \) is an open interval and \( \phi : I \to \mathbb{K}^n \) is a solution to \( y' = f(y) \), then one has \( \phi \in C^{p+1}(I, \mathbb{K}^n) \) and, for each \( x, h \in \mathbb{R} \) such that \( x, x + h \in I \), \( \phi \) admits the Taylor expansion

\[
\phi(x + h) = \phi(x) + \sum_{T \in T^p \#T \leq p} \frac{h^#T}{T!} \alpha(T) f^{(T)}(\phi(x)) + R_p(x, h),
\]

where \( \| R_p(x, h) \| = O(\| h \|^{p+1}) \) for \( \| h \| \to 0 \) and the map \( h \mapsto R_p(x, h) \) is continuous in a neighborhood of \( 0 \).

Proof. We have \( \phi \in C^{p+1}(I, \mathbb{K}^n) \) according to Prop. C.1 of the Appendix. We now prove (2.70) via induction on \( p \): For \( p = 0 \), the sum in (2.70) is empty and the formula holds due to Th. 2.53 (applied for \( p = 0 \) and with \( \phi \) instead of \( f \)). Now fix \( p \in \mathbb{N}_0 \), \( f \in C^{p+1}(\Omega, \mathbb{K}^n) \), and assume (2.70) to hold by induction hypothesis. We apply the induction hypothesis with \( h \) replaced by \( t \in [0, h] \) and set

\[
y := \phi(x), \quad y(t) := \sum_{T \in T^p \#T \leq p} \frac{t^#T}{T!} \alpha(T) f^{(T)}(y) + R_p(x, t)
\]
to obtain
\[
\begin{align*}
    f(\phi(x + t)) &= \text{ind. hyp.} \quad f(y + y(t)) \\
    \text{Th. 2.53} &= \sum_{k=0}^{p} \frac{f^{(k)}(y)(y(t), \ldots, y(t))}{k!} + R_{p,f}(y, y(t)) \\
    &= \sum_{k=0}^{p} \frac{1}{k!} f^{(k)}(y) \left( \sum_{T_1 \in T \# \leq p} \frac{t^{#T_1}}{T_1!} \alpha(T_1) f^{(T_1)}(y), \\
    &\quad \ldots, \sum_{T_k \in T \# \leq p} \frac{t^{#T_k}}{T_k!} \alpha(T_k) f^{(T_k)}(y) \right) \\
    &\quad + P_{p,f}(x, t) + R_{p,f}(y, y(t)),
\end{align*}
\]
where, for the \(R_{p,f}(x, t)\) occurring in \(y(t)\), we know \(\|R_{p,f}(x, t)\| = O(|t|^{p+1})\) for \(|t| \to 0\). Moreover, due to (2.69), \(P_{p,f}(x, t)\) has the form of a finite sum of terms \(v \prod_{l=1}^{p} X_l\), where \(v \in \mathbb{K}^n\) and each \(X_l \in \mathbb{K}\) is 1 or a component of \(\sum_{T \in T \# \leq p} \frac{t^{#T}}{T!} \alpha(T) f^{(T)}(y)\) or a component of \(R_{p,f}(x, t)\), where, in each summand, at least one \(X_l = (R_{p,f}(x, t))_j\). Thus, for each summand, we have \(\|v \prod_{l=1}^{p} X_l\| = O(|t|^{p+1})\) for \(|t| \to 0\), implying \(\|P_{p,f}(x, t)\| = O(|t|^{p+1})\) for \(|t| \to 0\) as well. We also know \(\|R_{p,f}(y, v)\| = O(\|v\|^{p+1})\) for \(\|v\| \to 0\), implying
\[
\lim_{t \to 0} \sup \frac{\|R_{p,f}(y, y(t))\|}{|t|^{p+1}} = \lim_{t \to 0} \sup \frac{\|g(t)\|^{p+1} \|R_{p,f}(y, y(t))\|}{|t|^{p+1} \|y(t)\|^{p+1}} \in \mathbb{R}_0^+.
\]
Thus, we have shown
\[
\begin{align*}
    f(\phi(x + t)) &= \sum_{k=0}^{p} \frac{1}{k!} f^{(k)}(y) \left( \sum_{T_1 \in T \# \leq p} \frac{t^{#T_1}}{T_1!} \alpha(T_1) f^{(T_1)}(y), \\
    &\quad \ldots, \sum_{T_k \in T \# \leq p} \frac{t^{#T_k}}{T_k!} \alpha(T_k) f^{(T_k)}(y) \right) \\
    &\quad + R_{p,f,1}(x, t),
\end{align*}
\]
where \(\|R_{p,f,1}(x, t)\| = O(|t|^{p+1})\) for \(|t| \to 0\). Then \(t \mapsto R_{p,f,1}(x, t)\) is continuous in \([0, h]\), due to the continuity of \(f, \phi, \) and the multilinear maps \(f^{(k)}(y)\). Using the multilinearity of the \(f^{(k)}(y)\) once again yields
\[
\begin{align*}
    f(\phi(x + t)) &= \sum_{k=0}^{p} \frac{1}{k!} \sum_{#T_1 + \cdots + #T_k \leq p} \frac{t^{#T_1 + \cdots + #T_k}}{T_1! \cdots T_k!} \alpha(T_1) \cdots \alpha(T_k) f^{(k)}(y) (f^{(T_1)}(y), \ldots, f^{(T_k)}(y)) \\
    &\quad + R_{p,f,2}(x, t),
\end{align*}
\]
where \( \|R_{p,f,2}(x,t)\| = O(|t|^{p+1}) \) for \( |t| \to 0 \) and \( t \mapsto R_{p,f,2}(x,t) \) is continuous in \([0,h]\).
Now, instead of summing over all ordered \( k \)-tuples, we can use the symmetry of \( f^{(k)}(y) \) to merely sum over all unordered \( k \)-tuples, using that each unordered \( k \)-tuple \( T = [T_1, \ldots, T_k] \) corresponds to \( \delta(T) \) ordered \( k \)-tuples. Thus, also making use of the recursive definitions of \( T, \#T, T!, \alpha(T) \), and \( f^{(T)} \), we obtain

\[
\phi(x + h) - \phi(x) = \sum_{k=0}^{p} \sum_{T \in \mathcal{T}^* \subseteq \{p+1\}, T = [T_1, \ldots, T_k]} \frac{\#T \cdot t^{#T-1}}{T!} \delta(T) \frac{\alpha(T_1) \cdots \alpha(T_k)}{k!} f^{(T)}(y) + R_{p,f,2}(x,t),
\]

As \( \phi \) is a solution to the ODE, we infer from the above equality

\[
\phi(x + h) - \phi(x) = \int_0^h \phi'(x + t) \, dt = \int_0^h f(\phi(x + t)) \, dt = \sum_{T \in \mathcal{T}^* \subseteq \{p+1\}} \frac{h^{#T}}{T!} \alpha(T) f^{(T)}(\phi(x)) + R_{p+1}(x,h),
\]

where

\[
R_{p+1}(x,h) := \int_0^h R_{p,f,2}(x,t) \, dt
\]

(the integral exists, as \( t \mapsto R_{p,f,2}(x,t) \) is continuous). Thus,

\[
\limsup_{h \to 0} \frac{\|R_{p+1}(x,h)\|}{|h|^{p+2}} =: C(x) \in \mathbb{R}^+_0,
\]

showing \( \|R_{p+1}(x,h)\| = O(|h|^{p+2}) \) for \( |h| \to 0 \) and we have verified (2.70) to hold for \( p + 1 \). The continuity of \( h \mapsto R_{p+1}(x,h) \) follows, again, by solving (2.70) for \( R_{p+1}(x,h) \).

In consequence, the induction and the proof of the proposition are complete. \( \blacksquare \)

The striking and immensely useful observation due to Butcher is the fact that one can give the Taylor expansion of the defining function of an RK method an analogous structure to that of the above Taylor expansion of the solution \( \phi \). It turns out that certain coefficients in the Taylor expansion of the defining function can be computed recursively, using the above trees \( T \), making use of auxiliary vectors \( v_A(T) \in \mathbb{K}^s \), depending on the RK matrix \( A \in \mathcal{M}(s, \mathbb{K}) \). Thus, before we can provide the Taylor expansion in Prop. 2.57 below, we still need to provide the recursive definition of the \( v_A(T) \):

**Definition 2.55.** Let \( s \in \mathbb{N}, A \in \mathcal{M}(s, \mathbb{K}) \). Recursively, we define a function \( v_A : \mathcal{T} \to \mathbb{K}^s \): Define

\[
v_A(\emptyset) := (1, \ldots, 1)^t \in \mathbb{K}^s,
\]

and, for \( d \in \mathbb{N}, \)

\[
\forall \mathcal{M} \in \mathcal{T}_{d+1}, \quad j \in \{1, \ldots, s\} \quad v_A(\mathcal{M})_j := \prod_{T \in \mathcal{T}_{\leq d}} (A v_A(T))_j^{\mathcal{M}(T)},
\]
where the definition of \( M \in T_{d+1} \) guarantees that only finitely many factors of the above products are \( \neq 1 \). In other words, if \( M = [T_1, \ldots, T_N] \), \( N \in \mathbb{N} \), then \( v_A(M)_j = (A v_A(T_1))_j \cdots (A v_A(T_N))_j \) for each \( j \in \{1, \ldots, s\} \).

**Example 2.56.** Let \( s \in \mathbb{N}, A \in \mathcal{M}(s, \mathbb{K}) \). We compute \( v_A(T) \) for all the trees \( T \) of Ex. 2.52, i.e. for all trees of order at most 4. We define

\[
\forall j \in \{1, \ldots, s\} \quad c_j := \sum_{l=1}^{s} a_{jl} \quad (2.71)
\]

(while, here, (2.71) is merely the definition of the \( c_j \), it is the same as the node condition (2.35), i.e. it is also satisfied if \( A \) and the \( c_j \) are parameters of an RK method that satisfies the node condition). The following trees \( T_{\mu\nu} \) are the same as in Ex. 2.52. We obtain, for each \( j \in \{1, \ldots, s\} \),

\[
\begin{align*}
T_{11} = 0 = & \cdot, \\ v_A(T_{11})_j = & 1, \\ T_{21} = & [0] = 1, \\ v_A(T_{21})_j = & (A v_A(T_{11}))_j = \sum_{l=1}^{s} a_{jl} \cdot 1 = c_j,
\end{align*}
\]

\[
\begin{align*}
T_{31} = & [T_{21}] = 1, \\ v_A(T_{31})_j = & (A v_A(T_{21}))_j = \sum_{l=1}^{s} a_{jl} c_l,
\end{align*}
\]

\[
\begin{align*}
T_{32} = & [0, 0] = \bigvee, \\ v_A(T_{32})_j = & (A v_A(T_{11}))_j = \sum_{l=1}^{s} a_{jl} c_l,
\end{align*}
\]

\[
\begin{align*}
T_{41} = & [T_{31}] = 1, \\ v_A(T_{41})_j = & (A v_A(T_{11}))_j = \sum_{l=1}^{s} a_{jl} a_{kl} c_l,
\end{align*}
\]

\[
\begin{align*}
T_{42} = & [T_{32}] = \bigvee, \\ v_A(T_{42})_j = & (A v_A(T_{32}))_j = \sum_{l=1}^{s} a_{jl} c_l^2,
\end{align*}
\]

\[
\begin{align*}
T_{43} = & [T_{11}, T_{21}] = \bigvee, \\ v_A(T_{43})_j = & (A v_A(T_{11}))_j (A v_A(T_{21}))_j = c_j \sum_{l=1}^{s} a_{jl} c_l,
\end{align*}
\]

\[
\begin{align*}
T_{44} = & [T_{11}, T_{11}, T_{11}, T_{11}] = \bigvee, \\ v_A(T_{44})_j = & (A v_A(T_{11}))_j \cdot (A v_A(T_{11}))_j \cdot (A v_A(T_{11}))_j = c_j^3.
\end{align*}
\]

**Proposition 2.57.** Let \( n \in \mathbb{N}, p \in \mathbb{N}_0, \Omega \subseteq \mathbb{K}^n \) open, \( f \in C^p(\Omega, \mathbb{K}^n) \). Let \( s \in \mathbb{N} \) and consider an \( s \)-stage RK method according to Def. 2.24 with weights vector \( b^i = (b_1, \ldots, b_s) \in \mathbb{R}^s \) and RK matrix \( A \in \mathcal{M}(s, \mathbb{K}) \). Assume

\[
\forall y \in \Omega \quad \forall j \in \{1, \ldots, s\} \quad \lim_{h \downarrow 0} k_j(y, h) = f(y),
\]

(sufficient conditions are that the RK method is explicit or that \( f \) satisfies the conditions of Th. 2.31 and is in standard form\(^1\); also note that the \( k_j \) do not depend on \( x \), as \( f \) does not depend on \( x \)). Then, for each \( y \in \Omega \) and each sufficiently small \( h \in \mathbb{R}^+ \), the \( k_j \) admit Taylor expansions

\[
k_j(y, h) = \sum_{T \in T_{\#_k \leq p}} h^{#T-1} \alpha(T) v_A(T)_j f(T)(y) + R_{p,j}(y, h), \quad (2.72a)
\]

\(^1\)Due to Prop. B.4 of the Appendix, the conditions of Th. 2.31 are automatically satisfied for \( p \geq 1 \).
where \( \| R_{p,j}(y,h) \| = O(|h|^p) \) for \(|h| \to 0 \). Thus, the defining function \( \varphi \) of the RK method admits the Taylor expansion

\[
\varphi(y,h) = \sum_{T \in T^{\#} \leq p} h^{\#T-1} \alpha(T) b^T A(T) f^{(T)}(y) + R_p(y,h),
\]

(2.72b)

where \( \| R_p(y,h) \| = O(|h|^p) \) for \(|h| \to 0 \).

**Proof.** Since

\[
\varphi(y,h) = \sum_{j=1}^{s} b_j k_j(y,h),
\]

(2.72b), clearly, follows from (2.72a). Thus, it suffices to show (2.72a). The proof of (2.72a) is conducted similar to the proof of Prop. 2.54, using induction on \((2.72b) \), clearly, follows from (2.72a). Thus, it suffices to show (2.72a). The proof of (2.72a) is conducted similar to the proof of Prop. 2.54, using induction on \( p \). For \( p = 0 \), the sum in (2.72a) is empty and the statement holds due to the assumption \( \lim_{h \downarrow 0} k_j(y,h) = f(y) \). Now fix \( p \in \mathbb{N}_0 \), \( f \in C^{p+1}(\Omega, \mathbb{K}^n) \), and assume (2.72a) to hold by induction hypothesis. We set

\[
R_{p,j,1}(y,h) := hAR_{p,j}(y,h), \quad y(h) := \sum_{T \in T^{\#} \leq p} h^{\#T} \alpha(T) (A v_A(T))_j f^{(T)}(y) + R_{p,j,1}(y,h)
\]

to obtain

\[
k_j(y,h) = f \left( y + h \sum_{l=1}^{s} a_{jl} k_l(y,h) \right) \overset{\text{ind.hyp.}}{=} f(y + y(h))
\]

Th. 2.53

\[
\sum_{k=0}^{p} f^{(k)}(y) \frac{(y(h), \ldots, y(h))}{k!} + R_{p,f}(y,y(h))
\]

\[
= \sum_{k=0}^{p} \frac{1}{k!} f^{(k)}(y) \left( \sum_{T_k \in T^{\#} \leq p} h^{\#T_k} \alpha(T_k) (A v_A(T_k))_j f^{(T_k)}(y), \ldots, \sum_{T_k \in T^{\#} \leq p} h^{\#T_k} \alpha(T_k) (A v_A(T_k))_j f^{(T_k)}(y) \right)
\]

where \( \| R_{p,j}(y,h) \| = O(|h|^p) \) for \(|h| \to 0 \). Moreover, due to (2.69), \( P_{p,f,j}(y,h) \) has the form of a finite sum of terms \( v \prod_{l=1}^{p} X_l \), where \( v \in \mathbb{K}^n \) and each \( X_l \in \mathbb{K} \) is 1 or a component of \( \sum_{T \in T^{\#} \leq p} h^{\#T} \alpha(T) (A v_A(T))_j f^{(T)}(y) \) or a component of \( R_{p,j,1}(y,h) \), where in each summand, at least one \( X_l = (R_{p,j,1}(y,h))_a \). Thus, for such summand, we have \( \| v \prod_{l=1}^{p} X_l \| = O(|h|^{p+1}) \) for \(|h| \to 0 \), implying \( \| P_{p,f,j}(y,h) \| = O(|h|^{p+1}) \) for \(|h| \to 0 \) as well. We also know \( \| R_{p,f}(y,v) \| = O(||v||^{p+1}) \) for \(||v|| \to 0 \), implying

\[
\limsup_{h \to 0} \frac{\| R_{p,f}(y,y(h)) \|}{|h|^{p+1}} = \limsup_{h \to 0} \frac{\| y(h) \|^{p+1} \| R_{p,f}(y,y(h)) \|}{|h|^{p+1} \| y(h) \|^{p+1}} \in \mathbb{R}_0^+.
\]
Thus, we have shown

$$k_j(y, h) = \sum_{k=0}^{p} \frac{1}{k!} f^{(k)}(y) \left( \sum_{T_i \in T^\# \leq p} h^{|T_i|} \alpha(T_i) (A v_A(T_i))_j f^{(T_i)}(y), \right.$$  

\[ \ldots, \sum_{T_k \in T^\# \leq p} h^{|T_k|} \alpha(T_k) (A v_A(T_k))_j f^{(T_k)}(y) \right) + R_{p,j}(y, h), \]

where \( \| R_{p,j}(y, h) \| = O(|h|^{p+1}) \) for \( |h| \to 0 \). Using the multilinearity of the \( f^{(k)}(y) \) once again yields

\[
k_j(y, h) = \sum_{k=0}^{p} \frac{1}{k!} \sum_{T^\# \leq p} h^{|T^\#|} \left( \prod_{l=1}^{k} \alpha(T_l)(A v_A(T_l))_j \right) f^{(l)}(y) \sum_{T^\# \leq p} h^{|T^\#|} \alpha(T) \right) f^{(T)}(y) + R_{p+1,j}(y, h), \]

where \( \| R_{p+1,j}(y, h) \| = O(|h|^{p+1}) \) for \( |h| \to 0 \). Now, instead of summing over all ordered \( k \)-tuples, we can use the symmetry of \( f^{(k)}(y) \) to merely sum over all unordered \( k \)-tuples, using that each unordered \( k \)-tuple \( T = [T_1, \ldots, T_k] \) corresponds to \( \delta(T) \) ordered \( k \)-tuples. Thus, also making use of the recursive definitions of \( T, \#T, v_A(T), \alpha(T), \) and \( f^{(T)} \), we obtain

\[
k_j(y, h) = \sum_{k=0}^{p} \frac{1}{k!} \sum_{T \in T^\# \leq (p+1)} h^{|T^\#|} v_A(T)_j \left( \prod_{T \in \delta(T)} \frac{\alpha(T_1) \cdots \alpha(T_k)}{k!} \right) f^{(T)}(y) + R_{p+1,j}(y, h), \]

Hence, we have verified (2.72a) to hold for \( p+1 \), completing the induction and the proof of the proposition.

\[ \blacksquare \]

**Definition 2.58.** Let \( s \in \mathbb{N} \), \( b^t = (b_1, \ldots, b_s) \in \mathbb{R}^s \), \( A \in \mathcal{M}(s, \mathbb{K}) \). We say that the \( s \)-stage RK method with weights vector \( b \) and RK matrix \( A \) satisfies the **consistency condition of order** \( p \in \mathbb{N} \) if, and only if,

\[
\forall_{T \in T^\# \leq p} b^t v_A(T) = \frac{1}{T!}. \tag{2.73}
\]

**Remark 2.59.** We note that (2.73) is a generalization of the consistency condition (2.34): Since \( \emptyset \) is the only tree of order 1, \( v_A(\emptyset) = (1, \ldots, 1)^t \), and \( \emptyset! = 1 \), the consistency condition of order 1 reads \( \sum_{j=1}^{s} b_j = 1 \), which is precisely (2.34).

**Theorem 2.60 (Butcher).** Let \( n, s \in \mathbb{N} \), \( b^t = (b_1, \ldots, b_s) \in \mathbb{R}^s \), \( A \in \mathcal{M}(s, \mathbb{K}) \). Consider an \( s \)-stage RK method with weight vector \( b \) and RK matrix \( A \).
(a) Autonomous Case: Let $\Omega \subseteq \mathbb{K}^n$ be open, and let $f \in C^p(\Omega, \mathbb{K}^n)$, $p \in \mathbb{N}$, be such that the local truncation error $\lambda$ is well-defined (for example, $\Omega = \mathbb{K}^n$ is sufficient) and such that

$$\forall y \in \Omega \quad \forall j \in \{1, \ldots, s\} \quad \lim_{h \to 0} k_j(y, h) = f(y)$$

(see the statement of Prop. 2.57 for sufficient conditions). Let $(\xi, \eta) \in \mathbb{R} \times \Omega$ and choose $b > \xi$ and $\phi : [\xi, b] \rightarrow \mathbb{K}^n$ such that $\phi$ is the unique solution to $y' = f(y)$, $y(\xi) = \eta$, on $[\xi, b]$. If the RK method satisfies the consistency condition of order $p$, i.e. (2.73), then it is consistent of order $p$.

(b) Nonautonomous Case: Let $G \subseteq \mathbb{R} \times \mathbb{K}^n$ be open, and let $f \in C^p(G, \mathbb{K}^n)$, $p \in \mathbb{N}$, be such that the local truncation error $\lambda$ is well-defined (for example, $G = \mathbb{R} \times \mathbb{K}^n$ is sufficient). Moreover, let the RK method be in standard form. Let $(\xi, \eta) \in G$ and choose $b > \xi$ and $\phi : [\xi, b] \rightarrow \mathbb{K}^n$ such that $\phi$ is the unique solution to $y' = f(x, y)$, $y(\xi) = \eta$, on $[\xi, b]$. If the RK method satisfies the consistency condition of order $p$, i.e. (2.73), plus the node condition (2.35) (with node vector $c = (c_1, \ldots, c_s) \in \mathbb{R}^s$), then it is consistent of order $p$.

(c) If the RK method is consistent of order $p \in \mathbb{N}$ for each autonomous initial value problem $y' = f(y)$, $y(0) = 0$, with $f \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$, $n \in \mathbb{N}$, then it must satisfy the consistency condition of order $p$, i.e. (2.73).

Proof. (a): The main work of the proof has already been carried out in the proofs of Prop. 2.54 and Prop. 2.57, respectively. It remains to use these propositions in (2.17): Let $x \in [\xi, b]$ and set $y := \phi(x)$. According to (2.17), for each sufficiently small $h \in [0, b - x]$, we obtain the local truncation error

$$\lambda(x, h) = y + h \phi(y, h) - \phi(x + h)$$

$$= y + h \sum_{T \in T^\# \leq p} h^{|T| - 1} \alpha(T) b^T v_A(T) f^{(T)}(y) + h R_{p, \phi}(y, h)$$

$$- y - \sum_{T \in T^\# \leq p} \frac{h^{|T|}}{T!} \alpha(T) f^{(T)}(y) + R_{p, \phi}(x, h)$$

$$\overset{(2.73)}{=} h R_{p, \phi}(y, h) - R_{p, \phi}(x, h), \quad (2.74)$$

where $\|R_{p, \phi}(y, h)\| = O(|h|^p)$ for $|h| \to 0$ and $\|R_{p, \phi}(x, h)\| = O(|h|^{p+1})$ for $|h| \to 0$. Thus,

$$\limsup_{h \to 0} \frac{\|\lambda(x, h)\|}{|h|^{p+1}} \in \mathbb{R}_0^+,$$

showing the method to be consistent of order $p$.

(b): We know from Th. 2.35 that $y' = f(x, y)$, $y(\xi) = \eta$, is equivalent to the autonomous initial value problem $y' = g(y)$, $y(\xi) = (\xi, \eta)$, where

$$g : G \rightarrow \mathbb{K}^{n+1}, \quad g(x, y_1, \ldots, y_n) := (1, f(x, y_1, \ldots, y_n)).$$
The RK method defining functions for the nonautonomous and the autonomous problem are, respectively,

\[ \varphi_f : D_{\varphi_f} \rightarrow \mathbb{K}^n, \quad \varphi_f(x, y, h) = \sum_{j=1}^{s} b_j k_{f,j}(x, y, h), \]

\[ \varphi_g : D_{\varphi_g} \rightarrow \mathbb{K}^{n+1}, \quad \varphi_g((x, y), h) = \sum_{j=1}^{s} b_j k_{g,j}((x, y), h), \]

where the \( k_{f,j}(x, y, h) \) satisfy

\[ \forall j \in \{1, \ldots, s\} \quad k_{f,j}(x, y, h) = f \left( x + c_j h, y + h \sum_{l=1}^{s} a_{jl} k_{f,l}(x, y, h) \right), \]

and the \( k_{g,j}((x, y), h) \) satisfy

\[ \forall j \in \{1, \ldots, s\} \quad k_{g,j}((x, y), h) = g \left( (x, y) + h \sum_{l=1}^{s} a_{jl} k_{g,l}((x, y), h) \right). \]

As we assume \( f \in C^p(G, \mathbb{K}^n) \) with \( p \geq 1 \), \( f \) is locally Lipschitz with respect to \( y \) by Prop. B.4 of the Appendix, i.e. \( f \) satisfies the conditions of Th. 2.31. Hence, the RK method has unique local solutions and, thus, we know from Lem. 2.36 (as we also assume the node condition and the RK method to be in standard form), for each sufficiently small \( h > 0 \),

\[ \forall j \in \{1, \ldots, s\} \quad k_{g,j}((x, y), h) = (1, k_{f,j}(x, y, h)). \]

We know from Th. 2.35 that, if \( \phi : [\xi, b] \rightarrow \mathbb{K}^n \) is the solution to \( y' = f(x, y), y(\xi) = \eta \), then \( \psi : [\xi, b] \rightarrow \mathbb{K}^{n+1}, \psi(x) = (x, \phi(x)) \), is the solution to \( y' = g(y), y(\xi) = (\xi, \eta) \), yielding, for the respective local truncation errors at \( x \in [\xi, b] \) and sufficiently small \( h \in ]0, b - x[ \):

\[ \lambda_f(x, h) = y + h \varphi_f(x, y, h) - \phi(x + h) \]

and

\[ \lambda_g(x, h) = \psi(x) + h \varphi_g(\psi(x), h) - \psi(x + h) \\
= (x, \phi(x)) + h (1, \varphi_f(x, \phi(x), h)) - (x + h, \phi(x + h)) = (0, \lambda_f(x, h)). \]  \hspace{1cm} (2.75)

According to (a), the RK method is consistent of order \( p \) for the autonomous problem and we obtain

\[ \limsup_{h \downarrow 0} \frac{\|\lambda_g(x, h)\|}{|h|^{p+1}} \in \mathbb{R}_0^+ \quad (2.75) \Rightarrow \limsup_{h \downarrow 0} \frac{\|\lambda_f(x, h)\|}{|h|^{p+1}} \in \mathbb{R}_0^+, \]

showing the method to be consistent of order \( p \) for the nonautonomous problem as well.
We will construct the \( f_T \) inductively (the construction will actually show we can obtain each component of each \( f_T \) to be a monomial): We start the construction by setting \( f_0 : \mathbb{R} \rightarrow \mathbb{R} \), \( f_0(y) := 1 \). Then \( f_0^{(0)}(0) = f_0(0) = 1 \) and, for \( S \in \mathcal{T} \setminus \{\emptyset\} \), \( f_0^{(S)}(0) = 0 \), since \( f_0^{(S)} \) involves a derivative of order \( \geq 1 \) of \( f_0 \), and all such derivatives vanish identically. Now let \( d \in \mathbb{N} \), assume \( f_T \in C^\infty(\mathbb{R}^\#T, \mathbb{R}^\#T) \) with the property of (2.76) has already been constructed for each \( T \in \mathcal{T}_{\leq d} \), and let \( T = [T_1, \ldots, T_N] \in \mathcal{T}_{d+1}, N \in \mathbb{N} \). We need to define
\[
f_T : \mathbb{R}^\#T \rightarrow \mathbb{R}^\#T, \quad y \mapsto f_T(y).
\]
As the following definition does actually depend on the order of \( T_1, \ldots, T_N \), we fix an enumeration \( \alpha : \mathcal{T} \rightarrow \mathbb{N} \) and assume \( T_1, \ldots, T_N \) to be ordered according to \( \alpha \) (i.e. \( \alpha(T_1) \leq \alpha(T_2) \leq \cdots \leq \alpha(T_N) \)). We now partition
\[
y = (y_1, \ldots, y_\#T) = (y_1, y_1^1, \ldots, y_N^N), \quad \text{where } y_1 \in \mathbb{R}, \quad \forall j \in \{1, \ldots, N\} \quad y^i_j \in \mathbb{R}^\#j,
\]
and define
\[
f_T(y) := (y_1^1 \cdots y_1^N, f_{T_1}(y_1^1), \ldots, f_{T_N}(y_N^N)) \in \mathbb{R}^\#T.
\]
Then, clearly, each component of \( f_T \) is a monomial, since we know each component of each \( f_{T_j}(y^j) \) to be a monomial by induction. Since \( T \neq \emptyset \), we have
\[
\left( f_T^{(0)}(0) \right)_1 = (f_T(0))_1 = 0.
\]
Now let \( S = [S_1, \ldots, S_M] \in \mathcal{T}, M \in \mathbb{N} \). We compute
\[
\left( f_T^{(S)}(0) \right)_1 = \left( f_T^{(M)}(0) \left( f_T^{(S_1)}(0), \ldots, f_T^{(S_M)}(0) \right) \right)_1
\]
\[
= \sum_{j_1, \ldots, j_M = 1}^{\#T} \left( \partial_{j_1} \cdots \partial_{j_M} f_T(0) \right) \left( f_T^{(S_1)}(0) \right)_{j_1} \cdots \left( f_T^{(S_M)}(0) \right)_{j_M}
\]
\[
= \sum_{j_1, \ldots, j_M = 1}^{\#T} \partial_{j_1} \cdots \partial_{j_M} (y_1^1 \cdots y_N^N)(0) \left( f_T^{(S_1)}(0) \right)_{j_1} \cdots \left( f_T^{(S_M)}(0) \right)_{j_M}
\]
\[
= \begin{cases} 
\left( f_T^{(S_1)}(0) \right)_1 \cdots \left( f_T^{(S_N)}(0) \right)_1 & \text{for } M = N, \\
0 & \text{for } M \neq N
\end{cases}
\]
\[
= \delta_{S_1T_1} \cdots \delta_{S_NT_N} \quad \text{for } M = N, \\
0 \quad \text{for } M \neq N
\]
\[
= \delta_{ST},
\]
completing the proof of (2.76). Now let $p \in \mathbb{N}$. If $S \in T^{\leq p}$, then, due to (2.76), for $f := f_S$ and $(\xi, \eta) = (0, 0)$, (2.74) becomes

$$\lambda(0, h) = h^{\#S} \left( \frac{\alpha(S)}{S!} v_A(S) - \frac{\alpha(S)}{S!} v_A(S) + h R_{p,\varphi}(0, h) - R_{p,\phi}(0, h), \right)$$

where $\|R_{p,\varphi}(0, h)\| = O(|h|^p)$ for $|h| \to 0$ and $\|R_{p,\phi}(0, h)\| = O(|h|^{p+1})$ for $|h| \to 0$. Hence, if

$$\limsup_{h \downarrow 0} \frac{\|\lambda(0, h)\|}{|h|^{p+1}} \in \mathbb{R}_0^+,$$

then (as $\#S \leq p$)

$$b^t v_A(S) = \frac{1}{S!},$$

proving (2.76).

**Example 2.61.** Let $s \in \mathbb{N}$, $b^t = (b_1, \ldots, b_s) \in \mathbb{R}^s$, $A \in M(s, \mathbb{K})$ and consider an $s$-stage RK method with weights vector $b$ and RK matrix $A$. We use the results of Ex. 2.52 and Ex. 2.56 above, to formulate the consistency conditions (2.73) of up to order 4, where, as in Ex. 2.56, we define

$$\forall j \in \{1, \ldots, s\} c_j := \sum_{l=1}^s a_{jl}.$$ 

We already know from Rem. 2.59 that

$$\sum_{j=1}^s b_j = 1 \quad (2.77a)$$

constitutes the consistency condition of order 1. In the following, the notation for the trees is the same as in Ex. 2.52 and Ex. 2.56 above. Since $T_{21}$ is the only tree of order 2,

$$b^t v_A(T_{21}) = \sum_{j=1}^s b_j c_j = \frac{1}{2} \quad (2.77b)$$

is the only additional consistency condition of order 2. As the trees of order 3 are precisely $T_{31}$ and $T_{32}$, there are precisely two additional equations for the consistency condition of order 3, namely

$$b^t v_A(T_{31}) = \sum_{j,l=1}^s b_j a_{jl} c_l = \frac{1}{6}, \quad (2.77c)$$

$$b^t v_A(T_{32}) = \sum_{j=1}^s b_j c_j^2 = \frac{1}{3}. \quad (2.77d)$$
The four trees of order 4 yield precisely the following four additional equations for the consistency condition of order 4:

\[
\begin{align*}
    b^t v_A(T_{41}) &= \sum_{j,k,l=1}^s b_j a_{jk} a_{kl} c_l = \frac{1}{24}, \\
    b^t v_A(T_{42}) &= \sum_{j,l=1}^s b_j a_{jl} c_l^2 = \frac{1}{12}, \\
    b^t v_A(T_{43}) &= \sum_{j,l=1}^s b_j c_j a_{jl} c_l = \frac{1}{8}, \\
    b^t v_A(T_{44}) &= \sum_{j=1}^s b_j c_j^3 = \frac{1}{4}.
\end{align*}
\]

**Example 2.62.** (a) Both the explicit Euler method of Ex. 2.28(a) and the implicit Euler method of Ex. 2.28(c) are 1-stage RK methods with \( b = (1) \), showing they satisfy the consistency condition (2.77a) of order 1. The implicit trapezoidal method of Ex. 2.43(d) has Butcher tableau

\[
\begin{array}{ccc}
    0 & 0 & 0 \\
    1 & \frac{1}{2} & \frac{1}{2} \\
    \frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{array}
\]

Thus, (2.77a) is clearly satisfied. Moreover,

\[
\sum_{j=1}^s b_j c_j = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2},
\]

showing (2.77b) to be satisfied as well. On the other hand

\[
\sum_{j,l=1}^s b_j a_{jl} c_l = b_2 a_{21} c_1 + b_2 a_{22} c_2 = \frac{1}{2} \cdot \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} \cdot 1 = \frac{1}{4} \neq \frac{1}{6},
\]

i.e. (2.77c) fails.

(b) As stated before in Ex. 2.28(b), the classical explicit RK method has Butcher tableau

\[
\begin{array}{ccc}
    0 \\
    \frac{1}{2} & \frac{1}{2} \\
    \frac{1}{2} & 0 & \frac{1}{2} \\
    1 & 0 & 0 & 1
\end{array}
\]

We verify that it satisfies all 8 equations of (2.77), i.e. it satisfies the consistency condition of order 4:

\[
\sum_{j=1}^s b_j = \frac{1}{6} + \frac{1}{3} + \frac{1}{3} + \frac{1}{6} = 1,
\]
\[
\sum_{j=1}^s b_j c_j = 2 \cdot \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{6} \cdot 1 = \frac{1}{2}, \\
\sum_{j,l=1}^s b_j a_{jl} c_l = \frac{1}{3} \cdot \frac{1}{2} \cdot 0 + \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{6} \cdot \frac{1}{2} = 1, \\
\sum_{j=1}^s b_j c_j^2 = 2 \cdot \frac{1}{3} \cdot \frac{1}{4} + \frac{1}{6} \cdot 1^2 = \frac{1}{3}, \\
\sum_{j,k,l=1}^s b_j a_{jk} a_{kl} c_l = b_3 a_{32} a_{21} c_1 + b_4 a_{43} a_{32} c_2 = \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot 0 + \frac{1}{6} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{24}, \\
\sum_{j,l=1}^s b_j a_{jl} c_l^2 = \frac{1}{3} \cdot \frac{1}{2} \cdot 0^2 + \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{4} = \frac{1}{12}, \\
\sum_{j,l=1}^s b_j c_j a_{jl} c_l = \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot 0 + \frac{1}{3} \cdot \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{6} \cdot 1 \cdot \frac{1}{2} = \frac{1}{8}, \\
\sum_{j=1}^s b_j c_j^3 = 2 \cdot \frac{1}{3} \cdot \frac{1}{8} + \frac{1}{6} \cdot 1^3 = \frac{1}{4}.
\]

In combination with Th. 2.60(b), this, finally, provides the proof of Th. 2.23(a).

(c) Consider the following implicit 2-stage RK method with Butcher tableau

\[
\begin{array}{ccc}
\frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\
\frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\
\frac{1}{2} & \frac{1}{2} & \frac{1}{2}
\end{array}
\]

It is a so-called Gauss method. It is an example of a 2-stage method, satisfying the consistency condition of order 4: As in (b), we check the validity of all 8 equations of (2.77):

\[
\sum_{j=1}^s b_j = \frac{1}{2} + \frac{1}{2} = 1, \\
\sum_{j=1}^s b_j c_j = \frac{1}{2} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) + \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) = 2 \cdot \frac{1}{4} = \frac{1}{2}, \\
\sum_{j,l=1}^s b_j a_{jl} c_l = \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) + \frac{1}{2} \left( \frac{1}{4} + \frac{1}{4} - \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) = 2 \cdot \frac{1}{2} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) = \frac{1}{4} - \frac{3}{36} = \frac{1}{6}, \\
\sum_{j=1}^s b_j c_j^2 = \frac{1}{2} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right)^2 + \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right)^2 = 2 \cdot \frac{1}{2} \cdot \frac{1}{4} + 2 \cdot \frac{1}{2} \cdot \frac{3}{36} = \frac{1}{3},
\]

\[
\sum_{j,k,l=1}^{s} b_j a_{jk} a_{kl} c_l = \frac{1}{2} \sum_{j=1}^{s} a_{j1} \left( \frac{1}{4} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) + \left( \frac{1}{4} - \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) \right) \\
+ \frac{1}{2} \sum_{j=1}^{s} a_{j2} \left( \frac{1}{4} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) + \left( \frac{1}{4} + \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) \right) \\
= \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) \left( \frac{1}{4} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) + \left( \frac{1}{4} - \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) \right) \\
+ \frac{1}{2} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) \left( \frac{1}{4} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) + \left( \frac{1}{4} + \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) \right) \\
= \frac{1}{4} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right)^2 + \frac{1}{4} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right)^2 = \frac{1}{12},
\]

\[
\sum_{j,l=1}^{s} b_j a_{jl} c_l^2 = \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right)^2 + \frac{1}{2} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right)^2 = \frac{1}{12},
\]

\[
\sum_{j,l=1}^{s} b_j c_j a_{jl} c_l = \frac{1}{2} \cdot \frac{1}{4} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right)^2 + \frac{1}{2} \cdot \frac{1}{4} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right)^2 \\
+ \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right) \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right) \left( \frac{1}{4} + \frac{1}{4} \right) = \frac{1}{12} \cdot \frac{1}{8} + \frac{1}{12} = \frac{1}{8},
\]

\[
\sum_{j=1}^{s} b_j c_j^3 = \frac{1}{2} \left( \frac{1}{2} - \frac{\sqrt{3}}{6} \right)^3 + \frac{1}{2} \left( \frac{1}{2} + \frac{\sqrt{3}}{6} \right)^3 \\
= 2 \cdot \frac{1}{2} \cdot \frac{1}{8} + 2 \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{3}{36} = \frac{1}{4}.
\]

Thus, as the method, clearly, also satisfies the node condition, it is consistent of order 4 for each \( f \) that satisfies the hypotheses of Th. 2.60(a) or Th. 2.60(b).

### 2.4 Extrapolation

Extrapolation is a method that can be used to improve the accuracy of single-step methods. As before, the goal is to approximate the solution to the initial value problem (2.9), i.e. \( y' = f(x, y), \quad y(\xi) = \eta \), which is assumed to have a unique exact solution \( \phi \) defined on some interval \([\xi, b]\), \( b > \xi \). We know that, given a defining function \( \varphi \) and a partition \( \Delta = (x_0, \ldots, x_N) \) of \([\xi, b]\), we obtain a sequence \((y_0, y_1, \ldots)\) of approximations to \((\phi(x_0), \phi(x_1), \ldots)\).

Assume we have a defining function \( \varphi \) that provides an explicit single-step method, being well-defined method in the sense of Def. 2.12(a). We now fix \( x \in [\xi, b] \) and are interested in approximating \( \phi(x) \). We consider partitions of \( \Delta^h \) of \([\xi, x]\) of decreasing (sufficiently small) equidistant stepsizes \( h \). If \((h_j)_{j \in \mathbb{N}_0}\) is a sequence of admissible stepsizes such that
lim_{j \to \infty} h_j = 0$, then, for a reasonable method (one of order of convergence $p \in \mathbb{N}$, say), we expect $\lim_{j \to \infty} y_{h_j}(x) = \phi(x)$, where $y_{h_j}(x)$ is the approximation at $x$ given by the partition $\Delta^h$ of $[\xi, x]$ that has equidistant stepsizes $h_j$ (only discrete values for $h_j > 0$ are admissible, namely those satisfying $\frac{x - \xi}{h_j} \in \mathbb{N}$).

The idea of extrapolation is to improve the approximation at $x$, by computing $y_{h_j}(x)$ for a number of different stepsizes $h_0 > h_1 > \cdots > h_M$, interpolating the points $(h_0, y_{h_0}(x)), \ldots, (h_M, y_{h_M}(x))$ via polynomial interpolation to obtain a polynomial $P$, $h \mapsto P(h)$, using $P(0)$ as the improved approximation of $\phi(x)$. The term extrapolation comes from the fact that $0 \notin [h_M, h_0]$, i.e., one extrapolates $P$ to some value outside the interval of interpolation points.

One can only expect the described idea to work well if $y_h(x)$ (as a function of $h$) behaves like a polynomial, at least asymptotically for $h \to 0$. This is the case, if the error $y_h(x) - \phi(x)$ has an asymptotic expansion in the sense of Def. 2.64 below.

**Notation 2.63.** Let $\xi, x \in \mathbb{R}$, $\xi < x$. Define the set of admissible stepsizes by

$$H_x := \left\{ h \in \mathbb{R}^+ : \frac{x - \xi}{h} \in \mathbb{N} \right\}.$$  

For each $h \in H_x$, let $N(h) := \frac{x - \xi}{h} \in \mathbb{N}$, i.e., $N(h)$ is the number of steps in the partition $\Delta^h \in \Pi([\xi, x])$ consisting of the $N(h) + 1$ equidistant points

$$x^h_k = \xi + kh, \quad k \in \{0, \ldots, N(h)\},$$

that means

$$\Delta^h = (x^h_0, x^h_1, \ldots, x^h_{N(h)}) = (\xi, \xi + h, \ldots, \overbrace{\xi + hN(h)}^{=x}).$$

**Definition 2.64.** In the situation of Def. 2.12, assume the explicit single-step method given by the defining function $\varphi$ to be well-defined, where $h(\varphi) \in \mathbb{R}^+$ is as in Def. 2.12(a) (i.e. the usual recursion provides all approximations, provided the partition $\Delta$ of $[\xi, b]$ has mesh size $h_{\max}(\Delta) < h(\varphi)$). Let $x \in [\xi, b]$. For each $h \in H_x \cap ]0, h(\varphi)[$, denote

$$y^h_0 := \eta,$$

$$\forall k \in \{0, \ldots, N(h) - 1\} \quad y^h_{k+1} = y^h_k + h \varphi(x^h_k, y^h_k, h).$$

Also define

$$U := \left\{ (x, h) \in [\xi, b] \times ]0, h(\varphi)[ : h \in H_x \right\}$$

and let

$$u : U \longrightarrow \mathbb{K}^n, \quad u(x, h) := y^h_{N(h)}.$$

We say that the method given by $\varphi$ admits an asymptotic expansion of the (global) error $u(x, h) - \phi(x)$ if, and only if, there exist $p, r \in \mathbb{N}$ such that

$$\forall j \in \{0, \ldots, r - 1\} \quad \exists_{\epsilon_{p+j} \in \mathbb{C}} \quad \lim_{h \to 0} \sup_{(x, h) \in U} \left\| u(x, h) - \phi(x) - \sum_{j=0}^{r-1} \epsilon_{p+j}(x) h^{p+j} \right\|_{h^{p+r}} =: C \in \mathbb{R}_{\geq 0}.$$  

(2.78a)
where the convergence in (2.78a) is often stated with a Landau symbol in the form
\[ u(x, h) - \phi(x) = c_p(x)h^p + c_{p+1}(x)h^{p+1} + \cdots + c_{p+r-1}(x)h^{p+r-1} + O(h^{p+r}) \]
(for \( h \to 0 \), uniformly in \( x \in [\xi, b] \)).

(2.78b)

**Remark 2.65.** Remaining in the setting of the previous Def. 2.64, we consider a finite sequence (i.e. a vector)
\[ \vec{h} := (h_0, \ldots, h_M) \in \left( H_x \cap [0, h(\phi)] \right)^{M+1}, \quad \varphi(h) > h_0 > \cdots > h_M > 0, \quad M \in \mathbb{N}, \]
of admissible stepsizes. Then, for each \( p \in \mathbb{N} \), there exists a unique (\( \mathbb{K}^n \)-valued) polynomial
\[ P := P(\varphi, x, \vec{h}, p), \quad P : \mathbb{R} \to \mathbb{K}^n, \]
\[ P(h) = P(\varphi, x, \vec{h}, p)(h) = a_0 + \sum_{j=0}^{M-1} a_{p+j}h^{p+j}, \quad a_0, a_p, \ldots, a_{p+M-1} \in \mathbb{K}^n, \]
satisfying the interpolation conditions
\[ \forall j \in \{0, \ldots, M\} \quad P(h_j) = u(x, h_j), \quad (2.79) \]
where \( u \) is the function from Def. 2.64: Indeed, this follows by componentwise Hermite interpolation: Let \( \alpha \in \{1, \ldots, n\} \) and let \( \text{Op} \) stand for either \( \text{Re} \) (real part) or \( \text{Im} \) (imaginary part). Then, for the component \( Q := \text{Op} P_\alpha \) of \( P \), (2.79) yields the \( M+1 \) conditions \( Q(h_j) = \text{Op} u_\alpha(x, h_j) \). In addition, \( Q \) also satisfies the \( p-1 \) conditions
\[ \forall j \in \{1, \ldots, p-1\} \quad Q^{(j)}(0) = 0. \]
Thus, by (1-dimensional) Hermite interpolation (cf. [Phi17b, Th. 3.25]), \( Q \) is the unique polynomial of degree at most \( p + M - 1 \), satisfying these \( p + M \) conditions.

**Theorem 2.66** (Asymptotic Expansion of the (Global) Error). In the situation of Def. 2.12, let \( f : G \to \mathbb{K}^n \) be defined on an open set \( G \subseteq \mathbb{R} \times \mathbb{K}^n \); let \( \varphi : \mathcal{D}_\varphi \to \mathbb{K}^n \) be defined on a suitable open set \( \mathcal{D}_\varphi \subseteq \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+ \) such that, in particular, the local truncation error \( \lambda \) is well-defined. Furthermore assume:

(i) The method given by \( \varphi \) is consistent of order \( p \) for some \( p \in \mathbb{N} \).
(ii) \( \varphi \) is globally Lipschitz with respect to \( y \).
(iii) \( f \in C^{p+r}(G, \mathbb{K}^n) \) and \( \varphi \in C^{p+r}(\mathcal{D}_\varphi, \mathbb{K}^n) \) for some \( r \in \mathbb{N} \).

Then the method given by \( \varphi \) admits an asymptotic expansion of the (global) error \( u(x, h) - \phi(x) \) in the sense of Def. 2.64.

**Proof.** The proof needs some work and we will not have the time to study it in this class. A proof can be found, e.g., in [Pla06, Sec. 7.5].
Corollary 2.67. In the situation of Def. 2.12, let \( f : G \to \mathbb{K}^n \) be defined on an open set \( G \subseteq \mathbb{R} \times \mathbb{K}^n \); let \( \varphi : D_\varphi \to \mathbb{K}^n \) be defined on a suitable open set \( D_\varphi \subseteq \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+ \) and assume \( \varphi \) defines an explicit RK method. If \( f \in C^{p+r}(G, \mathbb{K}^n) \) with \( p, r \in \mathbb{N} \) is globally Lipschitz with respect to \( y \) and such that the method is consistent of order \( p \), then \( \varphi \) admits an asymptotic expansion of the (global) error \( u(x, h) - \phi(x) \) in the sense of Def. 2.64.

Proof. One merely needs to combine Th. 2.66 with Rem. 2.29(b),(c). \( \blacksquare \)

Theorem 2.68. In the situation of Def. 2.12, assume the explicit single-step method given by the defining function \( \varphi \) to be well-defined. We will use the same notation as in Def. 2.64 and in Rem. 2.65; in particular, we consider the function \( u \) as in Def. 2.64 and the polynomial \( P \) as in Rem. 2.65. If the method given by \( \varphi \) admits an asymptotic expansion of the (global) error according to Def. 2.64 and we take some

\[
\tilde{N} := (N_0, \ldots, N_M) \in \mathbb{N}^{M+1}, \quad N_0 < \cdots < N_M, \quad 1 \leq M \leq r, \tag{2.80a}
\]

which, for each \( h \in H_\varphi \cap [0, h(\varphi)] \), yields some

\[
\tilde{h}(\tilde{N}) := (h_0, \ldots, h_M) \in \left( H_\varphi \cap [0, h(\varphi)] \right)^{M+1}, \quad h_j := h/N_j, \tag{2.80b}
\]

as in Rem. 2.65, then there exist coefficients \( b_j \in \mathbb{R}, \ j \in \{p + M, \ldots, p + r - 1\} \), not depending on \( \varphi, x, \) and \( h \) (but, in general, depending on \( \tilde{N} \) and \( p \)) such that the error between \( P(0) \) and \( \phi(x) \) can be written in the form

\[
P(\varphi, x, \tilde{h}(\tilde{N}), p)(0) - \phi(x) = \sum_{j=p+M}^{p+r-1} b_j c_j(x) h^j + O(h^{p+r}) \tag{2.81}
\]

(for \( h \to 0 \), uniformly in \( x \in [\xi, b] \)),

where the precise meaning of the convergence in (2.81) is the same as in (2.78).

Proof. It suffices to show (2.81) for each component \( \alpha \in \{1, \ldots, n\} \), as long as the resulting \( b_j \) do not depend on \( \alpha \). Thus, let \( \alpha \in \{1, \ldots, n\} \). Then the interpolation condition (2.79) can be written as a linear system in matrix-vector form as

\[
\begin{pmatrix}
1 & 1/N_0^p & 1/N_0^{p+1} & \cdots & 1/N_0^{p+M-1} \\
1 & 1/N_1^p & 1/N_1^{p+1} & \cdots & 1/N_1^{p+M-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 1/N_M^p & 1/N_M^{p+1} & \cdots & 1/N_M^{p+M-1}
\end{pmatrix}
\begin{pmatrix}
ad_{0\alpha} \\
ad_{p,\alpha} h^p \\
ad_{p+1,\alpha} h^{p+1} \\
\vdots \\
ad_{p+M-1,\alpha} h^{p+M-1}
\end{pmatrix}
= A_M \in \mathcal{M}(M+1, \mathbb{R})
\begin{pmatrix}
a \alpha(x, h_0) \\
\vdots \\
a \alpha(x, h_M)
\end{pmatrix}.
\tag{2.82a}
\]

As we know this linear system to have a unique solution according to Rem. 2.65, the matrix \( A_M \) must be invertible. On the other hand, we have the asymptotic expansion
(2.78b) and evaluating component $\alpha$ of the expression in (2.78b) at the $h_j$ yields, in matrix-vector form,

\[
A_M \begin{pmatrix} 
\phi_\alpha(x) \\
c_{p+1,\alpha}(x) h^{p+1} \\
\vdots \\
c_{p+M-1,\alpha}(x) h^{p+M-1}
\end{pmatrix} = \begin{pmatrix} 
u_\alpha(x,h_0) \\
u_\alpha(x,h_1) \\
\vdots \\
u_\alpha(x,h_M)
\end{pmatrix} - r_\alpha(x,h),
\] (2.82b)

where

\[
r_\alpha(x,h) = \sum_{j=p+M}^{p+r-1} \begin{pmatrix} 1/N_0^j \\
1/N_1^j \\
\vdots \\
1/N_M^j
\end{pmatrix} c_{j\alpha}(x) h^j + O(h^{p+r})
\]

(for $h \to 0$, uniformly in $x \in ]\xi, b]$).

Subtracting (2.82b) from (2.82a) yields

\[
A_M \begin{pmatrix} 
(a_{0\alpha} - \phi_\alpha(x) \\
(a_{p+1,\alpha} - c_{p+1,\alpha}(x)) h^{p+1} \\
\vdots \\
(a_{p+M-1,\alpha} - c_{p+M-1,\alpha}(x)) h^{p+M-1}
\end{pmatrix} = r_\alpha(x,h).
\]

We now multiply the above equation by $A_M^{-1}$ and claim that the first row of the resulting equation completes the proof of (2.81): Indeed, we obtain

\[
P_\alpha(\varphi, x, \tilde{h}(\tilde{N}), p)(0) - \phi_\alpha(x) = a_{0\alpha} - \phi_\alpha(x) = \sum_{l=0}^{M} \sum_{j=p+M}^{p+r-1} a_{M,1l}^{-1} c_{j\alpha}(x) h^j + O(h^{p+r})
\]

(for $h \to 0$, uniformly in $x \in ]\xi, b]$)

and, setting

\[
\forall j \in \{p+M, \ldots, p+r-1\} \quad b_j := \sum_{l=0}^{M} a_{M,1l}^{-1} N_l^j,
\]

we have (2.81) (noting the $b_j$ to be independent of $\alpha, \varphi, x$, and $\tilde{h}$).

\begin{remark}
Note that, according to (2.81), the approximation given by the extrapolation $P_\alpha(\varphi, x, \tilde{h}(\tilde{N}), p)(0)$ now has order of convergence $p + M$ in the sense that

\[
\exists C > 0 \quad \forall x \in ]\xi, b] \quad \lim_{h \to 0} \frac{\|P_\alpha(\varphi, x, \tilde{h}(\tilde{N}), p)(0) - \phi(x)\|}{h^{p+M}} = \lim_{h \to 0} \frac{\|\sum_{j=p+M}^{p+r-1} b_j c_j(x) h^j\|}{h^{p+M}} = |b_{p+M}| \|c_{p+M}(x)\| \leq C.
\]
\end{remark}
Remark 2.70. Among others, the following rules have been applied in the literature to obtain decreasing sequences \((h_j)_{j \in \mathbb{N}_0}\) of stepsizes for extrapolation (cf. Rem. 2.65 and (2.80)). All sequences start with a given base stepsize \(h_0 := h > 0\) and then apply different rules:

**Harmonic Sequence:**
\[
\forall j \in \mathbb{N} \quad h_j := \frac{h_0}{j+1}.
\]

**Romberg Sequence:**
\[
\forall j \in \mathbb{N} \quad h_j := \frac{h_{j-1}}{2}.
\]

**Bulirsch Sequence:**
\[
h_1 := \frac{h_0}{2}, \quad h_2 := \frac{h_0}{3}, \quad h_3 := \frac{h_0}{4}, \quad \forall j \geq 4 \quad h_j := \frac{h_{j-2}}{2}.
\]

In principle, one can carry out Hermite interpolation to obtain the interpolating polynomial \(P\) of Rem. 2.65. Alternatively, one can obtain the \(a_0, a_p, \ldots, a_{p+M-1}\) from the linear systems (2.82a). However, as we are only interested in the value \(P(0)\), if possible, it is more efficient to compute \(P(0)\) directly, without computing the polynomial’s coefficients. For \(p = 1\), one can achieve this using a so-called Neville tableau, making use of the following Th. 2.72.

**Notation 2.71.** For \(n \in \mathbb{N}_0\), we denote the set of all polynomials \(P : \mathbb{R} \rightarrow \mathbb{R}\) of degree at most \(n\) by \(\mathcal{P}_n\).

**Theorem 2.72.** Let \(M \in \mathbb{N}\) and consider points
\[
(x_0, y_0), (x_1, y_1), \ldots, (x_M, y_M) \in \mathbb{R}^2,
\]
where \(x_0, \ldots, x_M\) are all distinct. For \(k, m \in \mathbb{N}_0\) with \(k + m \leq M\), we let
\[
P_{k,k+1,\ldots,k+m} \in \mathcal{P}_m
\]
denote the unique interpolating polynomial of degree \(\leq m\), satisfying the conditions
\[
\forall j \in \{k, \ldots, k+m\} \quad P_{k,k+1,\ldots,k+m}(x_j) = y_j. \tag{2.83}
\]

Then, for each \(x \in \mathbb{R}\), the following recursion holds:
\[
\forall k \in \{0, \ldots, M\} \quad P_k(x) = y_k, \tag{2.84a}
\]
\[
\forall k, m \in \mathbb{N}_0, \quad k + m \leq M, \quad m \geq 1 \quad P_{k,k+1,\ldots,k+m}(x) = \frac{(x - x_k)P_{k+1,\ldots,k+m}(x) - (x - x_{k+m})P_{k,\ldots,k+m-1}(x)}{x_{k+m} - x_k}. \tag{2.84b}
\]

As one can apply the recursion (2.84) componentwise, it still holds if the \(y_j\) and the polynomials are \(\mathbb{R}^n\)-valued, \(n \in \mathbb{N}\).
Proof. For each \( k \in \{0, \ldots, M\}, \) \( P_k \in \mathcal{P}_0 \) is constant and satisfies \( P_k(x_k) = y_k \), proving (2.84a). Now let \( k, m \) satisfy the conditions of (2.84b). Denoting the polynomial given by the right-hand side of (2.84b) by \( Q \), we have to show \( P_{k,k+1,\ldots,k+m} = Q \). Clearly, both \( P_{k,k+1,\ldots,k+m} \) and \( Q \) are in \( \mathcal{P}_m \). Thus, the proof is complete, if we can show that \( Q \) also satisfies the conditions (2.83), that means

\[
\forall \ j \in \{k, \ldots, k+m\} \quad Q(x_j) = y_j.
\]  

(2.85)

To verify (2.85), we compute

\[
Q(x_k) = \frac{0 - (x_k - x_{k+m})P_{k,k+1,\ldots,k+m-1}(x_k)}{x_{k+m} - x_k} = y_k,
\]

\[
Q(x_{k+m}) = \frac{(x_{k+m} - x_k)P_{k+1,\ldots,k+m}(x_{k+m}) - 0}{x_{k+m} - x_k} = y_{k+m},
\]

and, for each \( j \in \{k+1, \ldots, k+m-1\}, \)

\[
Q(x_j) = \frac{(x_j - x_k)y_j - (x_j - x_{k+m})y_j}{x_{k+m} - x_k} = y_j,
\]

proving that (2.85) does, indeed, hold.

Remark 2.73. The recursion (2.84) can be organized and visualized in a so-called Neville tableau as follows:

\[
y_0 = P_0(x) \quad \downarrow \quad P_0(x) \\
y_1 = P_1(x) \quad \rightarrow \quad P_{01}(x) \\
y_2 = P_2(x) \quad \rightarrow \quad P_{12}(x) \quad \rightarrow \quad P_{012}(x) \\
\vdots \\
y_{M-1} = P_{M-1}(x) \quad \rightarrow \quad P_{M-2,M-1}(x) \quad \rightarrow \quad \ldots \quad \ldots \quad P_{0\ldots M-1}(x) \\
y_M = P_M(x) \quad \rightarrow \quad P_{M-1,M}(x) \quad \rightarrow \quad \ldots \quad \ldots \quad P_{1\ldots M}(x) \quad \rightarrow \quad P_{0\ldots M}(x)
\]

Example 2.74. Suppose, we want to improve a single-step method given by some \( \varphi \) at \( x = b \) by interpolation with \( p = 1 \) and \( M = 2 \), using the Romberg sequence according to Rem. 2.70 for some sufficiently small and admissible \( h > 0 \). We have to find \( P(0) = P_{012}(0) \), where \( P \in \mathcal{P}_2 \) satisfies conditions (2.79), which, in the current situation, are

\[
\forall \ j \in \{0,1,2\} \quad P(h_j) = u(b, h_j) = y_{h_j}^N(b_j), \quad N(h_j) = (b - \xi)/h_j.
\]

We obtain \( P(0) \) from a Neville tableau as follows:

\[
u(b, h) = P_0(0) \quad \downarrow \\
u(b, h/2) = P_1(0) \quad \rightarrow \quad P_{01}(0) = \frac{-hP_1(0) + \frac{h}{2}P_0(0)}{\frac{h}{2}} = 2u(b, h/2) - u(b, h) \quad \downarrow \\
u(b, h/4) = P_2(0) \quad \rightarrow \quad P_{12}(0) = \frac{-\frac{h}{2}P_2(0) + \frac{h}{4}P_1(0)}{\frac{h}{4}} = 2u(b, h/4) - u(b, h/2) \quad \rightarrow
\]
\[ P_{012}(0) = \frac{-hP_{12}(0) + \frac{1}{3} P_{01}(0)}{h - \frac{h}{4}} = \frac{4}{3} \left( 2u(b, h/4) - u(b, h/2) \right) - \frac{1}{3} \left( 2u(b, h/2) - u(b, h) \right) \]
\[ = \frac{1}{3} \left( 8u(b, h/4) - 6u(b, h/2) + u(b, h) \right). \]

If Th. 2.68 applies with \( r = M = 2 \), then (2.81) yields
\[ P(0) - \phi(b) = O(h^3) \quad \text{for} \quad h \to 0. \]

**Example 2.75.** Suppose, we want to improve a single-step method given by some \( \varphi \) at \( x \in [\xi, b] \) by interpolation with \( p \geq 1 \) and \( M = 1 \), using some sufficiently small and admissible \( h > 0 \), letting \( h_0 := h_1 := h_0/N, N \in \mathbb{N}, N \geq 2 \). We have to find \( P(0) \), where \( P \in P_p \) satisfies conditions (2.79), which, in the current situation, are
\[ P(h) = u(x, h), \quad P(h/N) = u(x, h/N). \]

For each \( \alpha \in \{1, \ldots, n\} \), we can obtain \( a_{0\alpha} \) and \( a_{p\alpha} \) from (2.82a), which, in the current situation, reads
\[ a_{0\alpha} + a_{p\alpha} h^p = u_{\alpha}(x, h) \]
\[ a_{0\alpha} + \frac{a_{p\alpha} h^p}{N^p} = u_{\alpha}(x, h/N) \]

with the solution
\[ P_{\alpha}(0) = a_{0\alpha} = u_{\alpha}(x, h/N) + \frac{u_{\alpha}(x, h/N) - u_{\alpha}(x, h)}{N^p - 1}, \quad a_{p\alpha} = \frac{u_{\alpha}(x, h) - u_{\alpha}(x, h/N)}{h^p (1 - \frac{1}{N^p})}. \]

If Th. 2.68 applies with \( r = 2 \), then (2.81) yields
\[ P(0) - \phi(x) = b_{p+1} c_{p+1}(x) h^{p+1} + O(h^{p+2}) \quad \text{for} \quad h \to 0. \]

### 2.5 Stepsize Control

As before, let us consider a well-defined explicit single-step method for the initial value problem (2.9), i.e. \( y' = f(x, y), y(\xi) = \eta \), given via a defining function \( \varphi \). We know that, given a sequence of admissible stepsizes \( (h_k)_{k \in \mathbb{N}} \) and setting \( x_{k+1} := x_k + h_k \), \( \varphi \) provides a sequence of approximations \( (y_k)_{k \in \mathbb{N}} \) according to (2.14).

We are now interested in using *adaptive* stepsizes \( h_k \). More precisely, we want to vary the stepsizes \( h_k \), aiming at keeping the error in a prescribed range (reducing \( h_k \) if the error is too large, enlarging \( h_k \) (e.g. to make the computation more efficient), if the error is smaller than required. In general, this is a difficult problem and often involves the use of heuristics. Many different variants exist in the literature. Here, we follow [Pla06, Sec. 7.7]. The idea is to modify (2.14) such that, instead of using \( h_k \) in the \( k \)th step, we use \( h_k/2 \) twice. Then we use an extrapolation as in the previous section to estimate the
local error, and adjust $h_k$ if necessary. We will now provide the details to carry out this idea. The described modification of (2.14) leads to the recursion

$$\forall k \in \{0, 1, \ldots\} \quad w_k := y_k + \frac{h_k}{2} \varphi\left(x_k, y_k, \frac{h_k}{2}\right),$$

$$\forall k \in \{0, 1, \ldots\} \quad y_{k+1} := w_k + \frac{h_k}{2} \varphi\left(x_k + \frac{h_k}{2}, w_k, \frac{h_k}{2}\right), \quad x_{k+1} = x_k + h_k.$$  

As before, we assume the initial value problem under consideration to have a unique exact solution $\phi$, which we assume to be defined at the current point $x_k \geq \xi$. Given an approximation $y_k \approx \phi(x_k)$, the desired new stepsize $h_k$ should be such that

$$\|y_{k+1} - \psi(x_k + h_k)\| \approx \epsilon,$$  

(2.86)

where $\|\cdot\|$ denotes some arbitrary norm on $\mathbb{K}^n$, $\epsilon > 0$ is given, and $\psi$ is the exact solution to the initial value problem

$$y' = f(t, x), \quad y(x_k) = y_k,$$

which is assumed to be unique and defined at $x_k + h_k$. As $\psi$ is not known, one will have to use a numerical approximation of $\psi$ to test the validity of (2.86). Thus, obtaining the new stepsize $h_k$ involves iteration, starting with some initial guess $h_k^{(0)}$ (where one would use $h_k^{(0)} := h_{k-1}$ for $k \geq 1$ and $h_k^{(0)}$ can be set according to Rem. 2.77(a) below):

(a) Compute

$$w_k^{(l)} := y_k + \frac{h_k^{(l)}}{2} \varphi\left(x_k, y_k, \frac{h_k^{(l)}}{2}\right),$$

$$y_{k+1}^{(l)} := w_k^{(l)} + \frac{h_k^{(l)}}{2} \varphi\left(x_k + \frac{h_k^{(l)}}{2}, w_k^{(l)}, \frac{h_k^{(l)}}{2}\right).$$

(b) Compute an estimation of the error

$$\delta_k^{(l)} \approx \|y_{k+1}^{(l)} - \psi(x_k + h_k^{(l)})\|$$

(this can be done by an extrapolation according to (2.87) below). Stop the iteration and return $h_k := h_k^{(l)}$ if

$$c_1 \epsilon \leq \delta_k^{(l)} \leq c_2 \epsilon,$$

where $0 < c_1 < 1 < c_2$ are prescribed constants.

(c) Set $h_k^{(l+1)} < h_k^{(l)}$ if $\delta_k^{(l)} > c_2 \epsilon$ and $h_k^{(l+1)} > h_k^{(l)}$ if $\delta_k^{(l)} < c_1 \epsilon$, using some suitable rule (a possible rule will be given in (2.90) below). Proceed to (a) for the next iteration step.
We now describe how to obtain $\delta_k^{(l)}$ for (b): First we approximate $z_k \approx \psi(x_k + h_k^{(l)})$, using an extrapolation as in Ex. 2.75 with $h_0 := h_k^{(l)}$ and $h_1 := \frac{h_0}{2}$. In the present situation, the formula for $P(0)$ of Ex. 2.75 becomes

$$z_k = P(0) = y_{k+1}^{(l)} + \frac{y_{k+1}^{(l)} - y_k^{(l)}}{2^p - 1}, \quad v_k^{(l)} := y_k + h_k^{(l)} \varphi(x_k, y_k, h_k^{(l)}),$$

and, thus,

$$\delta_k^{(l)} = \|y_{k+1}^{(l)} - z_k\| = \left\| \frac{y_{k+1}^{(l)} - v_k^{(l)}}{2^p - 1} \right\|. \quad (2.87)$$

The rule for obtaining $h_{(l+1)}^k$ in (c) above can be based on the following result:

**Lemma 2.76.** If Th. 2.68 applies with $p \in \mathbb{N}$ and $r = 2$ to $\varphi$ and the initial value problem $y' = f(x, y), y(x_k) = y_k$, then, writing

$$w(h) := y_k + \frac{h}{2} \varphi(x_k, y_k, \frac{h}{2}),$$

$$u(h) := w(h) + \frac{h}{2} \varphi(x_k + \frac{h}{2}, w(h), \frac{h}{2}),$$

and noting that, in (2.87), $\delta_k^{(l)} = \delta_k^{(l)}(h_k^{(l)})$ can be seen as a function of $h_k^{(l)}$,

$$\|u(h) - \psi(x_k + h)\| = \left( \frac{h}{h_k^{(l)}} \right)^{p+1} \delta_k^{(l)}(h_k^{(l)}) + O\left( (h_k^{(l)})^{p+2} \right) \quad (2.88)$$

(for $h, h_k^{(l)} \to 0$ with $0 < h \leq h_k^{(l)}$).

**Proof.** See [Pla06, Lem. 7.35].

If Lem. 2.76 applies and

$$\delta_k^{(l)} \approx \epsilon \gg (h_k^{(l)})^{p+2} \quad \text{i.e.} \quad h_k^{(l)} \ll \epsilon^{-\frac{1}{p+2}}, \quad (2.89)$$

then it makes sense to use (2.88) to obtain, by neglecting the remainder term,

$$h_k^{(l+1)} := \left( \frac{\epsilon}{\delta_k^{(l)}} \right)^{\frac{1}{p+1}} h_k^{(l)} \quad (2.90)$$

for the new test stepsize in (c) above.

**Remark 2.77.** (a) In view of (2.89), one might choose for the initial test stepsize $h_0^{(l)} := \alpha \epsilon^{-\frac{1}{p+2}}$ with $0 < \alpha < 1$, e.g. $\alpha = \frac{1}{2}$ or $\alpha = \frac{1}{10}$.

(b) In general, there is no guarantee that the above iteration for finding the new stepsize $h_k$ terminates. It terminates if, and only if, the condition in (b) is satisfied for some $l \in \mathbb{N}_0$. In practice, one would set some bound $L \in \mathbb{N}$ for $l$ and stop the iteration for $l = L$ even if the condition in (b) fails to hold. One might then stop the method, as the stepsize control has failed, or one might revert to some default value for $h_k$, depending on some criterion for the severity of the failure.
2.6 Stiff Equations

In Rem. 2.8(b), we discussed the relation between explicit and implicit methods. We mentioned that methods given in explicit form are typically computationally easier and are usually preferred, unless the ODE to be solved defies solution by explicit methods (which means that an accurate solution by explicit methods needs stepsizes that are unacceptably small). Such ODE, as also already mentioned in Rem. 2.8(b), are usually called stiff ODE. A mathematically precise definition of the notion of stiff ODE seems difficult, and the literature does not seem to have reached a consensus on this. Here, we will follow [Pla06, Sec. 8.9].

For simplicity, in this section, we will restrict ourselves to initial value problems, where the right-hand side is \( \mathbb{R}^n \)-valued and defined on the entire space, i.e.

\[
y' = f(x, y), \quad f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n, \\
y(\xi) = \eta, \quad \eta \in \mathbb{R}^n.
\]  

(2.91a)  

(2.91b)

Now, typically, (2.91) is stiff if there exists an equilibrium function \( \psi_e : I \to \mathbb{R}^n \) such that every solution \( \phi : I \to \mathbb{R}^n \) to (2.91), \( \xi \in I \), approaches \( \psi_e \) rapidly in the sense that \( \phi(x) \approx \psi_e(x) \) for each \( x > \xi + \epsilon \) \( (x \in I) \) with a small \( \epsilon > 0 \).

The following Def. 2.78 is an attempt at putting the notion of a stiff ODE into a somewhat more precise form. It is basically reproduced from [Pla06, Sec. 8.9.1].

**Definition 2.78.** Let \( \langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) denote a fixed scalar product on \( \mathbb{R}^n \), \( n \in \mathbb{N} \), with induced norm \( \| \cdot \| : \mathbb{R}^n \to \mathbb{R}_+^+ \). Let \( \xi \in \mathbb{R} \) and let \( I \subseteq \mathbb{R} \) be a (nontrivial) interval with \( \xi = \min I \).

(a) Then \( f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n \) (and the initial value problem (2.91)) satisfies an upper Lipschitz condition with respect to \( y \) and \( \langle \cdot, \cdot \rangle \) on \( I \) if, and only if, there exists a continuous function \( M : I \to \mathbb{R} \) such that

\[
\forall (x, y), (x, \bar{y}) \in I \times \mathbb{R}^n \quad \langle f(x, y) - f(x, \bar{y}), y - \bar{y} \rangle \leq M(x) \| y - \bar{y} \|^2.
\]  

(2.92)

Moreover, the initial value problem is called dissipative on \( I \) if, and only if, (2.92) holds with \( M \leq 0 \) (in the sense that \( M(x) \leq 0 \) for each \( x \in I \)).

(b) The initial value problem (2.91) is stiff provided that it satisfies the following two conditions (i) and (ii):

(i) The problem is dissipative or, at least, (2.92) holds with an \( M \) that does not surpass a moderate positive size, say \( M \leq 1 \).

(ii) The expression on the left-hand side of (2.92) divided by \( \| y - \bar{y} \|^2 \) can become strongly negative, i.e.

\[
\forall x \in I \\
m(x) := \inf \left\{ \frac{\langle f(x, y) - f(x, \bar{y}), y - \bar{y} \rangle}{\| y - \bar{y} \|^2} : y, \bar{y} \in \mathbb{R}^n, \ y \neq \bar{y} \right\} \ll 0.
\]
Remark 2.79. Note that the Cauchy-Schwarz inequality implies
\[\forall (x,y),(\bar{x},\bar{y}) \in \mathbb{R} \times \mathbb{R}^n, \quad \frac{|\langle f(x,y) - f(x,\bar{y}), y - \bar{y} \rangle|}{\|y - \bar{y}\|^2} \leq \frac{\|f(x,y) - f(x,\bar{y})\|}{\|y - \bar{y}\|}.\]
In consequence, the condition of Def. 2.78(b)(ii) means that \(f\) can be globally \(L\)-Lipschitz with respect to \(y\) only with a very large Lipschitz constant \(L \geq |m(x)|\). In particular, for the explicit Euler method and the explicit classical Runge-Kutta method, the constant \(K\) provided by (2.27) becomes very large (cf. Th. 2.20 and Th. 2.23), such that reasonable approximations from these methods can only be expected for exceedingly small stepsizes \(h\).

Example 2.80. Let \(\lambda \in \mathbb{R}\). We consider (2.91) with \(n = 1\), \(\xi = 0\), and \(f = f_\lambda\), where
\[f_\lambda : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \quad f_\lambda(x,y) := \lambda y - (1 + \lambda)e^{-x},\]
i.e. the initial value problem
\[
y' = \lambda y - (1 + \lambda)e^{-x}, \quad y(0) = \eta, \quad \eta \in \mathbb{R}. \tag{2.93a}
\]
To compare with Def. 2.78 and Rem. 2.79, we compute (using the Euclidean scalar product, which is just multiplication in one dimension)
\[
\forall (x,y),(\bar{x},\bar{y}) \in \mathbb{R}^2 \quad \langle f_\lambda(x,y) - f_\lambda(x,\bar{y}), y - \bar{y} \rangle = \lambda(y - \bar{y})^2 = \lambda\|y - \bar{y}\|^2,
\]
obtaining \(M(x) = m(x) = \lambda\). Thus Def. 2.78(b) is satisfied for \(\lambda \ll 0\). We consider \(\lambda = -10\) and \(\lambda = -1000\), and compute approximations using the explicit Euler method (2.12) as well as the implicit Euler method of Ex. 2.28(c), comparing the results with the exact solution
\[\phi_\lambda : \mathbb{R} \rightarrow \mathbb{R}, \quad \phi_\lambda(x) = e^{-x} + (\eta - 1)e^{\lambda x}\]
\((\phi_\lambda(0) = \eta\) is immediate and
\[\phi'_\lambda(x) = -e^{-x} + \lambda(\eta - 1)e^{\lambda x} = \lambda e^{-x} + \lambda(\eta - 1)e^{\lambda x} - (1 + \lambda)e^{-x} = \lambda \phi_\lambda(x) - (1 + \lambda)e^{-x},\]
shows \(\phi_\lambda\) is a solution to (2.93a)).
For the numerical comparison, we fix \(\eta := 1\). Note
\[\eta = 1 \implies \forall \lambda \in \mathbb{R} \forall x \in \mathbb{R} \quad \phi_\lambda(x) = e^{-x}.
\]
Given \(h > 0\), the recursion for the explicit Euler method is
\[y_0 = 1, \quad \forall k \in \mathbb{N}_0 \quad y_{k+1} = y_k + h \left(\lambda y_k - (1 + \lambda)e^{-x_k}\right).\]
For the implicit Euler method, the equation for \( y_{k+1} \) is

\[
y_{k+1} = y_k + h \left( \lambda y_{k+1} - (1 + \lambda) e^{-x_{k+1}} \right),
\]

i.e. the recursion is, for \( h\lambda \neq 1 \),

\[
y_0 = 1,
\]

\[
\forall k \in \mathbb{N}_0 \quad y_{k+1} = \frac{y_k - h (1 + \lambda) e^{-x_{k+1}}}{1 - h\lambda}.
\]

We now apply both methods with increasingly small equidistant stepsizes

\[
h = 2^{-4}, 2^{-6}, 2^{-8}, 2^{-10}, 2^{-12},
\]

### Table 1: Numerical results to (2.93) with \( \lambda = -10 \), computed for several stepsizes \( h \) by the explicit Euler method and by the implicit Euler method. The problem is not stiff and both methods perform equally well.

<table>
<thead>
<tr>
<th>( h )</th>
<th>( y_h(1) - \phi_{-10}(1) ) explicit Euler</th>
<th>( y_h(1) - \phi_{-10}(1) ) implicit Euler</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2^{-4} \approx 0.0625 )</td>
<td>(-1.247 \cdot 10^{-3})</td>
<td>(1.308 \cdot 10^{-3})</td>
</tr>
<tr>
<td>( 2^{-6} \approx 0.0156 )</td>
<td>(-3.174 \cdot 10^{-4})</td>
<td>(3.212 \cdot 10^{-4})</td>
</tr>
<tr>
<td>( 2^{-8} \approx 0.0039 )</td>
<td>(-7.971 \cdot 10^{-5})</td>
<td>(7.994 \cdot 10^{-5})</td>
</tr>
<tr>
<td>( 2^{-10} \approx 0.0010 )</td>
<td>(-1.995 \cdot 10^{-5})</td>
<td>(1.996 \cdot 10^{-5})</td>
</tr>
<tr>
<td>( 2^{-12} \approx 0.0002 )</td>
<td>(-4.989 \cdot 10^{-6})</td>
<td>(4.990 \cdot 10^{-6})</td>
</tr>
</tbody>
</table>

### Table 2: Numerical results to (2.93) with \( \lambda = -1000 \), computed for several stepsizes \( h \) by the explicit Euler method and by the implicit Euler method. The problem is stiff and the explicit method is unstable for the largest three values of \( h \).

<table>
<thead>
<tr>
<th>( h )</th>
<th>( y_h(1) - \phi_{-1000}(1) ) explicit Euler</th>
<th>( y_h(1) - \phi_{-1000}(1) ) implicit Euler</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2^{-4} \approx 0.0625 )</td>
<td>(1.283 \cdot 10^{24})</td>
<td>(1.175 \cdot 10^{-5})</td>
</tr>
<tr>
<td>( 2^{-6} \approx 0.0156 )</td>
<td>(2.865 \cdot 10^{69})</td>
<td>(2.892 \cdot 10^{-6})</td>
</tr>
<tr>
<td>( 2^{-8} \approx 0.0039 )</td>
<td>(8.014 \cdot 10^{112})</td>
<td>(7.202 \cdot 10^{-7})</td>
</tr>
<tr>
<td>( 2^{-10} \approx 0.0010 )</td>
<td>(-1.797 \cdot 10^{-7})</td>
<td>(1.799 \cdot 10^{-7})</td>
</tr>
<tr>
<td>( 2^{-12} \approx 0.0002 )</td>
<td>(-4.495 \cdot 10^{-8})</td>
<td>(4.496 \cdot 10^{-8})</td>
</tr>
</tbody>
</table>
recomputing the results of [Pla06, Table 8.3], reproduced in Tables 1 and 2 for the reader’s convenience, where we write

\[ y_h(1) := y_{1/h} \approx \phi_\lambda(1) = e^{-1}. \]

For \( \lambda = -10 \), the problem (2.93) is actually not stiff and the results in Table 1 show that both the explicit and the implicit Euler method produce reasonable results even for the largest stepsize \( h = 2^{-4} \).

For \( \lambda = -1000 \), (2.93) is stiff, and the results in Table 2 show that, while the implicit Euler method performs reasonably for each value of \( h \), the error of the explicit Euler method appears to tend to infinity for the first three values of \( h \), but then becomes reasonably small for \( h = 2^{-10} \) and \( h = 2^{-12} \) (one says the explicit Euler method is unstable for the larger values of \( h \)).

**Example 2.81.** Let us consider the 2-dimensional example, where \( n = 2 \), \( \xi = 0 \), and

\[ f : \mathbb{R} \times \mathbb{R}^2 \to \mathbb{R}^2, \quad f(x, y) := Ay = \begin{pmatrix} -100 y_1 + y_2 \\ -\frac{1}{10} y_2 \end{pmatrix}, \quad A := \begin{pmatrix} -100 & 1 \\ 0 & -\frac{1}{10} \end{pmatrix}, \]

i.e. the initial value problem

\[ \begin{aligned}
    y'_1 &= -100 y_1 + y_2, \\
    y'_2 &= -\frac{1}{10} y_2, \\
    y(0) &= \eta, \quad \eta \in \mathbb{R}^2.
\end{aligned} \tag{2.94a} \]

To compare with Def. 2.78 and Rem. 2.79, we compute (using the Euclidean scalar product)

\[ \forall (x, y), (x, \bar{y}) \in \mathbb{R}^2 \]

\[ \langle f(x, y) - f(x, \bar{y}), y - \bar{y} \rangle = \langle A(y - \bar{y}), y - \bar{y} \rangle 
\]

\[ = -100 (y_1 - \bar{y}_1)^2 + (y_2 - \bar{y}_2)(y_1 - \bar{y}_1) - \frac{1}{10} (y_2 - \bar{y}_2)^2 
\]

\[ \leq - \left( 10 |y_1 - \bar{y}_1| - \frac{1}{\sqrt{10}} |y_2 - \bar{y}_2| \right)^2 \leq 0 \leq \|y - \bar{y}\|^2_2, \]

and

\[ \inf \left\{ \frac{\langle f(x, y) - f(x, \bar{y}), y - \bar{y} \rangle}{\|y - \bar{y}\|^2_2} : y, \bar{y} \in \mathbb{R}^n, \quad y \neq \bar{y} \right\} = -\infty, \]

obtaining \( M(x) = 0 \) and \( m(x) = \infty \). Thus, the problem (2.94) is stiff according to Def. 2.78(b).

As in the previous example, the goal is to assess the performance of both the explicit Euler method (2.12) and the implicit Euler method of Ex. 2.28(c), comparing their respective results with the exact solution. As (2.94a) is a linear ODE with constant coefficients, the exact solution can be obtained using corresponding results from the theory of ODE. To this end, one observes \( A \) to be diagonalizable,

\[ D := W^{-1} AW = \begin{pmatrix} -100 & 0 \\ 0 & -\frac{1}{10} \end{pmatrix}, \quad W = \begin{pmatrix} 1 & 1 \\ 0 & \frac{999}{10} \end{pmatrix}, \quad W^{-1} = \begin{pmatrix} 1 & -\frac{10}{999} \\ 0 & \frac{999}{10} \end{pmatrix}. \]
Indeed,
\[
W^{-1}AW = W^{-1} \begin{pmatrix} -100 & 1 \\ 0 & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} 1 & 10 \\ 0 & \frac{999}{10} \end{pmatrix} = \begin{pmatrix} 1 & -\frac{10}{999} \\ 0 & \frac{10}{999} \end{pmatrix} \begin{pmatrix} -100 & -100 + \frac{999}{10} \\ 0 & -\frac{999}{10} \end{pmatrix} 
= \begin{pmatrix} -100 & 0 \\ 0 & -\frac{1}{10} \end{pmatrix}.
\]

As a consequence of [Phi16c, Th. 4.44(b)],
\[
\Psi : \mathbb{R} \rightarrow \mathcal{M}(2, \mathbb{R}), \quad \Psi(x) := \begin{pmatrix} e^{-100x} & 0 \\ 0 & e^{-x/10} \end{pmatrix},
\]
constitutes a fundamental matrix solution to \(y' = Dy\), and, by [Phi16c, Th. 4.47],
\[
\Phi : \mathbb{R} \rightarrow \mathcal{M}(2, \mathbb{R}), \quad \Phi(x) := W\Psi(x),
\]
constitutes a fundamental matrix solution to (2.94a). Thus, by variation of constants [Phi16c, (4.29)] with \(x_0 = 0, \ y_0 = \eta\) and \(b(t) \equiv 0\), we obtain the following solution to the initial value problem (2.94):
\[
\phi : \mathbb{R} \rightarrow \mathbb{R}^2, \quad \phi(x) = \Phi(x)\Phi^{-1}(0) \eta = W\Psi(x)W^{-1}\Psi^{-1}(0) \eta = W\Psi(x)W^{-1} \eta.
\]

Introducing the abbreviations
\[
v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} := W^{-1} \eta, \quad w_{22} := \frac{999}{10},
\]
one can rewrite \(\phi\) as
\[
\phi : \mathbb{R} \rightarrow \mathbb{R}^2, \quad \phi(x) = \begin{pmatrix} 1 & 1 \\ 0 & w_{22} \end{pmatrix} \begin{pmatrix} v_1 e^{-100x} \\ v_2 e^{-x/10} \end{pmatrix} = e^{-100x} \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + e^{-x/10} \begin{pmatrix} v_2 \\ w_{22}v_2 \end{pmatrix}.
\]

Given \(h > 0\), the recursion for the explicit Euler method is
\[
\forall k \in \mathbb{N}_0 \quad y_0 = \eta, \quad y_{k+1} = y_k + hAy_k = (\text{Id} + hA)y_k,
\]
implying
\[
\forall k \in \mathbb{N}_0 \quad y_k = (\text{Id} + hA)^k y_0 = W (\text{Id} + hD)^k W^{-1} \eta.
\]

Using
\[
\forall k \in \mathbb{N}_0 \quad (\text{Id} + hD)^k = \begin{pmatrix} (1 - 100h)^k & 0 \\ 0 & (1 - \frac{1}{10}h)^k \end{pmatrix}
\]
and the above definitions of \(v\) and \(w_{22}\), we obtain
\[
\forall k \in \mathbb{N}_0 \quad y_k = (1 - 100h)^k \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + \left(1 - \frac{1}{10}h\right)^k \begin{pmatrix} v_2 \\ w_{22}v_2 \end{pmatrix}.
\]

(2.95)
For $h > \frac{1}{50}$, we have $|1 - 100h| > 1$ and (2.96) implies
\[ v_1 \neq 0 \implies \lim_{k \to \infty} \|y_k\|_2 = \infty, \]
whereas $\lim_{x \to \infty} \|\phi(x)\|_2 = 0$, showing the instability of the explicit Euler method for $h > \frac{1}{50}$. When using (2.95) instead of (2.96), numerically, due to roundoff errors, $\lim_{k \to \infty} \|y_k\|_2 = \infty$ will even occur for $v_1 = 0$.

For the implicit Euler method, the equation for $y_{k+1}$ is
\[ y_{k+1} = y_k + hAy_{k+1}, \]
i.e. the recursion is
\[ y_0 = \eta, \quad \forall k \in \mathbb{N}_0 \quad y_{k+1} = (\text{Id} - hA)^{-1}y_k, \]
implying
\[ \forall k \in \mathbb{N}_0 \quad y_k = \left( (\text{Id} - hA)^{-1} \right)^k y_0 = W \left( (\text{Id} - hD)^{-1} \right)^k W^{-1} \eta. \]

Using
\[ \forall k \in \mathbb{N}_0 \quad \left( (\text{Id} - hD)^{-1} \right)^k = \begin{pmatrix} \frac{1}{1 + 100h} & 0 \\ 0 & \frac{1}{1 + \frac{1}{10}h} \end{pmatrix} \]
and the above definitions of $v$ and $w_{22}$, we obtain
\[ \forall k \in \mathbb{N}_0 \quad y_k = \begin{pmatrix} 1 + 100h \\ 0 \end{pmatrix}^k \begin{pmatrix} v_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 1 + \frac{1}{10}h \\ \frac{1}{10} \end{pmatrix}^k \begin{pmatrix} v_2 \\ w_{22}v_2 \end{pmatrix}, \]
such that
\[ \lim_{k \to \infty} \|y_k\|_2 = 0, \]
independently of the size of $h > 0$. While this does not prove the convergence of the implicit Euler method, a convergence result is provided by the following Th. 2.84.

**Lemma 2.82.** Consider the implicit Euler method of Ex. 2.28(c) for the initial value problem (2.91), i.e. $y' = f(x,y)$, $y(\xi) = \eta$, $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$, $\eta \in \mathbb{R}^n$. Let $b \in \mathbb{R}$, $b > \xi$, and assume $\phi : [\xi,b] \to \mathbb{R}^n$ to be the unique solution to $y' = f(x,y)$, $y(\xi) = \eta$, on $[\xi,b]$. We define
\[ \forall x \in [\xi,b] \quad \forall h \in [0,b-x] \quad \tilde{\lambda}(x,h) := \phi(x) + h f(x+h, \phi(x+h)) - \phi(x+h), \]
which constitutes a variant of the local truncation error $\lambda(x,h)$ of Def. 2.12(c)\(^2\). If $f \in C^1(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n)$, then
\[ \exists C \geq 0 \quad \forall x \in [\xi,b] \quad \forall h \in [0,b-x] \quad \|\tilde{\lambda}(x,h)\| \leq C h^2. \]
\(^2\)In general, $\tilde{\lambda}$ and $\lambda$ will not be identical, due to the fact that Def. 2.12(c) assumes the method to be written in explicit form. In particular, the result of the present lemma is not the same as the result obtained from Ex. 2.62(a) and Th. 2.60(b), which says that the method is consistent of order 1, when rewritten with a $\phi$ in explicit (standard) form.
Proof. If $f$ is $C^1$, then the solution $\phi$ is $C^2$ according to Prop. C.1 in the Appendix. Thus, we can apply Taylor’s theorem with the remainder term in integral form to obtain, for each $x \in [\xi, b]$ and each $h \in [0, b - x]$,
\[
\phi(x) = \phi(x + h - h) = \phi(x + h) - \phi'(x + h) h + \int_{x+h}^{x} (x-t) \phi''(t) \, dt.
\]
Applying this in the definition of $\tilde{\lambda}$ yields, for each $x \in [\xi, b]$ and each $h \in [0, b - x]$,
\[
\tilde{\lambda}(x, h) = \phi(x) + h f(x + h, \phi(x + h)) - \phi(x + h) = - \int_{x}^{x+h} (x-t) \phi''(t) \, dt,
\]
and, thus,
\[
\|\tilde{\lambda}(x, h)\| \leq C h^2,
\]
where
\[
C = \frac{1}{2} \max\{\|\phi''(t)\| : t \in [\xi, b]\},
\]
completing the proof of the lemma.

Lemma 2.83. Let $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ denote a fixed scalar product on $\mathbb{R}^n$, $n \in \mathbb{N}$, with induced norm $\| \cdot \|$. Let $\xi, b \in \mathbb{R}$ with $\xi < b$ and assume $f : [\xi, b] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ to satisfy an upper Lipschitz condition with respect to $y$ on $[\xi, b]$ as defined in Def. 2.78(a), however with a constant function $M(x) \equiv M \in \mathbb{R}$.

(a) If $M \leq 0$, then
\[
\forall x \in [\xi, b], \forall y, \tilde{y} \in \mathbb{R}^n, \forall h \in \mathbb{R}^+_0 \quad \|y - \tilde{y}\| \leq \|y - \tilde{y} - h(f(x, y) - f(x, \tilde{y}))\|.
\]
(b) If $M > 0$, then
\[
\forall x \in [\xi, b], \forall y, \tilde{y} \in \mathbb{R}^n, \forall h, H \in \mathbb{R}^+_0, \quad \|y - \tilde{y}\| \leq C_h \|y - \tilde{y} - h(f(x, y) - f(x, \tilde{y}))\|,
\]
where
\[
C_h := 1 + \frac{h M}{1 - HM} = \frac{1 + M(h - H)}{1 - HM}.
\]
Proof. Let $x \in [\xi, b]$, let $y, \tilde{y} \in \mathbb{R}^n$, and let $h \in \mathbb{R}^+_0$. Then, according to Def. 2.78(a),
\[
h \langle f(x, y) - f(x, \tilde{y}), y - \tilde{y} \rangle \leq h M \|y - \tilde{y}\|^2,
\]
implying
\[
(1 - h M) \|y - \tilde{y}\|^2 \leq \langle y - \tilde{y}, y - \tilde{y} \rangle - h \langle f(x, y) - f(x, \tilde{y}), y - \tilde{y} \rangle
\]
\[
= \langle y - \tilde{y} - h(f(x, y) - f(x, \tilde{y})), y - \tilde{y} \rangle
\]
\[
\leq \|y - \tilde{y} - h(f(x, y) - f(x, \tilde{y}))\| \|y - \tilde{y}\|,
\]
where the Cauchy-Schwarz inequality was used for the last estimate. If \( M \leq 0 \), then 
\[ \|y - \tilde{y}\| \leq (1 - hM) \|y - \tilde{y}\|, \]
proving (a). Now let \( M > 0 \) and \( h \leq H < 1/M \). Then 
\[ 1 - hM > 0 \]
and we may divide the above estimate by this term without changing the inequality. Since 
\[ \frac{1}{1 - hM} = 1 + \frac{hM}{1 - hM} \leq 1 + \frac{hM}{1 - HM} = C_k, \]
this proves (b).

**Theorem 2.84.** Let \( \langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R} \) denote a fixed scalar product on \( \mathbb{R}^n \), \( n \in \mathbb{N} \), with induced norm \( \| \cdot \| \). Let \( \xi, \beta \in \mathbb{R} \) with \( \xi < \beta \) and assume the initial value problem (2.91) satisfies an upper Lipschitz condition with respect to \( y \) on \([\xi, \beta]\) as defined in Def. 2.78(a), however with a constant function \( M(x) \equiv M \in \mathbb{R} \). Moreover, assume (2.91) has a unique solution \( \phi \) defined on \([\xi, \beta]\) and consider the implicit Euler method of Ex. 2.28(c) for a partition \( \Delta := (x_0, \ldots, x_N) \) of \([\xi, \beta]\), i.e. assume \( y_0, \ldots, y_N \in \mathbb{R}^n \) satisfy 
\[ \begin{align*}
y_0 &= \eta, \\
\forall \ k \in \{0, \ldots, N-1\} \quad y_{k+1} &= y_k + h_k f(x_{k+1}, y_{k+1}), \quad h_k := x_{k+1} - x_k.
\end{align*} \]

If \( \tilde{\lambda}(x, h) \) is defined as in Lem. 2.82 and
\[ \exists \ C \geq 0 \quad \forall \ x \in [\xi, \beta] \quad \forall \ h \in [0, b-x] \quad \|\tilde{\lambda}(x, h)\| \leq C h^2, \]
(according to Lem. 2.82, \( f \in C^1(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n) \) is sufficient), then the global truncation error can be estimated

\[ \max \left\{ \|y_k - \phi(x_k)\| : k \in \{0, \ldots, N\} \right\} \leq K h_{\max}(\Delta), \quad \text{where} \]

\[ K := \begin{cases} C(b - \xi) & \text{for } M \leq 0, \\ \frac{C}{M} \left( e^{M(b - \xi)}/(1 - h_{\max}(\Delta)M) - 1 \right) & \text{for } M > 0 \text{ and } 0 < h_{\max}(\Delta) < 1/M, \end{cases} \]

and where \( h_{\max}(\Delta) \) is the mesh size of \( \Delta \) as defined in (2.13a). Note that \( C \) (and, hence, \( K \)) does not depend on either \( f \) or the partition. Moreover, for \( M \leq 0 \), \( K \) tends to be of moderate size and it grows at most linearly with the length of the interval \([\xi, \beta]\).

**Proof.** Introducing the abbreviations
\[ \forall \ k \in \{0, \ldots, N\} \quad \phi_k := \phi(x_k), \quad e_k := y_k - \phi_k, \]
and
\[ \forall \ k \in \{0, \ldots, N-1\} \quad \lambda_k := \tilde{\lambda}(x_k, h_k) = \phi_k + h_k f(x_{k+1}, \phi_{k+1}) - \phi_{k+1}, \]
we obtain, for each \( k \in \{0, \ldots, N - 1\}, \)
\[ \begin{align*}
e_k + \lambda_k &= y_k - \phi_k + \phi_k + h_k f(x_{k+1}, \phi_{k+1}) - \phi_{k+1} \\
&= e_{k+1} - y_{k+1} + \phi_{k+1} + y_k + h_k f(x_{k+1}, \phi_{k+1}) - \phi_{k+1} \\
&= e_{k+1} - h_k (f(x_{k+1}, y_{k+1}) - f(x_{k+1}, \phi_{k+1})).
\end{align*} \]
If $M \leq 0$, then we apply (2.98) together with Lem. 2.83(a), yielding
\[
\|e_{k+1}\| \leq \|e_k + \lambda_k\| \leq \|e_k\| + \|\lambda_k\| \leq \|e_k\| + Ch_k^2.
\]
Using this in an induction on $k \in \{0, \ldots, N\}$ yields
\[
\forall k \in \{0, \ldots, N\} \quad \|e_k\| \leq C (x_k - \xi) h_{\text{max}}(\Delta):
\]
Indeed, $e_0 = y_0 - \phi_0 = \eta - \eta = 0$, and, for $k \in \{0, \ldots, N-1\}$,
\[
\|e_{k+1}\| \leq \|e_k\| + Ch_k^2, \quad \text{ind.hyp.}
\]
\[
\|e_k\| \leq C (x_k - \xi) h_{\text{max}}(\Delta) + C h_k h_{\text{max}}(\Delta) = C (x_{k+1} - \xi) h_{\text{max}}(\Delta),
\]
completing the induction and, in particular, proving (2.97) for $M \leq 0$. Analogously, for $M > 0$, we apply (2.98) together with Lem. 2.83(a), yielding, for $h_{\text{max}}(\Delta) < 1/M$,
\[
\|e_{k+1}\| \leq C h_k (\|e_k\| + \|\lambda_k\|) \leq C h_k \|e_k\| + \frac{1}{1 - h_{\text{max}}(\Delta) M} \|\lambda_k\|
\]
\[
\leq C h_k \|e_k\| + \frac{C h_{\text{max}}(\Delta)}{1 - h_{\text{max}}(\Delta) M} h_k,
\]
where
\[
C h_k = 1 + \frac{h_k M}{1 - h_{\text{max}}(\Delta) M} = \frac{1 + M(h_k - h_{\text{max}}(\Delta))}{1 - h_{\text{max}}(\Delta) M}.
\]
We can now apply Lem. 2.14 with
\[
a_k := \|e_k\|, \quad a_0 = \|e_0\| = 0, \quad L := \frac{M}{1 - h_{\text{max}}(\Delta) M}, \quad \beta := \frac{C h_{\text{max}}(\Delta)}{1 - h_{\text{max}}(\Delta) M},
\]
implying
\[
\forall k \in \{0, \ldots, N\} \quad \|e_k\| \leq \frac{e^{L(x_k - \xi)} - 1}{L} \beta \left( e^{M(x_k - \xi)/(1 - h_{\text{max}}(\Delta) M) - 1} \right) \frac{C h_{\text{max}}(\Delta)}{M},
\]
in particular, proving (2.97) for $M > 0$. \qed

### 2.7 Collocation Methods

An important idea for constructing useful implicit methods (actually, implicit RK methods, as we will see below) is so-called collocation (explicit RK methods can also result, see, e.g., Ex. 2.91(a), but this is not the main importance of collocation methods). Given the initial-value problem
\[
y' = f(x, y), \quad y(\xi) = \eta, \quad f : G \rightarrow \mathbb{K}^n, \quad G \subseteq \mathbb{R} \times \mathbb{K}^n, \quad (\xi, \eta) \in G,
\]
the idea is to approximate the exact solution $\phi : [\xi, \xi + h] \rightarrow \mathbb{K}^n$ via a $\mathbb{K}^n$-valued polynomial of degree at most $s$, that satisfies the ODE at least at the $s$ collocation points $\xi + c_j h$, where $0 \leq c_1 < \cdots < c_s \leq 1$.  

---

**2 NUMERICAL SOLUTION OF ODE**
**Definition 2.85.** Let \( s, n \in \mathbb{N} \), \( G \subseteq \mathbb{R} \times \mathbb{K}^n \), \( f : G \rightarrow \mathbb{K}^n \), let \( c := (c_1, \ldots, c_s)^t \in [0, 1]^s \) with \( 0 \leq c_1 < \cdot \cdot \cdot < c_s \leq 1 \).

(a) Let \((x, y, h) \in \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+ \). We call a \( \mathbb{K}^n \)-valued polynomial \( P \) of degree at most \( s \) (i.e. each component of \( P \) is a \( \mathbb{K} \)-valued polynomial of degree at most \( s \)) *collocation polynomial*, determined by \( c, f, x, y, \) and \( h \), if, and only if, \( P \) satisfies the following conditions (i) – (ii):

(i) \( P(x) = y. \)

(ii) \( P'(x + c_j h) = f(x + c_j h, P(x + c_j h)) \) for each \( j \in \{1, \ldots, s\} \) (which, in particular, is supposed to mean, for each \( j \in \{1, \ldots, s\} \), that the used argument of \( f \) is in \( G \)).

(b) Let \((\xi, \eta) \in G\). We call a numerical method as defined in Def. 2.7 *collocation method*, determined by \( c \), if, and only if, \( m = 1, \alpha_0 = 1 \), and the defining function has the form

\[
\varphi : D_\varphi \rightarrow \mathbb{K}^n, \quad \varphi(x, y, h) = -y + \frac{P(x, y, h)(x + h)}{h} \quad \text{(recall } h > 0),
\]

where, for each \((x, y, h) \in D_\varphi\), \( P(x, y, h) \) is a collocation polynomial determined by \( c, f, x, y, \) and \( h \) according to (a).

With each collocation method as defined above, we will now associate an implicit RK method. Then, in Th. 2.89 below, we will show that the resulting defining functions are identical.

**Definition 2.86.** Let \( s \in \mathbb{N} \) and \( c := (c_1, \ldots, c_s)^t \in [0, 1]^s \) with \( 0 \leq c_1 < \cdot \cdot \cdot < c_s \leq 1 \). Let \( L_1, \ldots, L_s : \mathbb{R} \rightarrow \mathbb{R} \) be the Lagrange basis polynomials (cf. [Phi17b, Th. 3.2]) of degree at most \( s - 1 \), satisfying

\[
\forall_{j,l \in \{1, \ldots, s\}} L_j(c_l) = \delta_{jl} = \begin{cases} 1 & \text{for } j = l, \\ 0 & \text{for } j \neq l. \end{cases}
\]

Define

\[
\forall_{j \in \{1, \ldots, s\}} b_j := \int_0^1 L_j(t) \, dt, \quad \forall_{j,l \in \{1, \ldots, s\}} a_{jl} := \int_0^{c_j} L_l(t) \, dt.
\]

Then the RK method with weights \( b_1, \ldots, b_s \), nodes \( c_1, \ldots, c_s \), and RK matrix \( (a_{jl}) \) is called the RK method defined by \( c \) and collocation.

**Lemma 2.87.** Let \( s, n \in \mathbb{N} \), \( G \subseteq \mathbb{R} \times \mathbb{K}^n \), \( f : G \rightarrow \mathbb{K}^n \), let \( c := (c_1, \ldots, c_s)^t \in [0, 1]^s \) with \( 0 \leq c_1 < \cdot \cdot \cdot < c_s \leq 1 \). Let \((x, y, h) \in \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+ \) and let \( P := P(x, y, h) \) be a collocation polynomial determined by \( c, f, x, y, \) and \( h \), according to Def. 2.85(a). Defining

\[
\forall_{j \in \{1, \ldots, s\}} k_j := k_j(x, y, h) := P'(x + c_j h),
\]
we have

\[ \forall t \in \mathbb{R} \quad P'(x + th) = \sum_{j=1}^{s} k_j L_j(t), \quad (2.99a) \]

\[ \forall t \in \mathbb{R} \quad P(x + th) = y + h \sum_{j=1}^{s} k_j \int_{0}^{t} L_j(\theta) \, d\theta, \quad (2.99b) \]

where the \( L_j \) are defined as in Def. 2.86 above.

**Proof.** We have that (2.99a) holds, as each component of both sides constitutes a \( K \)-valued polynomial on \( \mathbb{R} \) of degree at most \( s - 1 \), where both sides agree at the \( s \) distinct values \( t = c_1, \ldots, c_s \) (cf. [Phi17b, Th. 3.2]). For (2.99b), we compute, for each \( t \in \mathbb{R} \),

\[ P(x + th) = y + h \int_{0}^{t} P'(x + \theta h) \, d\theta \overset{\text{(2.99a)}}{=} y + h \sum_{l=1}^{s} k_j \int_{0}^{t} L_j(\theta) \, d\theta, \]

thereby establishing the case. \( \blacksquare \)

**Lemma 2.88.** Let \( s, n \in \mathbb{N}, G \subseteq \mathbb{R} \times \mathbb{K}^n, f : G \rightarrow \mathbb{K}^n, \) let \( c := (c_1, \ldots, c_s)^t \in [0, 1]^s \) with \( 0 \leq c_1 < \cdots < c_s \leq 1 \). Let \( (x, y, h) \in \mathbb{R} \times \mathbb{K}^n \times \mathbb{R}^+ \). Define

\[ K(x, y, h) := \left\{ k := (k_1, \ldots, k_s) \in (\mathbb{K}^n)^s : k \text{ satisfies (2.32b)} \right\}, \]

\[ P(x, y, h) := \left\{ (P : \mathbb{R} \rightarrow \mathbb{K}^n) : P \text{ satisfies Def. 2.85(a)(i),(ii)} \right\}. \]

Then the map \( \Phi : P(x, y, h) \rightarrow K(x, y, h) \), given by the definition in Lem. 2.87, does, indeed, map into \( K(x, y, h) \), and it is bijective with the inverse given by \( \Phi^{-1} : K(x, y, h) \rightarrow P(x, y, h) \), \( k \mapsto P \), where

\[ P : \mathbb{R} \rightarrow \mathbb{K}^n, \quad P(t) := y + h \sum_{j=1}^{s} k_j \int_{0}^{t} L_j(\theta) \, d\theta. \quad (2.100) \]

**Proof.** If \( P \in P(x, y, h) \), then, from (2.99b) with \( t = c_j \), we obtain

\[ \forall j \in \{1, \ldots, s\} \quad P(x + c_j h) = y + h \sum_{l=1}^{s} a_j k_l \quad (2.101) \]

and, thus, for each \( j \in \{1, \ldots, s\} \),

\[ k_j = P'(x + c_j h) \overset{\text{Def. 2.85(a)(ii)}}{=} f(x + c_j h, P(x + c_j h)) \]

\[ = f \left( x + c_j h, y + h \sum_{l=1}^{s} a_j k_l \right), \]

showing \( k = \Phi(P) \) to satisfy (2.32b), i.e. \( k = \Phi(P) \in K(x, y, h) \).
If \( k \in \mathcal{K}(x, y, h) \) and \( P := \Phi^{-1}(k) \) is given by (2.100), then, since each \( L_j \) is a polynomial of degree at most \( s-1 \), clearly, \( P \) is a \((\mathbb{K}^n\text{-valued})\) polynomial of degree at most \( s \). If \( t = x \) in the definition of \( P \), then all the integrals vanish, showing \( P(x) = y \), in accordance with Def. 2.85(a)(i). Moreover, 

\[
P'(t) = \sum_{j=1}^{s} k_j L_j \left( \frac{t - x}{h} \right), \tag{2.102a}
\]

implying 

\[
\forall j \in \{1, \ldots, s\} \quad P'(x + c_j h) = \sum_{l=1}^{s} k_l L_l(c_j) = k_j. \tag{2.102b}
\]

In consequence, Lem. 2.87 applies, showing (2.101) to hold, once again. Thus, as \( k \) satisfies (2.32b), we obtain 

\[
P'(x + c_j h) = k_j = f \left( x + c_j h, y + h \sum_{l=1}^{s} a_{jl} k_l \right) = f(x + c_j h, P(x + c_j h)),
\]

which is Def. 2.85(a)(ii), showing \( P = \Phi^{-1}(k) \in \mathcal{P}(x, y, h) \).

\( \Phi \circ \Phi^{-1} = \text{Id} \): Let \( k \in \mathcal{K}(x, y, h) \). Then \( P := \Phi^{-1}(k) \) is given by (2.100) and 

\[
\Phi(P) = (\bar{k}_1, \ldots, \bar{k}_s), \quad \text{where } \forall j \in \{1, \ldots, s\} \quad \bar{k}_j := P'(x + c_j h).
\]

On the other hand, from the above argument, we know (2.102) to hold as well, showing 

\[
\forall j \in \{1, \ldots, s\} \quad \bar{k}_j = P'(x + c_j h) = k_j,
\]

proving \( \Phi \circ \Phi^{-1} = \text{Id} \).

\( \Phi^{-1} \circ \Phi = \text{Id} \): Let \( P \in \mathcal{P}(x, y, h) \). Then \( k := \Phi(P) \) is given by 

\[
\forall j \in \{1, \ldots, s\} \quad k_j := P'(x + c_j h)
\]

and \( \bar{P} := \Phi^{-1}(k) \) is given by 

\[
\bar{P} : \mathbb{R} \to \mathbb{K}^n, \quad \bar{P}(t) := y + h \sum_{j=1}^{s} k_j \int_{0}^{\frac{t-x}{h}} L_j(\theta) \, d\theta.
\]

As both \( P' \) and \( \bar{P}' \) satisfy (2.102), they are both polynomials of degree at most \( s-1 \) that agree at the \( s \) distinct points \( t = x + c_j h \), implying \( P' = \bar{P}' \). Since, also \( P(x) = y = \bar{P}(x) \), we conclude \( P = \bar{P} \), proving \( \Phi^{-1} \circ \Phi = \text{Id} \) and completing the proof of the lemma. \( \blacksquare \)

**Theorem 2.89.** Let \( s, n \in \mathbb{N}, G \subseteq \mathbb{R} \times \mathbb{K}^n, f : G \to \mathbb{K}^n \), let \( c := (c_1, \ldots, c_s)^t \in [0, 1]^s \) with \( 0 \leq c_1 < \cdots < c_s \leq 1 \).
(a) If $\varphi : D_\varphi \rightarrow \mathbb{K}^n$, $\varphi(x, y, h) = \frac{-y + P(x, y, h)(x + h)}{h}$, is the defining function of a collocation method, determined by $c$, as in Def. 2.85(b), then, defining $b \in \mathbb{R}^s$ and $(a_{jl}) \in M(s, \mathbb{R})$ as in Def. 2.86 plus the $k_j(x, y, h)$ as in Lem. 2.87, we have that $\varphi$ satisfies (2.32), i.e.

$$\forall (x,y,h)\in D_\varphi \quad \varphi(x, y, h) = \sum_{j=1}^{s} b_j k_j(x, y, h)$$

and

$$\forall (x,y,h)\in D_\varphi \quad \forall j\in\{1, \ldots, s\} \quad k_j(x, y, h) = f \left( x + c_jh, y + h \sum_{l=1}^{s} a_{jl} k_l(x, y, h) \right),$$

i.e. $\varphi$ is the defining function of the RK method defined by $c$ and collocation.

(b) If $\varphi : D_\varphi \rightarrow \mathbb{K}^n$ is the defining function of the RK method defined by $c$ and collocation, then it satisfies (2.32) with certain $k_j := k_j(x, y, h)$ for each $(x, y, h) \in D_\varphi$. With the notation of Lem. 2.88, we have $k := (k_1, \ldots, k_s) \in K(x, y, h)$ and, by Lem. 2.88, $P := \Phi^{-1}(k) \in \mathcal{P}(x, y, h)$ is a collocation polynomial determined by $c$, $f$, $x$, $y$, and $h$, according to Def. 2.85(a). Moreover, $\varphi$ is the defining function of a collocation method, satisfying Def. 2.85(b), and

$$\sum_{k\in\{1, \ldots, s\}} \sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \quad \sum_{k\in\{1, \ldots, s\}} \sum_{j=1}^{s} a_{jl} c_j^{k-1} = \frac{c_l^{j}}{k}, \quad \text{for each } (x,y,h) \in D_\varphi$$

i.e., in particular, the RK method defined by $c$ and collocation satisfies the consistency condition (2.34) as well as the node condition (2.35).

Proof. (a): Let $(x, y, h) \in D_\varphi$. Then we obtain, from (2.99b) with $t = 1$,

$$P(x, y, h)(x + h) = y + h \sum_{j=1}^{s} k_j(x, y, h) \int_{0}^{1} L_j(\theta) \, d\theta = y + h \sum_{j=1}^{s} b_j k_j(x, y, h), \quad (2.104)$$

showing $\varphi$ to satisfy (2.32a). This proves (a), since we know from Lem. 2.88 that $\varphi$ satisfies (2.32b) as well.

(b): As in the proof of (a), Lem. 2.87 applies, yielding, once again, the validity of (2.104). Thus

$$\varphi(x, y, h) = \sum_{j=1}^{s} b_j k_j(x, y, h) = \frac{-y + P(x, y, h)(x + h)}{h},$$

as claimed. To prove (2.103), we first note that

$$\forall t \in \mathbb{R} \quad \forall k\in\{1, \ldots, s\} \quad t^{k-1} = \sum_{j=1}^{s} c_j^{k-1} L_j(t), \quad (2.105)$$
which holds, as both sides of (2.105) constitute polynomials of degree at most \( s - 1 \) that agree at the \( s \) distinct points \( t = c_1, \ldots, c_s \). Thus,

\[
\sum_{j=1}^{s} b_j c_j^{k-1} = \sum_{j=1}^{s} \int_0^1 c_j^{k-1} L_j(t) \, dt \quad (2.105) = \int_0^1 t^{k-1} \, dt = \frac{1}{k},
\]

proving (2.103a), also yielding the consistency condition by setting \( k := 1 \). Similarly,

\[
\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \sum_{j=1}^{s} \int_0^1 c_j^{k-1} L_j(t) \, dt \quad (2.105) = \int_0^1 t^{k-1} \, dt = \frac{c^k}{k},
\]

proving (2.103b), also yielding the node condition by setting \( k := 1 \).

\[\text{Corollary 2.90. Let } s, n \in \mathbb{N}, G \subseteq \mathbb{R} \times \mathbb{K}^n, f : G \to \mathbb{K}^n, \text{ let } c := (c_1, \ldots, c_s) \in [0, 1]^s \text{ with } 0 \leq c_1 < \cdots < c_s \leq 1. \text{ Moreover, let } G \text{ be open and assume } f \text{ to be continuous and} \\
\text{locally Lipschitz with respect to } y. \\
\]

\[\text{(a) For each } (x, y) \in G, \text{ there exist } \eta(x, y) \in [0, \infty] \text{ and } r(x, y) \in \mathbb{R}^+ \text{ such that, if } \\
\text{if } h \in ] - \eta(x, y), \eta(x, y) [, \text{ there exists a unique collocation polynomial } P, \text{ determined} \\
\text{by } c, f, x, y, \text{ and } h, \text{ with the additional property} \\
\forall j \in \{1, \ldots, s\} \quad \| P'(x + c_j h) - f(x, y) \| < r(x, y). \]

\[\text{(b) If, additionally, } G = \mathbb{R} \times \mathbb{K}^n \text{ and } f \text{ is globally } L\text{-Lipschitz with respect to } y (L \in \mathbb{R}_0^+), \\
\text{and } \| A \|_\infty \text{ denotes the operator norm of } A = (a_{jl}) (a_{jl} \text{ defined as in Def. 2.86) with} \\
\text{respect to } \| \cdot \|_\infty \text{ on } \mathbb{K}^s, \text{ then, for each } h \in ] - \eta, \eta [, \text{ with} \\
\eta := \frac{1}{L \| A \|_\infty} (1/0 := \infty), \]

\[\text{there exists a unique collocation polynomial } P, \text{ determined by } c, f, x, y, \text{ and } h \text{ (i.e.} \\
\text{uniqueness is guaranteed without the additional condition of (a)).} \]

\[\text{Proof. Let } (x, y) \in G. \text{ Under the hypotheses of the corollary, according to Th. 2.31} \\
\text{and Lem. 2.25, there exist } \eta(x, y) \in [0, \infty] \text{ and } r(x, y) \in \mathbb{R}^+ \text{ such that, for each } h \in \\
] - \eta(x, y), \eta(x, y) [, \text{ the system (2.32b) has a unique solution} \\
k := (k_1, \ldots, k_s) \in B_r(x,y)(y_f, \ldots, y_f) \subseteq (\mathbb{K}^n)^s, \text{ where } y_f := f(x, y), \\
\text{and this solution is even unique in the entire space } (\mathbb{K}^n)^s \text{ under the additional hypotheses} \]
\[\text{of (b). Now, by Lem. 2.88 each solution } k \text{ to (2.32b) uniquely corresponds to a} \\
\text{collocation polynomial } P = \Phi^{-1}(k), \text{ determined by } c, f, x, y, \text{ and } h, \text{ where the corre-} \\
\text{spendence is given by the bijective map } \Phi \text{ of Lem. 2.88. Since, for each } j \in \{1, \ldots, s\}, \\
k_j = P'(x + c_j h), \text{ this proves the corollary.}\]

\[\text{Example 2.91. (a) Consider } s = 1 \text{ with the collocation point } c_1 := 0. \text{ From the} \\
\text{consistency and the node condition, we directly obtain } b_1 = 1 \text{ and } a_{11} = 0, \text{ showing} \\
\text{the resulting collocation method to be the explicit Euler method.}\]
(b) Consider $s = 1$ with the collocation point $c_1 := 1$. From the consistency and the node condition, we directly obtain $b_1 = a_{11} = 1$, showing the resulting collocation method to be the implicit Euler method.

(c) Consider $s = 2$ with the collocation points $c_1 := 0$, $c_2 := 1$. Then

$$L_1(t) = 1 - t, \quad L_2(t) = t,$$

$$b_1 = \int_0^1 (1 - t) \, dt = \left[ t - \frac{t^2}{2} \right]_0^1 = \frac{1}{2}, \quad b_2 = \int_0^1 t \, dt = \left[ \frac{t^2}{2} \right]_0^1 = \frac{1}{2},$$

$$a_{11} = \int_0^0 L_1(t) \, dt = a_{12} = \int_0^0 L_2(t) \, dt = 0,$$

$$a_{21} = \int_0^1 L_1(t) \, dt = b_1 = \frac{1}{2}, \quad a_{22} = \int_0^1 L_2(t) \, dt = b_2 = \frac{1}{2},$$

showing this collocation method to be the implicit trapezoidal method of Ex. 2.43(d).

(d) The classical RK method is not a collocation method, since it has $c_2 = c_3 = \frac{1}{2}$.

(e) Let $s \in \mathbb{N}$ and let $0 < c_1 < \cdots < c_s < 1$ be the zeros of the $s$th orthogonal polynomial with respect to the $L^2[0,1]$-scalar product (these orthogonal polynomials are obtained from $t^0, t^1, t^2, \ldots$ via the usual orthonormalization procedure, see, e.g., [Phi17b, Def. 4.46]). Then the collocation method defined by these collocation points is called the Gauss method corresponding to $s$ (the name comes from the fact that the $c_j$ also define a so-called Gaussian quadrature rule, cf. [Phi17b, Def. 4.49]). Let us check that, for $s = 2$, we obtain the Gauss method of Ex. 2.62(c): The first 3 orthogonal polynomials are (using [Phi17b, (4.76)]) $p_0 \equiv 1$,

$$p_1(t) = t - \frac{\langle t, 1 \rangle}{\|p_0\|^2} = t - \int_0^1 t \, dt = t - \frac{1}{2},$$

$$p_2(t) = t^2 - \frac{\langle t^2, 1 \rangle}{\|p_0\|^2} p_0(x) - \frac{\langle t^2, t - \frac{1}{2} \rangle}{\|p_1\|^2} p_1(t)$$

$$= t^2 - \int_0^1 t^2 \, dt - \int_0^1 \left( t^3 - \frac{1}{2} t^2 \right) \, dt \left( t - \frac{1}{2} \right)$$

$$= t^2 - \frac{1}{3} - 12 \cdot \frac{1}{12} \left( t - \frac{1}{2} \right) = t^2 - t + \frac{1}{6},$$

The zeros of $p_2$ are

$$c_1 = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{1}{6}} = \frac{1}{2} - \sqrt{\frac{1}{12} - \frac{1}{2}} = \frac{\sqrt{3}}{6}, \quad c_2 = \frac{1}{2} + \frac{\sqrt{3}}{6}.$$
Then
\[ L_1(t) = -\sqrt{3} t + \frac{\sqrt{3}}{2} + \frac{1}{2}, \quad L_2(t) = \sqrt{3} t - \frac{\sqrt{3}}{2} + \frac{1}{2}, \]
\[ b_1 = \int_0^1 L_1(t) \, dt = \frac{1}{2}, \quad b_2 = \int_0^1 L_2(t) \, dt = \frac{1}{2}, \]
\[ a_{11} = \int_0^{c_1} L_1(t) \, dt = -\frac{\sqrt{3}}{2} c_1^2 + \left( \frac{\sqrt{3}}{2} + \frac{1}{2} \right) c_1 = -\frac{1}{4} + \frac{\sqrt{3}}{4} - \frac{\sqrt{3}}{12} = \frac{1}{4}, \]
\[ a_{22} = \int_0^{c_2} L_2(t) \, dt = \frac{\sqrt{3}}{2} c_2^2 + \left( -\frac{\sqrt{3}}{2} + \frac{1}{2} \right) c_2 = \frac{1}{4} + \frac{\sqrt{3}}{4} + \frac{\sqrt{3}}{12} = \frac{1}{4}, \]
\[ a_{12} = c_1 - a_{11} = \frac{1}{4} - \frac{\sqrt{3}}{6}, \quad a_{21} = c_2 - a_{22} = \frac{1}{4} + \frac{\sqrt{3}}{6}, \]
showing that, for \( s = 2 \), we obtain the Gauss method of Ex. 2.62(c).

**Definition and Remark 2.92.** Let \( s \in \mathbb{N} \) and \( c := (c_1, \ldots, c_s)^t \in [0,1]^s \) with \( 0 \leq c_1 < \cdots < c_s \leq 1 \). We assume \( b_1, \ldots, b_s \in \mathbb{R} \) to be defined as in Def. 2.86. We call the functional
\[ Q_c : \mathcal{F}([0,1], \mathbb{R}) \rightarrow \mathbb{R}, \quad Q_c(f) := \sum_{j=1}^s b_j f(c_j), \]
the quadrature rule defined by \( c \) and collocation \( \mathcal{F}([0,1], \mathbb{R}) \) denotes the set of functions mapping from \([0,1] \) into \( \mathbb{R} \). We remark that, for \( f \) integrable, \( Q_c(f) \) can, indeed, be considered as an approximation of \( \int_0^1 f(t) \, dt \).

**Theorem 2.93.** Let \( s \in \mathbb{N} \), \( c := (c_1, \ldots, c_s)^t \in [0,1]^s \) with \( 0 \leq c_1 < \cdots < c_s \leq 1 \), and consider the RK method defined by \( c \) and collocation. If the quadrature rule defined by \( c \) and collocation according to Def. and Rem. 2.92 is exact for each polynomial of degree at most \( p \in \mathbb{N}_0 \), then the RK method satisfies the consistency condition of order \( p + 1 \) (as defined in Def. 2.58). In particular, the RK method is consistent of order \( p + 1 \) for each ODE, satisfying the hypotheses of Th. 2.60(a),(b).

**Proof.** The theorem follows from [DB08, Th. 6.40] in combination with Th. 2.60(c). ☐

**Example 2.94.** Let \( s \in \mathbb{N} \). Then the Gauss method corresponding to \( s \), as defined in Ex. 2.91(e), is an example of an implicit \( s \)-stage RK method, satisfying the consistency condition of order \( 2s \) (which is the maximum by Prop. 2.45(a)): Comparing the definition of \( Q_c \) in Def. and Rem. 2.92 with [Phi17b, Def. 4.49], shows \( Q_c \) to be a Gaussian quadrature rule, which, by [Phi17b, Th. 4.50(a)], is exact for each polynomial of degree at most \( 2s - 1 \). Then Th. 2.93 yields that the Gauss method satisfies the consistency condition of order \( 2s \) (for the Gauss method with \( s = 2 \), we checked this via an explicit calculation in Ex. 2.62(c)).
A LINEAR ALGEBRA

A Linear Algebra

A.1 Inner Products and Orthogonality

Definition and Remark A.1. Let $X$ be a vector space over $\mathbb{C}$. A function $\langle \cdot, \cdot \rangle : X \times X \rightarrow \mathbb{C}$ is called an inner product or a scalar product on $X$ if, and only if, the following three conditions are satisfied:

(i) $\langle x, x \rangle \in \mathbb{R}^+$ for each $0 \neq x \in X$.

(ii) $\langle \lambda x + \mu y, z \rangle = \lambda \langle x, z \rangle + \mu \langle y, z \rangle$ for each $x, y, z \in X$ and each $\lambda, \mu \in \mathbb{C}$.

(iii) $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for each $x, y \in X$.

If $\langle \cdot, \cdot \rangle$ is an inner product on $X$, then $(X, \langle \cdot, \cdot \rangle)$ is called an inner product space and the map

$$\| \cdot \| : X \rightarrow \mathbb{R}^+_0, \quad \| x \| := \sqrt{\langle x, x \rangle},$$

(A.1)

defines a norm on $X$, called the norm induced by the inner product.

Example A.2. Let $n \in \mathbb{N}$. Clearly, the map $\langle \cdot, \cdot \rangle : \mathbb{C}^n \times \mathbb{C}^n \rightarrow \mathbb{C}$, where

$$\forall a, b \in \mathbb{C}^n \quad \langle a, b \rangle := \sum_{j=1}^{n} a_j \overline{b_j}$$

(A.2)

defines an inner product on $\mathbb{C}^n$. The norm induced by this inner product is called the 2-norm and is denoted by $\| \cdot \|_2$.

Definition A.3. Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space.

(a) Vectors $x, y \in X$ are called orthogonal or perpendicular (denoted $x \perp y$) if, and only if, $\langle x, y \rangle = 0$. An orthogonal system is a family $(x_\alpha)_{\alpha \in I}, x_\alpha \in X,$ $I$ being some index set, such that $\langle x_\alpha, x_\beta \rangle = 0$ for each $\alpha, \beta \in I$ with $\alpha \neq \beta$. An orthogonal system is called an orthonormal system if, and only if, it consists entirely of unit vectors (with respect to the induced norm).

(b) An orthonormal system $(x_\alpha)_{\alpha \in I}$ is called an orthonormal basis if, and only if, $x = \sum_{\alpha \in I} \langle x, x_\alpha \rangle x_\alpha$ for each $x \in X$.

(c) If $V$ is a linear subspace of $X$, then we define

$$V^\perp := \{ x \in X : x \perp v \text{ for each } v \in V \}.$$  

(A.3)

Lemma A.4. Let $(X, \langle \cdot, \cdot \rangle)$ be an inner product space and let $V$ be a linear subspace of $X$.

(a) $V^\perp$ is a linear subspace of $X$. 

(b) \( V \cap V^\perp = \{0\} \).

(c) If \( I \) is some index set and \((x_\alpha)_{\alpha \in I}\), \(x_\alpha \in X\), is an orthogonal system such that \( x_\alpha \neq 0 \) for each \( \alpha \in I \), then the \( x_\alpha \) are all linearly independent.

(d) Let \( 0 < \dim X = n < \infty \). Then an orthonormal system \((x_\alpha)_{\alpha \in I}\) is an orthonormal basis if, and only if, it is a basis in the usual sense of linear algebra.

Proof. (a): Let \( x, y \in V^\perp \) and \( \lambda \in \mathbb{C} \). Then, for each \( v \in V \), it holds that \( \langle x + y, v \rangle = \langle x, v \rangle + \langle y, v \rangle = 0 \) and \( \langle \lambda x, v \rangle = \lambda \langle x, v \rangle = 0 \), showing that \( x + y \in V^\perp \) as well as \( \lambda x \in V^\perp \), implying that \( V^\perp \) is a linear subspace.

(b): If \( v \in V \) and \( v \in V^\perp \), then \( \langle v, v \rangle = 0 \) by the definition of \( V^\perp \). However, \( \langle v, v \rangle = 0 \) implies \( v = 0 \) according to A.1(i).

(c): Suppose \( n \in \mathbb{N} \) and \( \lambda_1, \ldots, \lambda_n \in \mathbb{C} \) together with \( x_{a_1}, \ldots, x_{a_n} \) are such that \( \sum_{k=1}^n \lambda_k x_{a_k} = 0 \). Then, for each \( j \in \{1, \ldots, n\} \), the relations \( \langle x_{a_k}, x_{a_j} \rangle = 0 \) for each \( k \neq j \) imply 0 = \( \langle 0, x_{a_j} \rangle = (\sum_{k=1}^n \lambda_k x_{a_k}, x_{a_j}) = \sum_{k=1}^n \lambda_k \langle x_{a_k}, x_{a_j} \rangle = \lambda_j \langle x_{a_j}, x_{a_j} \rangle \), which yields \( \lambda_j = 0 \) by A.1(i). Thus, we have shown that \( \lambda_j = 0 \) for each \( j \in \{1, \ldots, n\} \), which establishes that the \( x_\alpha \) are all linearly independent.

(d): If \((x_\alpha)_{\alpha \in I}\) is an orthonormal basis, then the \( x_\alpha \) are linearly independent by (c) (since \( \|x_\alpha\| = 1 \) for each \( \alpha \in I \)). In particular, \( \dim X = n \) implies \( \# I \leq n \). As Def. A.3(b) implies \( \text{span}\{x_\alpha : \alpha \in I\} = X \) (here we use \( n < \infty \), otherwise the sums from Def. A.3(b) could have infinitely many nonzero terms), we see that \((x_\alpha)_{\alpha \in I}\) is a basis in the usual sense of linear algebra. Conversely, if \((x_\alpha)_{\alpha \in I}\) is a basis in the usual sense of linear algebra, then, for each \( x \in X \), there are \( \lambda_\alpha \in \mathbb{C}, \alpha \in I \), such that \( x = \sum_{\alpha \in I} \lambda_\alpha x_\alpha \). Thus,

\[
\sum_{\alpha \in I} \langle x, x_\alpha \rangle x_\alpha = \sum_{\alpha \in I} \left( \sum_{\beta \in I} \lambda_\beta x_\beta, x_\alpha \right) x_\alpha = \sum_{\alpha \in I} \sum_{\beta \in I} \lambda_\beta \langle x_\beta, x_\alpha \rangle x_\alpha \\
= \sum_{\alpha \in I} \sum_{\beta \in I} \lambda_\beta \delta_{\beta, \alpha} x_\alpha = \sum_{\alpha \in I} \lambda_\alpha x_\alpha = x, \tag{A.4}
\]

showing that \((x_\alpha)_{\alpha \in I}\) is an orthonormal basis. ■

### A.2 Unitary Matrices

The set of unitary matrices is, in many respects, the analogue for matrices over the field of complex numbers to the set of orthogonal matrices over the field of real numbers (cf. [Phi17b, Sec. H.2]). Before we get to unitary matrices, we provide some general notation for matrices and sets of matrices.

**Notation A.5.** For \( m, n \in \mathbb{N} \), let \( \mathcal{M}(m, n, F) \) denote the set of \( m \times n \) matrices over the field \( F \), where \( F = \mathbb{R} \) and \( F = \mathbb{C} \) are the fields of interest for this class. The abbreviation \( \mathcal{M}(n, F) := \mathcal{M}(n, n, F) \) is used for the set of quadratic \( n \times n \) matrices over \( F \).
Notation A.6. If \( A = (a_{kj})_{(k,j)\in\{1,...,m\}\times\{1,...,n\}} \in \mathcal{M}(m,n,\mathbb{C}) \), then

\[
\overline{A} := (\overline{a}_{kj})_{(k,j)\in\{1,...,m\}\times\{1,...,n\}} \in \mathcal{M}(m,n,\mathbb{C}),
\]

\( A^t := (a_{kj})_{(j,k)\in\{1,...,n\}\times\{1,...,m\}} \in \mathcal{M}(n,m,\mathbb{C}), \)

\( A^* := (\overline{A})^t \in \mathcal{M}(n,m,\mathbb{C}) \),

(A.5a) (A.5b) (A.5c)

where \( \overline{A} \) is called the \textit{(complex) conjugate} of \( A \), \( A^t \) is the \textit{transpose} of \( A \), and \( A^* \) is the conjugate transpose, also called the \textit{adjoint}, of \( A \).

Definition A.7. Let \( n \in \mathbb{N} \). Then \( U \in \mathcal{M}(n,\mathbb{C}) \) is called \textit{unitary} if, and only if, \( U \) is invertible with \( U^{-1} = U^* \). The set of all unitary \( n \times n \) matrices is denoted by \( U(n) \).

Theorem A.8. Let \( n \in \mathbb{N} \). For \( U \in \mathcal{M}(n,\mathbb{C}) \), the following statements are equivalent:

(i) \( U \) is unitary.

(ii) \( U^* \) is unitary.

(iii) The columns of \( U \) form an orthonormal basis of \( \mathbb{C}^n \) with respect to the inner product of Ex. A.2.

(iv) The rows of \( U \) form an orthonormal basis of \( \mathbb{C}^n \) with respect to the inner product of Ex. A.2.

(v) \( U^t \) is unitary.

(vi) \( U \) is unitary.

(vii) \( \langle Ux, Uy \rangle = \langle x, y \rangle \) holds for each \( x, y \in \mathbb{C}^n \), where \( \langle \cdot, \cdot \rangle \) denotes the inner product of Ex. A.2.

(viii) \( U \) is isometric with respect to the 2-norm \( \| \cdot \|_2 \) of Ex. A.2, i.e. \( \|Ux\|_2 = \|x\|_2 \) for each \( x \in \mathbb{C}^n \).

Proof. \( "(i)\Leftrightarrow(ii)" : U^{-1} = U^* \) is equivalent to \( \text{Id} = UU^* \), which is equivalent to \( (U^*)^{-1} = U^* \).

\( "(i)\Leftrightarrow(iii)" : U^{-1} = U^* \) implies

\[
\left( \begin{array}{c}
\sum_{l=1}^{n} u_{lk} \bar{u}_{lj} = 0 \\
\sum_{l=1}^{n} u_{lj}^* u_{lk} = \begin{cases}
0 & \text{for } k \neq j, \\
1 & \text{for } k = j,
\end{cases}
\end{array} \right)
\]

(A.6)

showing that the columns of \( U \) form an orthonormal system, i.e. \( n \) linearly independent vectors according to Lem. A.4(c), i.e. an orthonormal basis of \( \mathbb{C}^n \) according to Lem. A.4(d). Conversely, if the columns of \( U \) form an orthonormal basis of \( \mathbb{C}^n \), then they satisfy (A.6), which implies \( U^*U = \text{Id} \).
“(i)⇔(iv)”: $U^{-1} = U^*$ implies
\[
\langle \left( \begin{array}{c} u_{k1} \\ \vdots \\ u_{kn} \end{array} \right), \left( \begin{array}{c} u_{j1} \\ \vdots \\ u_{jn} \end{array} \right) \rangle = \sum_{l=1}^{n} u_{kl} \overline{u}_{jl} = \sum_{l=1}^{n} u_{kl} u_{lj}^* = \begin{cases} 0 & \text{for } k \neq j, \\ 1 & \text{for } k = j, \end{cases}
\]
showing that the rows of $U$ form an orthonormal system, i.e. $n$ linearly independent vectors according to Lem. A.4(c), i.e. an orthonormal basis of $\mathbb{C}^n$ according to Lem. A.4(d). Conversely, if the rows of $U$ form an orthonormal basis of $\mathbb{C}^n$, then they satisfy (A.7), which implies $UU^* = \text{Id}$.

“(i)⇔(v)”: Since the rows of $U$ are the columns of $U^\dagger$, the equivalence of (i) and (v) is immediate from (iii) and (iv).

“(i)⇔(vi)”: Since $U = (U^*)^\dagger$, the equivalence of (i) and (vi) is immediate from (ii) and (v).

“(i)⇒(vii)”: For each $x,y \in \mathbb{C}^n$: $\langle Ux, Uy \rangle = \langle U^*Ux, y \rangle = \langle \text{Id} x, y \rangle = \langle x, y \rangle$.

“(vii)⇒(viii)”: For each $x \in \mathbb{C}^n$: $\|Ux\|_2 = \langle Ux, Ux \rangle = \langle x, x \rangle = \|x\|_2$.

“(viii)⇒(i)”: The equality $\langle Ux, Uy \rangle = \langle U^*Ux, y \rangle = \langle \text{Id} x, y \rangle = \langle x, y \rangle$ holding for each $x,y \in \mathbb{C}^n$ implies $U^*U = \text{Id}$, i.e. $U$ is unitary.

\[\Box\]

**Proposition A.9.** Let $n \in \mathbb{N}$.

(a) If $A, B \in U(n)$, then $AB \in U(n)$.

(b) $U(n)$ constitutes a group.

**Proof.** (a): If $A, B \in U(n)$, then
\[
(AB)^* AB = B^* A^* AB = B^* \text{Id} B = B^* B = \text{Id},
\]
showing $AB \in U(n)$.

(b): Due to (a) and Th. A.8(ii), $U(n)$ is a subgroup of the general linear group $GL(n)$ of all invertible complex $n \times n$ matrices.

\[\Box\]

\section{Lipschitz Continuity}

**Definition B.1.** Let $m,n \in \mathbb{N}$, $G \subseteq \mathbb{R} \times \mathbb{K}^m$, and $f : G \longrightarrow \mathbb{K}^n$.

(a) The function $f$ is called \textit{(globally) Lipschitz continuous} or just \textit{(globally) Lipschitz} with respect to $y$ if, and only if,
\[
\exists L \geq 0 \quad \forall (x,y),(x,\bar{y}) \in G \quad \|f(x,y) - f(x,\bar{y})\| \leq L\|y - \bar{y}\|. \tag{B.1}
\]
(b) The function \( f \) is called \textit{locally Lipschitz continuous} or just \textit{locally Lipschitz} with respect to \( y \) if, and only if, for each \((x_0, y_0) \in G\), there exists a (relative) open set \( U \subseteq G \) such that \((x_0, y_0) \in U \) (i.e. \( U \) is a (relative) open neighborhood of \((x_0, y_0)\)) and \( f \) is Lipschitz continuous with respect to \( y \) on \( U \), i.e. if, and only if,

\[
\forall (x_0, y_0) \in G \quad \exists \exists L \geq 0 \quad \forall (x, y), (x, y) \in U \quad \|f(x, y) - f(x, \bar{y})\| \leq L\|y - \bar{y}\|.
\]

(B.2)

The number \( L \) occurring in (a),(b) is called \textit{Lipschitz constant}. The norms on \( \mathbb{K}^m \) and \( \mathbb{K}^n \) in (a),(b) are arbitrary. If one changes the norms, then one will, in general, change \( L \), but not the property of \( f \) being (locally) Lipschitz.

\textbf{Caveat B.2.} It is emphasized that \( f : G \rightarrow \mathbb{K}^n, (x, y) \mapsto f(x, y) \), being Lipschitz with respect to \( y \) does \textit{not} imply \( f \) to be continuous: Indeed, if \( I \subseteq \mathbb{R}, \emptyset \neq A \subseteq \mathbb{K}^m \), and \( g : I \rightarrow \mathbb{K}^n \) is an arbitrary \textit{discontinuous} function, then \( f : I \times A \rightarrow \mathbb{K}^n \), \( f(x, y) := g(x) \) is not continuous, but satisfies (B.1) with \( L = 0 \).

While the local neighborhoods \( U \), where a function locally Lipschitz (with respect to \( y \)) is actually Lipschitz continuous (with respect to \( y \)) can be very small, we will now show that a continuous function is locally Lipschitz (with respect to \( y \)) on \( G \) if, and only if, it is Lipschitz continuous (with respect to \( y \)) on \textit{every} compact set \( K \subseteq G \).

\textbf{Proposition B.3.} Let \( m, n \in \mathbb{N}, G \subseteq \mathbb{R} \times \mathbb{K}^m \), and \( f : G \rightarrow \mathbb{K}^n \) be continuous. Then \( f \) is locally Lipschitz with respect to \( y \) if, and only if, \( f \) is (globally) Lipschitz with respect to \( y \) on \textit{every} compact subset \( K \) of \( G \).

\textbf{Proof.} First, assume \( f \) is not locally Lipschitz with respect to \( y \). Then there exists \((x_0, y_0) \in G\) such that

\[
\forall N \in \mathbb{N} \quad \exists (x_N, y_{N,1}), (x_N, y_{N,2}) \in G \cap \mathcal{B}(x_0, y_0) \quad \|f(x_N, y_{N,1}) - f(x_N, y_{N,2})\| > N\|y_{N,1} - y_{N,2}\|.
\]

(B.3)

The set

\[
K := \{(x_0, y_0)\} \cup \{(x_N, y_{N,j}) : N \in \mathbb{N}, j \in \{1, 2\}\}
\]

is clearly a compact subset of \( G \) (e.g. by the Heine-Borel property of compact sets, since every open set containing \((x_0, y_0)\) must contain all, but finitely many, of the elements of \( K \)). Due to (B.3), \( f \) is not (globally) Lipschitz with respect to \( y \) on the compact set \( K \) (so, actually, continuity of \( f \) was not used for this direction).

Conversely, assume \( f \) to be locally Lipschitz with respect to \( y \), and consider a compact subset \( K \) of \( G \). Then, for each \((x, y) \in K\), there is some (relatively) open \( U_{(x, y)} \subseteq G \) with \((x, y) \in U_{(x, y)}\) and such that \( f \) is Lipschitz with respect to \( y \) in \( U_{(x, y)} \). By the Heine-Borel property of compact sets, there are finitely many \( U_1 := U_{(x_1, y_1)}, \ldots, U_N := U_{(x_N, y_N)} \), \( N \in \mathbb{N} \), such that

\[
K \subseteq \bigcup_{j=1}^N U_j.
\]

(B.4)
For each \( j = 1, \ldots, N \), let \( L_j \) denote the Lipschitz constant for \( f \) on \( U_j \) and set \( L' := \max\{L_1, \ldots, L_N\} \). As \( f \) is assumed continuous and \( K \) is compact, we have
\[
M := \max\{\|f(x,y)\| : (x,y) \in K\} < \infty. \tag{B.5}
\]

Using the compactness of \( K \) once again, there exists a Lebesgue number \( \delta > 0 \) for the open cover \((U_j)_{j \in \{1, \ldots, N\}}\) of \( K \), i.e. \( \delta > 0 \) such that
\[
\forall \, (x,y), (x,\bar{y}) \in K \quad (\|y - \bar{y}\| < \delta \implies \exists \, j \in \{1, \ldots, N\} \; \{x,y\}, \{x,\bar{y}\} \subseteq U_j). \tag{B.6}
\]

Define \( L := \max\{L', 2M/\delta\} \). Then, for every \((x,y), (x,\bar{y}) \in K\):
\[
\begin{align*}
\|y - \bar{y}\| < \delta & \implies \|f(x,y) - f(x,\bar{y})\| \leq L_j\|y - \bar{y}\| \leq L\|y - \bar{y}\|, \tag{B.7a} \\
\|y - \bar{y}\| \geq \delta & \implies \|f(x,y) - f(x,\bar{y})\| \leq 2M = \frac{2M\delta}{\delta} \leq L\|y - \bar{y}\|, \tag{B.7b}
\end{align*}
\]
completing the proof that \( f \) is Lipschitz with respect to \( y \) on \( K \).

The following Prop. B.4 provides a useful sufficient condition for \( f : G \longrightarrow K^n, \; G \subseteq \mathbb{R} \times K^m \) open, to be locally Lipschitz with respect to \( y \):

**Proposition B.4.** Let \( m,n \in \mathbb{N} \), let \( G \subseteq \mathbb{R} \times K^m \) be open, and \( f : G \longrightarrow K^n \). A sufficient condition for \( f \) to be locally Lipschitz with respect to \( y \) is \( f \) being continuously (real) differentiable with respect to \( y \), i.e., \( f \) is locally Lipschitz with respect to \( y \) provided that all partials \( \partial_{y_k} f_l; k, l = 1, \ldots, n \) \((\partial_{y_k} f_l, \partial_{y_{k,2}} f_l \text{ for } K = \mathbb{C})\) exist and are continuous.

**Proof.** We consider the case \( K = \mathbb{R} \); the case \( K = \mathbb{C} \) is included by using the identifications \( C^m \cong \mathbb{R}^{2m} \) and \( C^n \cong \mathbb{R}^{2n} \). Given \((x_0, y_0) \in G\), we have to show \( f \) is Lipschitz with respect to \( y \) on some open set \( U \subseteq G \) with \((x_0, y_0) \in U\). Since \( G \) is open,
\[
\exists \, b > 0 \quad B := \{(x,y) \in \mathbb{R} \times \mathbb{R}^m : |x - x_0| \leq b \text{ and } \|y - y_0\|_1 \leq b\} \subseteq G,
\]
where \( \| \cdot \|_1 \) denotes the 1-norm on \( \mathbb{R}^m \). Since the \( \partial_{y_k} f_l, (k,l) \in \{1, \ldots, m\} \times \{1, \ldots, n\} \), are all continuous on the compact set \( B \),
\[
M := \max\{|\partial_{y_k} f_l(x,y)| : (x,y) \in B, \; (k,l) \in \{1, \ldots, m\} \times \{1, \ldots, n\}\} < \infty. \tag{B.8}
\]
Applying the mean value theorem to the \( n \) components of the function
\[
f_x : \{y \in \mathbb{R}^m : (x,y) \in B\} \longrightarrow \mathbb{R}^n, \quad f_x(y) := f(x,y),
\]
we obtain \( \eta_1, \ldots, \eta_n \in \mathbb{R}^m \) such that
\[
f_l(x,y) - f_l(x,\bar{y}) = \sum_{k=1}^m \partial_{y_k} f_l(x,\eta_l)(y_k - \bar{y}_k), \tag{B.9}
\]
and, thus,
\[
\left\| f(x, y) - f(x, \bar{y}) \right\|_1 = \sum_{l=1}^{n} |f_l(x, y) - f_l(x, \bar{y})|
\]
\[
\forall (x,y),(x\bar{y}) \in B \quad (B.8),(B.9) \sum_{l=1}^{n} \sum_{k=1}^{m} M|y_k - \bar{y}_k| = \sum_{l=1}^{n} M\left|y - \bar{y}\right|_1 = nM\left|y - \bar{y}\right|_1,
\]
i.e. \( f \) is Lipschitz with respect to \( y \) on \( B \) (where
\[
\{ (x, y) \in \mathbb{R} \times \mathbb{R}^m : |x - x_0| < b \text{ and } \|y - y_0\|_1 < b \} \subseteq B
\]
is an open neighborhood of \((x_0, y_0)\), showing \( f \) is locally Lipschitz with respect to \( y \). ■

C Ordinary Differential Equations (ODE)

Proposition C.1. Let \( G \subseteq \mathbb{R} \times \mathbb{K}^n \) be open, \( n \in \mathbb{N} \), and \( f : G \to \mathbb{K}^n \). Let \( I \subseteq \mathbb{R} \) be an open interval and let \( \phi : I \to \mathbb{K}^n \) be a solution to the ODE
\[
y' = f(x, y).
\]
\( (C.1) \)

If \( f \) has continuous partials up to order \( k \in \mathbb{N}_0 \), then \( \phi \in C^{k+1}(I, \mathbb{K}^n) \).

Proof. As we are only considering real differentiability, we may assume \( \mathbb{K} = \mathbb{R} \) without loss of generality (the case \( \mathbb{K} = \mathbb{C} \) is included via the identification \( \mathbb{K}^n \cong \mathbb{R}^{2n} \)). By the chain rule, if \( g : G \to \mathbb{R} \) is differentiable and
\[
\alpha : I \to \mathbb{R}, \quad \alpha(x) := g(x, \phi(x)),
\]
then \( \alpha \) is differentiable with
\[
\alpha' : I \to \mathbb{R}, \quad \alpha'(x) = \partial_x g(x, \phi(x)) + \sum_{\nu=1}^{n} \partial_{y_{\nu}} g(x, \phi(x)) \phi_{\nu}'(x)
\]
\[
= \partial_x g(x, \phi(x)) + \sum_{\nu=1}^{n} \partial_{y_{\nu}} g(x, \phi(x)) f_{\nu}(x, \phi(x)).
\]
\( (C.3) \)

Thus, an induction shows that, for each \( 1 \leq j \leq k + 1 \) and each \( x \in I \), \( \phi^{(j)}(x) \) is a polynomial in partial derivatives of order at most \( j - 1 \) of \( f \), all evaluated at \((x, \phi(x))\). In particular, if all partials up to order \( k \) of \( f \) are continuous, then so is \( \phi^{(k)} \). ■

References

REFERENCES

[DB08] PETER DEUFLHARD and FOLKMAR BORNEMANN. Numerische Mathematik 2, 3rd ed. Walter de Gruyter, Berlin, Germany, 2008 (German).


[Heu08] HARRO HEUSER. Lehrbuch der Analysis Teil 2, 14th ed. Vieweg+Teubner, Wiesbaden, Germany, 2008 (German).


