Volker John

Numerical Methods for Ordinary Differential Equations

1. Explicit One-Step Methods 3
2. Numerical Methods for Stiff Ordinary Differential Equations 27
3. Multi-StepMethods 49
4. Summary and Outlook 73
Appendix A: Topics on the Theory of Ordinary Differential Equations 77
References 99

Chapter 1
Explicit One-Step Methods

Remark 1.1. Contents. This course presents methods for the numerical solution of explicit systems of initial value problems for ordinary differential equations of first order

$$\boldsymbol{y}'(x) = \boldsymbol{f}(x, \boldsymbol{y}(x)), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0.$$

For the most part, only initial value problems for scalar ordinary differential equations of first order

$$y'(x) = f(x, y(x)), \quad y(x_0) = y_0, \tag{1.1}$$

are considered, for simplicity of presentation. The extension of the results and the methods to systems is generally straightforward.

It will be always assumed that there is a unique solution of the initial value problem in a neighborhood of the initial value. In applications, the independent variable is often the time. □

1.1 Consistency and Convergence

Definition 1.2. Grid, step size. A grid is a decomposition I_h of the interval $I = [x_0, x_e]$

$$I_h = \{x_0, x_1, \ldots, x_N = x_e\}$$

with $x_0 < x_1 < \ldots < x_N$. The differences between neighboring grid points $h_k = x_{k+1} - x_k$ are called step sizes. For an equidistant grid, the notation $h = h_k$ will be used for the step size, see Figure 1.1. □

Remark 1.3. Explicit and implicit methods. Let $y(x_k)$ denote the solution of (1.1) in the node x_k and y_k a numerical approximation of $y(x_k)$. A numerical

Fig. 1.1 Equidistant grid.

method for the solution of (1.1) on a grid I_h is called explicit, if an approximation y_{k+1} in x_{k+1} can be calculated directly from already computed values y_i, $i \leq k$. Otherwise, the method is called implicit method. Implicit methods require in each step the solution of a generally nonlinear equation for computing y_{k+1}. □

Definition 1.4. One-step method, incremental function. A one-step method for the computation of an approximation y_{k+1} of the solution of (1.1) on a grid I_h has the form

$$y_{k+1} = y_k + h_k \Phi\left(x, y, h_k\right), \; k = 0, 1, \ldots, \quad y_0 = y(x_0). \tag{1.2}$$

Here, $\Phi(\cdot, \cdot, \cdot)$ is called incremental function of the one-step method. □

Example 1.5. One-step methods, incremental functions. The explicit or forward Euler method

$$y_{k+1} = y_k + h_k f\left(x_k, y_k\right), \; k = 0, 1, 2, \ldots, \quad y_0 = y(x_0),$$

is an explicit one-step method with the incremental function

$$\Phi\left(x, y, h_k\right) = f\left(x_k, y_k\right).$$

The computation of y_{k+1} requires only the substitution of already computed values in the function $f(x, y)$ from the initial value problem (1.1).

The implicit or backward Euler method

$$y_{k+1} = y_k + h_k f\left(x_{k+1}, y_{k+1}\right), \; k = 0, 1, 2, \ldots, \quad y_0 = y(x_0),$$

is an implicit one-step method with the incremental function

$$\Phi\left(x, y, h_k\right) = f\left(x_{k+1}, y_{k+1}\right).$$

One has to solve an equation for computing y_{k+1}. The complexity of this step depends on $f(x, y)$. □

Remark 1.6. Representation of implicit one-step methods. Explicit one-step methods require only that known values are inserted in the incremental function. Hence, their incremental function can be written finally in the form

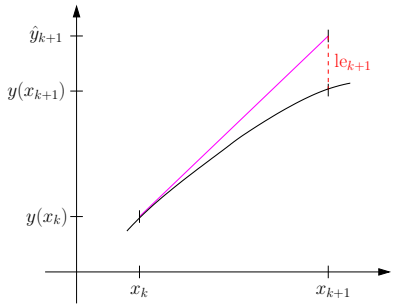

Fig. 1.2 The local error.

$\Phi(x_k, y_k, h_k)$. For the considerations in this section, one can adopt the point of view that also implicit one-step methods can be written as explicit one-step methods, because the data for the nonlinear equation are x_k, y_k, and h_k. However, generally one does not know the concrete form of the incremental function. □

Example 1.7. Incremental function of the implicit Euler method. The incremental function of the implicit Euler method on an equidistant grid can be written in the form

$$\Phi(x, y, h) = f(x + h, y + h\Phi(x, y, h)),$$

which allows formally the representation of this method as explicit one-step scheme. □

Definition 1.8. Local error. Let $\hat{y}_{k+1}$ be the result of one step of an explicit one-step method (1.2) with the initial value $y(x_k)$, i.e.,

$$\hat{y}_{k+1} = y(x_k) + h_k\Phi(x_k, y(x_k), h_k).$$

Then,

$$\begin{aligned} \mathrm{le}(x_{k+1}) = \mathrm{le}_{k+1} &= y(x_{k+1}) - \hat{y}_{k+1} \\ &= y(x_{k+1}) - (y(x_k) + h_k\Phi(x_k, y(x_k), h_k)) \end{aligned} \tag{1.3}$$

is called local error, see Figure 1.2 □

Remark 1.9. The local error. In the literature, sometimes

$$\frac{y(x_{k+1}) - y(x_k)}{h_k} - \Phi(x_k, y(x_k), h_k)$$

is defined to be the local error.

For the local error, one starts from the solution of the initial value problem and considers the error after one step of the numerical method.

One should require for a reasonable method that the local error is small in an appropriate sense. □

Definition 1.10. Consistent method. Let $y(x)$ be the solution of the initial value problem (1.1), $h_{\max} = \max_k h_k$, and

$$S := \{(x, y) \ : \ x \in [x_0, x_e], \ y \in \mathbb{R}\} .$$

The one-step method (1.2) is said to be consistent, if for all $f \in C(S)$, which satisfy in S a Lipschitz condition with respect to y, it holds

$$\lim_{h_{\max} \to 0} \left(\max_{x_k \in I_h} \frac{|\mathrm{le}\,(x_{k+1})|}{h_k} \right) = 0$$

or

$$\lim_{h_{\max} \to 0} \left(\max_{x_k \in I_h} |f(x_k, y(x_k)) - \Phi\,(x_k, y(x_k), h_k)| \right) = 0. \tag{1.4}$$

Both conditions are equivalent, compare Remark 1.11. □

Remark 1.11. Approximation of the derivative with the incremental function. For bounded incremental functions, it is obvious that the local error converges to zero if $h_{\max} \to 0$, because in this case it holds $h_k \to 0$ and $\hat{y}_{k+1} \to y(x_k)$, such that this statement follows from (1.3). Consistency requires more, namely that the incremental function approximates the derivative of the solution sufficiently well. Applying (1.3) and (1.1) yields

$$\begin{aligned} \frac{\mathrm{le}\,(x_{k+1})}{h_k} &= \frac{y\,(x_{k+1}) - y(x_k)}{h_k} - \Phi\,(x_k, y(x_k), h_k) \\ &\approx y'(x_k) - \Phi\,(x_k, y(x_k), h_k) \\ &= f\,(x_k, y_k) - \Phi\,(x_k, y(x_k), h_k)\,, \end{aligned}$$

compare (1.4). □

Example 1.12. Consistency of the explicit Euler method. For the explicit Euler method, it is $\Phi\,(x_k, y(x_k), h_k) = f\,(x_k, y(x_k))$. Hence, condition (1.4) from Definition 1.10 is satisfied and the method is consistent. □

Remark 1.13. Quality of the approximation of the incremental function. For practical purposes, not only the consistency itself but the quality of the approximation of the derivative by the incremental function is essential. The quality allows a comparison of different one-step methods. For simplicity of presentation, let $h_k = h$ for all k. □

Definition 1.14. Order of consistency. An explicit one-step method (1.2) has the consistency order $p \in \mathbb{N}$, if p is the largest natural number such that for all functions $f \in C(S)$, which satisfy a Lipschitz condition with respect to y, it holds

$$|\mathrm{le}\,(x_k + h)| \leq Ch^{p+1}$$

for all $x_k \in I_h$, for all I_h with $h \in (0, H]$, and with the constant $C > 0$ being independent of h. The constant C might depend on derivatives of $y(x)$, on $f(x, y)$, and on partial derivatives of $f(x, y)$. □

Example 1.15. Order of consistency of the explicit Euler method. Consider the explicit Euler method and assume that the function $y(x)$ is two times continously differentiable. Then, it follows with Taylor series expansion and using the differential equation that

$$\begin{aligned}
|\mathrm{le}\,(x_k + h)| &= |y(x_k + h) - \hat{y}_{k+1}| \\
&= |y(x_k) + hy'(x_k) + \frac{h^2}{2} y''(x_k + \theta h) - y(x_k) - h \underbrace{f\,(x_k, y(x_k))}_{=y'(x_k)}| \\
&= \frac{h^2}{2} \left|y''(x_k + \theta h)\right| \leq \frac{h^2}{2} \left\|y\right\|_{C^2([x_0, x_e])},
\end{aligned}$$

with $\theta \in (0, 1)$. Since there is no way to replace the term on the right-hand side by a term with a larger power of h, the method has consistency order 1. □

Remark 1.16. Consistency and convergence. The consistency is a local property of a one-step method. For practical purposes, it is important that the computed solution converges to the analytic solution if the grid becomes finer and finer. Of course, the order of convergence is of importance, too.

It will be shown that, under certain conditions, the convergence of a one-step method follows from its consistency and that the order of convergence equals the consistency order. □

Definition 1.17. Convergent method, order of convergence. A one-step method (1.2) converges for the initial value problem (1.1) on the interval $I = [x_0, x_e]$, if for each sequence of grids $\{I_h\}$ with $h_{\max} = \max_{h_k} h_k \to 0$ for the global error

$$e(x_k, h) = y(x_k) - y_k, \quad x_k \in I_h,$$

it follows that

$$\max_{x_k \in I_h} |e(x_k, h)| \to 0 \quad \text{for} \quad h_{\max} \to 0.$$

The one-step method has the order of convergence p^*, if p^* is the largest natural number such that for all step lengths $h_{\max} \in (0, H]$, it holds

$$|e(x_k, h)| \leq Ch_{\max}^{p^*} \quad \forall\ x_k \in I_h,$$

where $C > 0$ is independent of $h_{\max}$. □

Lemma 1.18. Estimate for a sequence of real numbers. *Assume that for real numbers x_n, $n = 0, 1, \ldots$, the inequality*

$$|x_{n+1}| \leq (1+\delta)\,|x_n| + \beta$$

holds with constants $\delta > 0$, $\beta \geq 0$. Then, it holds that

$$|x_n| \leq e^{n\delta}\,|x_0| + \frac{e^{n\delta}-1}{\delta}\beta, \quad n = 0, 1, \ldots.$$

Proof. With induction, problem for exercises. ■

Theorem 1.19. Connection of consistency and convergence. *Let $y(x)$ be the solution of the initial value problem (1.1) with $f \in C(S)$. Let a Lipschitz condition hold for the second argument of the incremental function*

$$|\Phi(x, y_1, h) - \Phi(x, y_2, h)| \leq M\,|y_1 - y_2| \quad \forall\ (x, y) \in S,\ h \in (0, H], \tag{1.5}$$

with $M \in \mathbb{R}$ fixed. Assume that for the local error the estimate

$$|\mathrm{le}\,(x_k + h)| \leq Ch^{p+1} \quad \forall\ x_k \in I_h, h \in (0, H] \tag{1.6}$$

is valid and assume that $y_0 = y(x_0)$.

Then, it follows for the global error that

$$|e(x_{k+1}, h)| \leq C\frac{e^{M\left(x_{k+1}-x_0\right)}-1}{M}h^p,$$

where C is independent of h.

Proof. It holds that

$$\begin{aligned} y_{k+1} &= y_k + h\Phi\,(x_k, y_k, h)\,, \\ y(x_{k+1}) &= y(x_k) + h\Phi\,(x_k, y(x_k), h) + \mathrm{le}\,\left(x_{k+1}\right), \quad k = 0, 1, \ldots. \end{aligned}$$

Then, it follows with the triangle inequality, the assumption on the local error (1.6), and the Lipschitz condition of the incremental function (1.5) that

$$\begin{aligned} \left|e(x_{k+1}, h)\right| &= \left|y(x_{k+1}) - y_{k+1}\right| \\ &= \left|y(x_k) - y_k + \mathrm{le}\,\left(x_{k+1}\right) + h\big(\Phi\,(x_k, y(x_k), h) - \Phi\,(x_k, y_k, h)\big)\right| \\ &= \left|e(x_k, h) + \mathrm{le}\,\left(x_{k+1}\right) + h\big(\Phi\,(x_k, y(x_k), h) - \Phi\,(x_k, y_k, h)\big)\right| \\ &\leq |e(x_k, h)| + \left|\mathrm{le}\,\left(x_{k+1}\right)\right| + h\,|\Phi\,(x_k, y(x_k), h) - \Phi\,(x_k, y_k, h)| \\ &\leq |e(x_k, h)| + Ch^{p+1} + hM\,|y(x_k) - y_k| \\ &= (1 + hM)\,|e(x_k, h)| + Ch^{p+1}. \end{aligned}$$

This sequence of inequalities has the form that was considered in Lemma 1.18. One obtains with $e(x_0) = 0$

$$\left|e(x_{k+1},h)\right| \le e^{(k+1)hM}\left|e(x_0)\right| + C\frac{e^{(k+1)hM}-1}{hM}h^{p+1} = C\frac{e^{M(x_{k+1}-x_0)}-1}{M}h^p.$$

■

Remark 1.20. To Theorem 1.19.

- The consideration of a constant step length is only for simplicity of presentation. The result of the theorem holds also for non-constant step lengths with $h = \max_k h_k$.
- One-step methods compute an approximation y_k of the solution in the grid points x_k, $k = 0, 1, \ldots, N$. To enable a better comparison with the analytic solution, one connects these points linearly from (x_k, y_k) to (x_{k+1}, y_{k+1}). In this way, one obtains a piecewise linear approximation of the solution that is defined on $[x_0, x_e]$. This function is called $y^h(x)$. The considerations from above can be extended to $y^h(x)$.

□

1.2 Explicit Runge–Kutta Schemes

Remark 1.21. Idea. The Euler methods are only of first order. The idea of Runge[1]–Kutta[2] methods consists in using an incremental function $\Phi(x, y, h)$ that is a linear combination of values of $f(x, y)$ in different points. With this approach, one obtains methods of higher order for the cost of evaluating more values of $f(x, y)$.

This approach can be illustrated well at the integral equation that is equivalent to the initial value problem (1.1). For simplicity, let the right-hand side of (1.1) depend only on x. Then, the integral equation has the form

$$y(x) = y_0 + \int_{x_0}^{x} f(t)\ dt. \tag{1.7}$$

The idea of the Runge–Kutta methods consists in approximating the right-hand side by a quadrature rule, e.g., in the interval $[x_k, x_{k+1}]$ by

$$\int_{x_k}^{x_{k+1}} f(t)\ dt \approx h_k \sum_{j=1}^{s} b_j f\left(x_k + c_j h_k\right)$$

with the weights b_j and the nodes $x_k + c_j h$.

In the following, only $h_k = h$ for all k will be considered for the reason of simplicity. □

[1] Carle David Tolmé Runge (1856 – 1927)

[2] Martin Kutta (1867 – 1944)

Definition 1.22. Runge–Kutta methods, increments, and stages. A Runge–Kutta method has the form

$$y_{k+1} = y_k + h\Phi(x, y, h),\ k = 0, 1, \ldots, \quad y_0 = y(x_0),$$

where the incremental function is defined with the help of

$$K_i(x, y, h) = f\left(x_k + c_i h, y_k + h\sum_{j=1}^{s} a_{ij} K_j(x, y, h)\right)$$

by

$$\Phi(x, y, h) = \sum_{i=1}^{s} b_i K_i(x, y, h),$$

with $c_1, \ldots, c_s, b_1, \ldots, b_s, a_{ij} \in \mathbb{R}$, $i, j = 1, \ldots, s$. The quantities $K_i(x, y, h)$, $i = 1, \ldots, s$, are called increments. The natural number $s \in \mathbb{N}$ is the number of stages of the method.

An equivalent definition is as follows

$$y_{k+1}^{(i)} = y_k + h\sum_{j=1}^{s} a_{ij} f\left(x_k + c_j h, y_{k+1}^{(j)}\right), \tag{1.8}$$

$$\Phi(x, y, h) = \sum_{i=1}^{s} b_i f\left(x_k + c_i h, y_{k+1}^{(i)}\right). \tag{1.9}$$

The intermediate values $y_{k+1}^{(i)}$ are called stages. □

Remark 1.23. Butcher[3] tableau. For the reason of clarity, one writes a Runge–Kutta scheme in general in form of a tableau, the so-called Butcher tableau

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ c_3 & a_{31} & a_{32} & \cdots & a_{3s} \\ \vdots & \vdots & \vdots & & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \quad \Longrightarrow \quad \begin{array}{c|c} \mathbf{c} & A \\ \hline & \boldsymbol{b}^T \end{array}. \tag{1.10}$$

Here, $\mathbf{c}$ are the nodes, A is the matrix of the method, and $\boldsymbol{b}$ are the weights. □

Remark 1.24. Increments and Butcher tableau. For explicit Runge–Kutta schemes, the increments can be computed one after the other

[3] John C. Butcher, born. 1933

$$
\begin{aligned}
K_1(x,y,h) &= f(x_k, y_k),\\
K_2(x,y,h) &= f\left(x_k + c_2 h, y_k + h a_{21} K_1(x,y,h)\right),\\
&\vdots\\
K_s(x,y,h) &= f\left(x_k + c_s h, y_k + h \sum_{j=1}^{s-1} a_{sj} K_j(x,y,h)\right).
\end{aligned} \tag{1.11}
$$

The Butcher tableau has the form

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

□

Example 1.25. Explicit Euler scheme. The explicit Euler scheme is an explicit Runge–Kutta scheme with the Butcher tableau

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

In the integral equation, the approximation

$$
\int_{x_k}^{x_{k+1}} f(t, y(t))\, dt \approx h f\left(x_k, y_k(x_k)\right)
$$

is used, see the proof of the Theorem of Peano, lectures notes of Numerical Mathematics I. □

Theorem 1.26. Consistency of explicit Runge–Kutta schemes. *Let $f \in C(S)$, see Definition 1.10. An explicit Runge–Kutta scheme is consistent if and only if*

$$
\sum_{i=1}^{s} b_i = 1. \tag{1.12}
$$

Proof. From the continuity of $f(x,y)$ and the definition (1.11) of the increments of an explicit Runge–Kutta scheme, it follows that

$$
\lim_{h\to 0} K_i(x,y,h) = f(x_k, y(x_k)), \quad \forall\ (x,y) \in S,\ i = 1, \ldots, s,
$$

for the case that the initial value of this step is $y_k = y(x_k)$. The continuity of the absolute value function gives

$$
\lim_{h\to 0} \left| f(x_k, y(x_k)) - \Phi\left(x_k, y(x_k), h\right) \right| = \lim_{h\to 0} \left| f(x_k, y(x_k)) - \sum_{i=1}^{s} b_i K_i(x,y,h) \right|
$$

$$= \left| f(x_k, y(x_k)) - \sum_{i=1}^{s} b_i \lim_{h\to 0} K_i(x, y, h) \right|$$
$$= \left| f(x_k, y(x_k)) \left(1 - \sum_{i=1}^{s} b_i \right) \right| = 0$$

if and only if $\sum_{i=1}^{s} b_i = 1$. Hence, the condition (1.4) in Definition 1.10 is satisfied. ■

Theorem 1.27. Interpretation of the increments. *Let for the solution of* (1.1) *hold* $y \in C^2([x_0, x_e])$*, let* $f \in C(S)$*, and let* f *be Lipschitz continuous in the second argument. If* $y_k = y(x_k)$ *and*

$$c_i = \sum_{j=1}^{i-1} a_{ij}, \quad i \geq 2, \tag{1.13}$$

holds, then $K_i(x, y, h)$ *is an approximation of at least first order (of consistency) to* $y'(x_k + c_i h)$*, i.e.,*

$$y'(x_k + c_i h) - K_i(x, y, h) = \mathcal{O}\left(h^2\right).$$

Proof. The proof follows by induction.

$i = 2$. For $i = 2$, it follows with (1.1), the Lipschitz continuity, and Taylor series expansion that

$$\begin{aligned}
&\left| y'(x_k + c_2 h) - K_2(x, y, h) \right| \\
&= \left| f\big(x_k + c_2 h, y(x_k + c_2 h)\big) - f\big(x_k + c_2 h, y(x_k) + h a_{21} f(x_k, y(x_k))\big) \right| \\
&\leq L \left| y(x_k + c_2 h) - y(x_k) - h a_{21} f(x_k, y(x_k)) \right| \\
&= L \left| y(x_k) + c_2 h y'(x_k) + \mathcal{O}(h^2) - y(x_k) - h a_{21} y'(x_k) \right| \\
&= L \left| (c_2 - a_{21}) h y'(x_k) + \mathcal{O}(h^2) \right|.
\end{aligned}$$

Hence, in the case $c_2 = a_{21}$, the error is of order $\mathcal{O}(h^2)$.

$i > 2$. Let the asymptotic order of the errors be proved for all indices $2, \ldots, i-1$. Then, one gets in the same way as for $i = 2$

$$\begin{aligned}
&\left| y'(x_k + c_i h) - K_i(x, y, h) \right| \\
&= \left| f\left(x_k + c_i h, y(x_k + c_i h)\right) - f\left(x_k + c_i h, y(x_k) + h \sum_{j=1}^{i-1} a_{ij} K_j(x, y, h) \right) \right| \\
&\leq L \left| y(x_k + c_i h) - y(x_k) - h \sum_{j=1}^{i-1} a_{ij} K_j(x, y, h) \right| \\
&= L \left| y(x_k) + c_i h y'(x_k) + \mathcal{O}(h^2) - y(x_k) - h \sum_{j=1}^{i-1} \left(a_{ij} \left(y'(x_k + c_j h) + \mathcal{O}(h^2) \right) \right) \right| \\
&= L \left| c_i h y'(x_k) + \mathcal{O}(h^2) - h \sum_{j=1}^{i-1} \left(a_{ij} \left(y'(x_k) + \mathcal{O}(h) \right) \right) \right|
\end{aligned}$$

$$= L \left| h \left(c_i - \sum_{j=1}^{i-1} a_{ij} \right) y'(x_k) + \mathcal{O}(h^2) \right|.$$

The order of the error $\mathcal{O}(h^2)$ is given, if $c_i = \sum_{j=1}^{i-1} a_{ij}$. ■

Remark 1.28. Conditions on the coefficients for certain orders of convergence. The conditions from Theorems 1.26 and 1.27 are satisfied for many explicit Runge–Kutta schemes. The goal consists in determining the coefficients $b_1, \ldots, b_s$, and a_{ij} in such a way that one obtains an order of consistency as high as possible. The consistency order of a Runge–Kutta scheme with s stages can be derived from the Taylor series expansion of the local error. Let (1.12) be valid, then one obtains, e.g.,

- A Runge–Kutta scheme with the parameters $(A, \boldsymbol{b}, \boldsymbol{c})$ has at least consistency order $p = 2$ if
$$\sum_{j=1}^{s} b_j c_j = \frac{1}{2}. \tag{1.14}$$
This condition will be shown in Example 1.29 for $s = 2$.
- If in addition
$$\sum_{j=1}^{s} b_j c_j^2 = \frac{1}{3} \quad \text{and} \quad \sum_{j=1}^{s} b_j \sum_{k=1}^{s} a_{jk} c_k = \frac{1}{6}$$
hold, then the order of consistency is at least $p = 3$.

The proof of the last statement and conditions for even higher order consistency can be found in the literature, e.g. in (Strehmel & Weiner, 1995; Strehmel *et al.*, 2012, Section 2.4.2). □

Example 1.29. Runge–Kutta methods with 2 *stages.* For the investigation of 2-stage Runge–Kutta schemes, one considers for simplicity the so-called autonomous initial value problem

$$y'(x) = f(y(x)), \quad y(x_0) = y_0.$$

One has for the increments

$$\begin{aligned} K_1(y,h) &= f(y_k), \\ K_2(y,h) &= f\left(y_k + h a_{21} K_1(y_k,h)\right) = f\left(y_k + h a_{21} f(y_k)\right) \\ &= f(y_k) + h a_{21} f(y_k) \partial_y f(y_k) + \mathcal{O}(h^2). \end{aligned}$$

If the initial value is exact, it follows for the incremental function that

$$\begin{aligned} \Phi(y(x_k)) &= b_1 K_1(y,h) + b_2 K_2(y,h) \\ &= (b_1 + b_2) f(y(x_k)) + h b_2 a_{21} f(y(x_k)) \partial_y f(y(x_k)) + \mathcal{O}(h^2). \end{aligned} \tag{1.15}$$

The Taylor series expansion of the solution has the form

$$y(x_k+h) = y(x_k) + h \underbrace{y'(x_k)}_{=f(y(x_k))} + \frac{h^2}{2} y''(x_k) + \mathcal{O}\left(h^3\right).$$

One obtains with the chain rule

$$y''(x) = \frac{d}{dx} y'(x) = \frac{d}{dx} f(y(x)) = \partial_y f(y) y'(x) = \partial_y f(y) f(y(x)).$$

Now, it follows for the local error, using Taylor series expansion and (1.15), that

$$\begin{aligned}
\mathrm{le}(x_k+h) &= y(x_k+h) - y(x_k) - h\Phi(y(x_k)) \\
&= y(x_k) + hf(y(x_k)) + \frac{h^2}{2}\left(\partial_y f(y(x_k)) f(y(x_k))\right) + \mathcal{O}\left(h^3\right) - y(x_k) \\
&\quad -h\Big((b_1+b_2) f(y(x_k)) + hb_2 a_{21} f(y(x_k)) \partial_y f(y(x_k)) + \mathcal{O}(h^2)\Big) \\
&= h\big(1-(b_1+b_2)\big) f(y(x_k)) + h^2 \left(\frac{1}{2} - b_2 a_{21}\right) f(y(x_k)) \partial_y f(y(x_k)) \\
&\quad + \mathcal{O}\left(h^3\right).
\end{aligned}$$

To achieve an order of consistency as large as possible, the first two terms have to vanish. One obtains with the condition $c_2 = a_{21}$ that

$$b_1 + b_2 = 1, \quad b_2 a_{21} = \frac{1}{2} \iff b_2 c_2 = \frac{1}{2}.$$

The first equation is the general condition for consistency (1.12) and the second condition is exactly (1.14) for $s = 2$. These two conditions characterize all 2-stage explicit Runge–Kutta methods that possess consistency and convergence order 2

$$\begin{array}{c|cc} c_2 & c_2 & \\ \hline & 1 - \frac{1}{2c_2} & \frac{1}{2c_2} \end{array}, \quad \text{with } c_2 \neq 0.$$

In the case $c_2 = 1/2$, one obtains the method of Runge (1895)

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

This method corresponds with respect to the approximation of the integral in (1.7) to the application of the mid point rule.

For $c_2 = 1$, one gets the method of Heun[4] (1900)

[4] Karl Heun (1859 – 1929)

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

which corresponds to the use of the trapezoidal rule for the numerical quadrature in (1.7). □

Remark 1.30. Autonomous ordinary differential equations. Every explicit first order ordinary differential equation

$$\boldsymbol{y}'(x) = \boldsymbol{f}(x, \boldsymbol{y}(x))$$

can be transformed into an autonomous form

$$\tilde{\boldsymbol{y}}'(x) = \tilde{\boldsymbol{f}}\left(\tilde{\boldsymbol{y}}(x)\right) = \begin{pmatrix} \boldsymbol{f}(x, \boldsymbol{y}(x)) \\ 1 \end{pmatrix}$$

by introducing the function

$$\overline{y}(x) := x \quad \text{and} \quad \tilde{\boldsymbol{y}}(x) := \begin{pmatrix} \boldsymbol{y}(x) \\ \overline{y}(x) \end{pmatrix}$$

and noting that $(\boldsymbol{y}(x), x)$ are just the components of $\tilde{\boldsymbol{y}}(x)$. □

Theorem 1.31. Consistency and convergence of explicit Runge–Kutta methods. *Let $y(x)$ be the solution of the initial value problem* (1.1) *with $f \in C(S)$ and let $f(x, y)$ satisfy a Lipschitz condition in the second argument. Then, an explicit Runge–Kutta scheme that is consistent of order p converges also with order p.*

Proof. The incremental function of an explicit Runge–Kutta scheme is a linear combination of values of the right-hand side $f(x, y)$. Thus, the assumptions of Theorem 1.19 are satisfied, since the Lipschitz condition in this theorem follows from the required Lipschitz condition on the right-hand side of the differential equation. The statement of the theorem follows now directly from Theorem 1.19. ■

Remark 1.32. Explicit Runge–Kutta methods of higher order. Analogously to 2-stage methods, it is possible to derive conditions on the coefficients of an explicit Runge–Kutta scheme in order to construct methods of higher order. An important question is the minimal number of stages that is necessary to be able to reach a certain order. Some answers to this question are from Butcher (1963, 1965, 1985)

$$\begin{array}{c|cccccccc} p & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline \min s & 1 & 2 & 3 & 4 & 6 & 7 & 9 & 11 \end{array}.$$

□

Example 1.33. Classical Runge–Kutta scheme (1901). The so-called classical Runge–Kutta scheme has four stages and the Butcher tableau

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

It is based on the Simpson[5] rule. The center node of the Simpson rule is used twice, $c_2 = c_3$, but with a different second argument for the computation of the increments. This method is of fourth order. □

1.3 Step Length Control

Remark 1.34. Motivation. The considerations so far did not provide a way for estimating a good step length for solving a given initial value problem with prescribed accuracy and with as little work as possible.

- If the steps are too large, then the numerical solution might be too inaccurate.
- If the steps are too small, then the numerical simulation might take much longer than necessary.

A good step length depends certainly on the concrete problem and generally it will change within the considered interval. For these reasons, the step length should be controlled during the numerical simulation of the initial value problem.

A typical approach consists in computing two approximations of the solution in a node with different methods and to draw conclusions on the size of the local error based on the difference of these approximations. Of course, the consideration of the global error would be better. However, Theorem 1.19 shows that on the one hand, the global error is influenced by problem-dependent terms, like the length of the interval $[x_0, x_e]$ or the Lipschitz constant. On the other hand, the global error is expected to be small only if the local errors are small. □

1.3.1 The Richardson Method

Remark 1.35. Idea. Given a numerical method for solving an initial value problem and given a step length h. The Richardson[6] method consists of the following steps, see also Figure 1.3:

[5] Thomas Simpson (1710 – 1761)

[6] Lewis Fry Richardson (1881 – 1953)

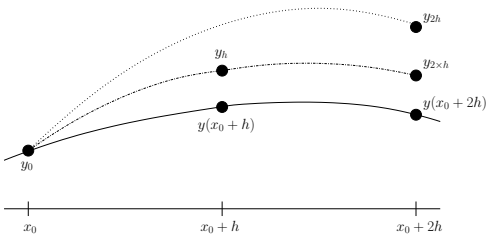

Fig. 1.3 Sketch of the Richardson method.

1. Starting from a node (x_0, y_0) and using a step length of $2h$, an approximation y_{2h} at the node $x_0 + 2h$ will be computed.
2. Two approximations y_h and $y_{2\times h}$ in $x_0 + h$ and $x_0 + 2h$ are computed with two steps of length h.
3. The step length will be controlled by comparing y_{2h} and $y_{2\times h}$.

In general, the more accurate approximation will be $y_{2\times h}$. In addition, it will be demonstrated that it is possible to improve the accuracy of $y_{2\times h}$ with the information obtained by this method. □

Example 1.36. Richardson method for an explicit 2*-stage Runge–Kutta method.* Consider an explicit 2-stage Runge–Kutta scheme. One obtains in the first step of the Richardson method, using (1.8), (1.9),

$$\begin{aligned} y_{2h}^{(1)} &= y_0, \\ y_{2h}^{(2)} &= y_0 + 2ha_{21}f(x_0, y_0), \\ y_{2h} &= y_0 + 2h\left[b_1K_1(x,y) + b_2K_2(x,y)\right] \\ &= y_0 + 2h\left[b_1 f(x_0, y_{2h}^{(1)}) + b_2 f\left(x_0 + 2c_2h, y_{2h}^{(2)}\right)\right], \end{aligned}$$

or written as Butcher tableau

$$\begin{array}{c|cc} 0 & & \\ 2c_2 & 2a_{21} & \\ \hline & 2b_1 & 2b_2 \end{array}.$$

Note that because of the step length $2h$, the weights sum up to 2.

The second step of the Richardson method yields

$$\begin{aligned} y_{2\times h}^{(1)} &= y_0, \\ y_{2\times h}^{(2)} &= y_0 + ha_{21}f(x_0, y_0), \\ y_{2\times h}^{(3)} = y_h &= y_0 + h\left[b_1 f\left(x_0, y_{2\times h}^{(1)}\right) + b_2 f\left(x_0 + c_2h, y_{2\times h}^{(2)}\right)\right], \\ y_{2\times h}^{(4)} &= y_h + ha_{21}f(x_0 + h, y_h), \end{aligned}$$

$$y_{2\times h} = y_h + h\left[b_1 f(x_0+h, y_h) + b_2 f\left(x_0 + h + c_2 h, y_{2\times h}^{(4)}\right)\right].$$

Inserting the formula for y_h in the last two lines, one sees that the Butcher tableau of this method is

$$\begin{array}{c|cccc} 0 & & & & \\ c_2 & a_{21} & & & \\ 1 & b_1 & b_2 & & \\ 1+c_2 & b_1 & b_2 & a_{21} & \\ \hline & b_1 & b_2 & b_1 & b_2 \end{array}.$$

That means, the computation of $y_{2\times h}$ is equivalent to the computation of an approximation with the help of an explicit 4-stage Runge–Kutta scheme.

Altogether, five function evaluations are needed:

$$\begin{aligned} &f(x_0, y_0),\ f\left(x_0 + 2c_2 h, y_{2h}^{(2)}\right), f\left(x_0 + c_2 h, y_{2\times h}^{(2)}\right), \\ &f(x_0+h, y_h),\quad f\left(x_0 + h + c_2 h, y_{2\times h}^{(4)}\right). \end{aligned}$$

In the case of a s-stage Runge–Kutta method, $(3s-1)$ function evaluations are required. This number is rather large and the high costs per time step are a disadvantage of the Richardson method. □

Remark 1.37. Comparison of both approximations. Consider a one-step method

$$y_{k+1} = y_k + h\Phi(x, y, h)$$

of order p. Let the initial value $y(x_0)$ be exact, then it follows for the local error in $x_0 + 2h$ that

$$y(x_0+2h) - y_{2h} = C(x_0)(2h)^{p+1} + \mathcal{O}\left(h^{p+2}\right). \tag{1.16}$$

For estimating the local error of $y_{2\times h}$ it will be assumed that the incremental function $\Phi(x, y, h)$ is Lipschitz continuous in the second argument. This assumption is always satisfied for explicit Runge–Kutta schemes if $f(x, y)$ possesses this property, see the proof of Theorem 1.31. It is

$$y_{2\times h} = y_h + h\Phi\left(x + h, y_h, h\right). \tag{1.17}$$

Let

$$\hat{y}_{2\times h} = y(x_0+h) + h\Phi\left(x + h, y(x_0+h), h\right) \tag{1.18}$$

be the iterate that is computed with the exact starting value in $x_0 + h$. Using the definition of the consistency order, one obtains with (1.17) and (1.18)

$$\begin{aligned} &y(x_0+2h) - y_{2\times h} \\ &= \left(y(x_0+2h) - \hat{y}_{2\times h}\right) + \left(\hat{y}_{2\times h} - y_{2\times h}\right) \end{aligned}$$

$$= \left[C(x_0+h)h^{p+1} + \mathcal{O}\left(h^{p+2}\right)\right] + \Big[y(x_0+h) + h\Phi\left(x+h, y(x_0+h), h\right) - y_h - h\Phi\left(x+h, y_h, h\right)\Big].$$

For the terms with the incremental function, one gets from the Lipschitz continuity and the consistency order

$$|h\Phi\left(x+h, y(x_0+h), h\right) - h\Phi\left(x+h, y_h, h\right)| \leq hL \underbrace{|y(x_0+h) - y_h|}_{\mathcal{O}\left(h^{p+1}\right)} = \mathcal{O}\left(h^{p+2}\right).$$

It follows, applying again the consistency error, that

$$\begin{aligned} & y(x_0+2h) - y_{2\times h} \\ & = C(x_0+h)h^{p+1} + y(x_0+h) - y_h + \mathcal{O}\left(h^{p+2}\right) \\ & = C(x_0+h)h^{p+1} + C(x_0)h^{p+1} + \mathcal{O}\left(h^{p+2}\right) + \mathcal{O}\left(h^{p+2}\right) \\ & = 2C(x_0)h^{p+1} + \mathcal{O}\left(h^{p+2}\right), \end{aligned} \tag{1.19}$$

where one assumes that $C(x_0+h) = C(x_0) + \mathcal{O}(h)$, i.e., that the constants do not change too rapidly.

Neglecting in (1.16) and (1.19) the higher order terms allows to eliminate $y(x_0+2h)$ and solve for the constant, yielding

$$C(x_0) = \frac{1}{2}\left(\frac{y_{2\times h} - y_{2h}}{2^p - 1}\right)\frac{1}{h^{p+1}}. \tag{1.20}$$

From (1.19), it follows for the local error of the more accurate method that

$$y(x_0+2h) - y_{2\times h} = \frac{y_{2\times h} - y_{2h}}{2^p - 1} + \mathcal{O}\left(h^{p+2}\right). \tag{1.21}$$

The first term on the right-hand side is a computable approximation of this local error. □

Remark 1.38. Increasing the accuracy. Rearranging terms in (1.21) gives

$$y(x_0+2h) - \left(y_{2\times h} + \frac{y_{2\times h} - y_{2h}}{2^p - 1}\right) = \mathcal{O}\left(h^{p+2}\right).$$

Then,

$$\overline{y}_{2\times h} = y_{2\times h} + \frac{y_{2\times h} - y_{2h}}{2^p - 1}$$

is an approximation of the solution of order $p+1$. □

Remark 1.39. Automatic step length control. From (1.21) and (1.20), it follows that

$$\mathrm{err} = \frac{|y_{2\times h} - y_{2h}|}{2^p - 1} \approx 2C(x_0)h^{p+1} \tag{1.22}$$

is a computable approximation of the local error. This approximation will be compared with a prescribed tolerance. Often, a so-called scaled tolerance sc is used, (Hairer *et al.*, 1993, p. 167) or (Strehmel *et al.*, 2012, p. 61). The scaled tolerance is a combination of an absolute tolerance atol and a relative tolerance rtol

$$\mathrm{sc} = \mathrm{atol} + \max\left\{|y_0|, |y_{2\times h}|\right\} \mathrm{rtol}.$$

Then, the scaled error

$$\mathrm{err}_{\mathrm{sc}} = \frac{|y_{2\times h} - y_{2h}|}{(2^p - 1)\mathrm{sc}}$$

is defined.

- If $\mathrm{err}_{\mathrm{sc}} \leq 1 \iff \mathrm{err} \leq \mathrm{sc}$, then the performed step will be accepted. Starting from $y_{2\times h}$ or $\overline{y}_{2\times h}$, the next step will be performed. If $\overline{y}_{2\times h}$ is used, this approach is also called local Richardson extrapolation.
  An important aspect is the choice of the step length h_{new} for the next step. The guideline is that the scaled error for the next step should be on the one hand still smaller than 1 but on the other hand as close to 1 as possible. Following (1.22), it should hold

$$\begin{aligned} 1 = \frac{\mathrm{err}_{\mathrm{new}}}{\mathrm{sc}} &= \frac{2C\left(x_0 + 2h\right)h_{\mathrm{new}}^{p+1}}{\mathrm{sc}} \approx \frac{2C(x_0)h_{\mathrm{new}}^{p+1}}{\mathrm{sc}} \\ &= \frac{2C(x_0)h^{p+1}}{\mathrm{sc}}\left(\frac{h_{\mathrm{new}}}{h}\right)^{p+1} \approx \mathrm{err}_{\mathrm{sc}}\left(\frac{h_{\mathrm{new}}}{h}\right)^{p+1}, \end{aligned}$$

  i.e., h_{new} has to be chosen such that

$$h_{\mathrm{new}} \approx \left(\frac{1}{\mathrm{err}_{\mathrm{sc}}}\right)^{1/(p+1)} h. \tag{1.23}$$

- If $\mathrm{err}_{\mathrm{sc}} > 1$, then the performed step will be rejected. The Richardson method is repeated from (x_0, y_0) with a step length $h_{\mathrm{new}} < h$.
  That means, the work that was spent for performing the step with step length h was wasted. One likes to avoid this situation.

□

Remark 1.40. Issues of the practical implementation. In practical simulations, one uses some modifications of (1.23) for stabilizing the algorithm.

- A safety factor $\alpha \in (0, 1)$ is introduced

$$h_{\mathrm{new}} = \alpha\left(\frac{1}{\mathrm{err}_{\mathrm{sc}}}\right)^{1/(p+1)} h,$$

often $\alpha \in [0.8, 0.9]$.

- One likes to avoid large oscillations of the sizes of subsequent steps. For this reason, a factor for the maximal increase $\alpha_{\max}$ of the new step size with respect to the current step size and a factor for the maximal decrease $\alpha_{\min} < \alpha_{\max}$ are used. Then, one obtains

$$h_{\text{new}} = h \min \left\{ \alpha_{\max}, \max \left\{ \alpha_{\min}, \alpha \left(\frac{1}{\text{err}_{\text{sc}}} \right)^{1/(p+1)} \right\} \right\}.$$

  If a very large step length is proposed

$$\alpha \left(\frac{1}{\text{err}_{\text{sc}}} \right)^{1/(p+1)} > \alpha_{\max},$$

  then the factor $\alpha_{\max}$ becomes effective and similarly for the case that a very small step length is proposed.
- Usually, one prescribes a minimal step length $h_{\min}$ and a maximal step length $h_{\max}$ and requires for all step lengths that $h_k \in [h_{\min}, h_{\max}]$.
- In the first step, one has to estimate h. Generally, this estimate has to be corrected. In practice, this correction is done very fast by algorithms for automatic step length control. An algorithm for determining a good initial step length can be found in (Hairer *et al.*, 1993, p. 168).

□

1.3.2 Embedded Runge–Kutta Schemes

Remark 1.41. Motivation, embedded Runge–Kutta schemes. Richardson extrapolation is quite expensive in terms of evaluations of the incremental function. It is possible to construct a step length control that needs less evaluations, with so-called embedded Runge–Kutta schemes.

The idea of embedded Runge–Kutta schemes consists in computing numerical approximations of the solution at the next time with two one-step methods with different order. The methods are chosen in such a way that it is possible to use certain evaluations of the incremental function for both of them. That means, one has to construct a Runge–Kutta scheme of the form

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ \vdots & & \ddots & \\ c_s & a_{s1} & \cdots & a_{s,s-1} \\ \hline & b_1 & \cdots & b_{s-1} \quad b_s \\ & \tilde{b}_1 & \cdots & \tilde{b}_{s-1} \quad \tilde{b}_s \end{array},$$

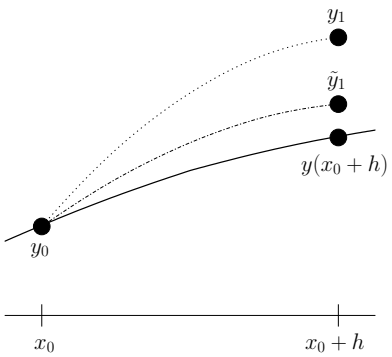

Fig. 1.4 Sketch of embedded Runge–Kutta schemes.

such that

$$y_1 = y_0 + h \sum_{i=1}^{s} b_i K_i(x, y)$$

is order of p and

$$\tilde{y}_1 = y_0 + h \sum_{i=1}^{s} \tilde{b}_i K_i(x, y)$$

is of order q, see Figure 1.4. In general, it is $q = p - 1$ or $q = p + 1$. □

Example 1.42. Runge–Kutta–Fehlberg 2(3). Consider explicit Runge–Kutta schemes with 3 stages

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & a_{21} & & \\ c_3 & a_{31} & a_{32} & \\ \hline & b_1 & b_2 & b_3 \qquad p=2 \\ & \tilde{b}_1 & \tilde{b}_2 & \tilde{b}_3 \qquad q=3 \end{array}.$$

One of the schemes should be of order 2 and the other one of third order. There are 11 parameters to choose. From Theorem 1.26, Theorem 1.27, and Remark 1.28, it follows that 8 equations have to be satisfied

$$\begin{aligned} c_2 &= a_{21}, \\ c_3 &= a_{31} + a_{32}, \\ b_1 + b_2 + b_3 &= 1, \\ b_2 c_2 + b_3 c_3 &= \frac{1}{2}, \\ \tilde{b}_1 + \tilde{b}_2 + \tilde{b}_3 &= 1, \\ \tilde{b}_2 c_2 + \tilde{b}_3 c_3 &= \frac{1}{2}, \end{aligned}$$

$$\tilde{b}_2 c_2^2 + \tilde{b}_3 c_3^2 = \frac{1}{3},$$
$$\tilde{b}_3 a_{32} c_2 = \frac{1}{6}.$$

That means, one has to set three parameters. First, one can choose $c_2 = 1$, $b_3 = 0$. Then, it follows from the first equation that $a_{21} = 1$, from the fourth equation that $b_2 = 1/2$, and from the third equation that $b_1 = 1/2$. Now, one chooses $c_3 = 1/2$. From the sixth and seventh equation, it follows that $\tilde{b}_2 = 1/6$ and $\tilde{b}_3 = 4/6$. Then, one gets from the fifth equation $\tilde{b}_1 = 1/6$ and from the eighth equation $a_{32} = 1/4$. Finally, the second equation gives $a_{31} = 1/4$. The resulting methods have the form

$$\begin{array}{c|ccc|l} 0 & & & & \\ 1 & 1 & & & \\ 1/2 & 1/4 & 1/4 & & \\ \hline & 1/2 & 1/2 & 0 & p=2 \\ & 1/6 & 1/6 & 4/6 & q=3 \end{array}.$$

The method with order $q = 3$ is the Simpson rule. The complete embedded approach is called Runge–Kutta–Fehlberg[7] 2(3) method (RKF 2(3)). □

Remark 1.43. Error estimate, theoretical drawback. By construction, it holds for the embedded scheme that

$$y_1 = y(x_0 + h) + \mathcal{O}\left(h^{p+1}\right), \quad \tilde{y}_1 = y(x_0 + h) + \mathcal{O}\left(h^{q+1}\right).$$

It follows that

$$\begin{aligned} |\text{err}| := |\tilde{y}_1 - y_1| &= \left| y(x_0 + h) + \mathcal{O}\left(h^{q+1}\right) - y(x_0 + h) + \mathcal{O}\left(h^{p+1}\right) \right| \\ &= \left| \mathcal{O}\left(h^{p+1}\right) + \mathcal{O}\left(h^{q+1}\right) \right| \end{aligned} \tag{1.24}$$

is an estimate of the main error term of the Runge–Kutta scheme of order $q^* = \min\{p, q\}$. That means, one obtains only an estimate of the error of the lower order method. To obtain information only on the lower order method is the main theoretical drawback of this approach, since one is interested actually in the higher order method and one will continue the computation also from the higher order approximation. □

Remark 1.44. Automatic step length control, I Controller. Let h be the step size that was used for computing y_1 of order p and $\tilde{y}_1$ of order q with $p < q$. From (1.24), one has

$$|\text{err}| = |y_1 - \tilde{y}_1| = Ch^{p+1}. \tag{1.25}$$

[7] Erwin Fehlberg (1911 – 1972)

Given a tolerance tol for the maximal local error.

- One approach consists in controlling the error per step (EPS). Then, one requires that
$$r_1 = |\mathrm{err}| \leq \mathrm{tol}. \tag{1.26}$$
If this condition is satisfied, then the current step is accepted. Next, one requires for the new step size that the local error is equal to the tolerance
$$Ch_{\mathrm{new}}^{p+1} = \mathrm{tol},$$
with C from (1.25) This requirement gives
$$h_{\mathrm{new}} = \left(\frac{\mathrm{tol}}{C}\right)^{1/(p+1)} = \left(\frac{\mathrm{tol}}{Ch^{p+1}}\right)^{1/(p+1)} h.$$
With (1.25) and (1.26), the new step length is computed by
$$h_{\mathrm{new}} = \left(\frac{\alpha\ \mathrm{tol}}{|\mathrm{err}|}\right)^{1/k} h = \left(\frac{\alpha\ \mathrm{tol}}{r_1}\right)^{1/k} h, \tag{1.27}$$
where $k = p + 1$ and $\alpha \in (0, 1)$ is again a safety factor.
- Another way is the consideration of the error relative to the current step length, the so-called error per unit step (EPUS),
$$r_1 = \frac{|\mathrm{err}|}{h} \leq \mathrm{tol}. \tag{1.28}$$
The satisfaction of a condition of form (1.28) leads to a new step of form (1.27) with $k = p$.
- If (1.26) or (1.28) is not satisfied, then the step is rejected and it will be repeated with a step length smaller than h.
- A generalization of this approach is the so-called I Controller. Replacing in (1.27) $1/k$ by k_I gives
$$h_{\mathrm{new}} = \left(\frac{\alpha\ \mathrm{tol}}{r_1}\right)^{k_I} h.$$
For obtaining a useful automatic step length control mechanism, the choice $k_I = 1/k$ or equivalently $kk_I = 1$ is not necessary. The following choices can be found in the literature

$$\begin{array}{rll} kk_I \in [0,2] \Longleftrightarrow k_I \in [0, 2/k] & & \text{stable control,} \\ kk_I \in (1,2) \Longleftrightarrow k_I \in (1/k, 2/k) & & \text{fast and oscillating control,} \\ kk_I \in (0,1) \Longleftrightarrow k_I \in (0, 1/k) & & \text{slow and smooth control,} \\ kk_I = 1 \Longleftrightarrow k_I = 1/k & & \text{standard I Controller.} \end{array}$$

There are more sophisticated controllers that are used in practical simulations, see Söderlind (2002) for an overview.

□

Remark 1.45. Methods used in practice. In practice, one uses, e.g.,

- RKF 4(5), $s = 6$, Fehlberg (1964),
- RKF 7(8), $s = 13$, Fehlberg (1969),
- DOPRI 4(5) (or DOPRI 5(4) or DOPRI5), $s = 6$, Dormand[8], Prince[9]: Dormand & Prince (1980),
- DOPRI 7(8), $s = 13$, Prince & Dormand (1981).

The standard routine `ode45` from MATLAB uses DOPRI 4(5). □

Remark 1.46. Fehlberg trick. The Fehlberg trick requires that

$$K_s = f\left(x_k + c_s h, y_k + h\sum_{i=1}^{s-1} a_{si}K_i\right) \stackrel{!}{=} f\left(x_k + h, \underbrace{y_k + h\sum_{i=1}^{s} b_i K_i}_{y_{k+1}}\right),$$

i.e., the last evaluation of the incremental function of the old step can be used as first value of the incremental function in the new step. The conditions for applying this trick are

$$a_{si} = b_i,\ i = 1, \ldots, s-1, \quad b_s = 0, \quad c_s = 1.$$

It can be applied, e.g., in DOPRI 4(5). This trick works only if $h_{\text{old}} \approx h_{\text{new}}$. □

[8] John R. Dormand

[9] P. J. Prince

Chapter 2
Numerical Methods for Stiff Ordinary Differential Equations

2.1 Stiff Ordinary Differential Equations

Remark 2.1. Stiffness. It was observed in Curtiss & Hirschfelder (1952) that explicit methods failed for the numerical solution of initial value problems for ordinary differential equations that model certain chemical reactions. They introduced the notation stiffness for such chemical reactions where the fast reacting components arrive in a very short time in their equilibrium and the slowly changing components are more or less fixed, i.e., stiff. In 1963, Dahlquist found out that the reason for the failure of explicit Runge–Kutta methods is their bad stability, see Section 2.3. It should be emphasized that the stability properties of the equations themselves are good, it is in fact a problem of the explicit methods.

There is no unique definition of stiffness in the literature. However, essential properties of stiff systems are as follows:

- There exist, for certain initial conditions, solutions that change slowly.
- Solutions in a neighborhood of these smooth solutions converge quickly to them.

A definition of stiffness can be found in (Strehmel & Weiner, 1995, p. 202), (Strehmel *et al.*, 2012, p. 208). This definition involves a certain norm that depends on the equation and it might be complicated to evaluate this norm. If the solution of (1.1) is sought in the interval $[x_0, x_e]$ and if the right-hand side of (1.1) is Lipschitz continuous in the second argument with Lipschitz constant L, then an approximation of this definition is as follows. A system of ordinary differential equations is called stiff if

$$L\,(x_e - x_0) \gg 1. \tag{2.1}$$

Another definition of stiffness will be given in Definition 2.28. □

Example 2.2. Stiff system of ordinary differential equations. Consider the system

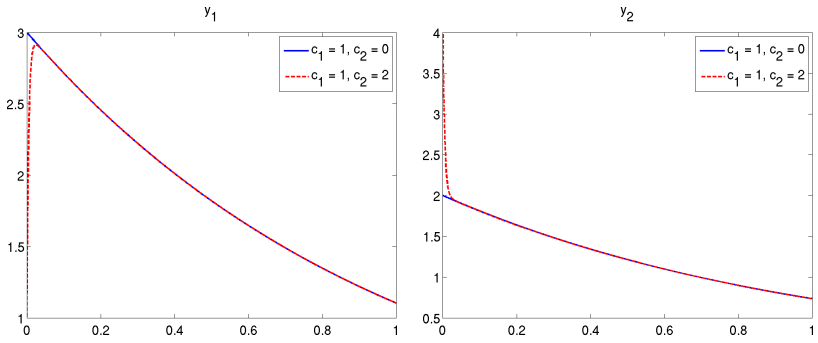

Fig. 2.1 Solutions of Example 2.2, left: first component, right: second component.

$$y_1' = -80.6y_1 + 119.4y_2,$$
$$y_2' = 79.6y_1 - 120.4y_2,$$

in $(0, 1)$. This system is a linear system of ordinary differential equations that can be written in the form

$$\boldsymbol{y}' = \begin{pmatrix} -80.6 & 119.4 \\ 79.6 & -120.4 \end{pmatrix} \boldsymbol{y}.$$

Taking as Lipschitz constant, e.g., the l_1 norm of the system matrix (column sums), one gets $L = 239.8$ and condition (2.1) is satisfied. The general solution of this system is, compare Appendix A.2.3,

$$\boldsymbol{y}(x) = c_1 \begin{pmatrix} 3 \\ 2 \end{pmatrix} e^{-x} + c_2 \begin{pmatrix} -1 \\ 1 \end{pmatrix} e^{-200x}.$$

The constants are determined by the initial condition. If the initial condition is such that $c_2 = 0$, then the solution is smooth for all $x > 0$. Otherwise, if $c_2 \neq 0$, then the solutions changes rapidly for small x while approaching the smooth solution, see Figure 2.1 □

2.2 Implicit Runge–Kutta Schemes

Remark 2.3. Motivation. If the upper triangular part of the matrix of a Runge–Kutta method, see Definition 1.22, is not identical to zero, the Runge–Kutta method is called implicit. That means, there are increments that depend not only on previously computed increments but also on not yet computed increments. Thus, one has to solve a nonlinear problem for computing these increments. Consequently, the implementation of implicit Runge–Kutta

methods is much more involved compared with the implementation of explicit Runge–Kutta methods. Generally, performing one step of an implicit method is much more time-consuming than for an explicit method. However, the great advantage of implicit methods is that they can be used for the numerical simulation of stiff systems, see the stability theory in Section 2.3. □

Remark 2.4. Derivation of implicit Runge–Kutta methods. Implicit Runge–Kutta schemes can be derived from the integral representation (1.7) of the initial value problem. One can show that for each implicit Runge–Kutta scheme with the weights b_j and the nodes $x_k + c_j h$ there is a corresponding quadrature rule with the same weights and the same nodes, see the section on Gaussian quadrature in Numerical Mathematics I. □

Example 2.5. Gauss–Legendre quadrature. Consider the interval $[x_k, x_k+h] = [x_k, x_{k+1}]$. Let $c_1, \ldots, c_s$ be the roots of the Legendre polynomial $P_s(t)$ with the arguments

$$t = \frac{2}{h}\,(x - x_k) - 1 \quad \Longrightarrow \quad t \in [-1, 1].$$

There are s mutually distinct real roots in $(-1, 1)$. After having computed $c_1, \ldots, c_s$, one can determine the coefficients a_{ij}, b_j such that one obtains a method of order $2s$, see Example 2.8. □

Remark 2.6. Simplifying order conditions. The order conditions for an implicit Runge–Kutta scheme with s stages are the same as given in Theorems 1.26, 1.27, and Remark 1.28. These conditions lead to a nonlinear system of equations for computing the parameters of the scheme. These computations are generally quite complicated.

A useful tool for solving this problem are the so-called simplifying order conditions, introduced in Butcher (1964):

$$\begin{aligned}
B(p) &: \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, p, \\
C(l) &: \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad i = 1, \ldots, s,\ k = 1, \ldots, l, \\
D(m) &: \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{1}{k} b_j \left(1 - c_j^k\right), \quad j = 1, \ldots, s,\ k = 1, \ldots, m,
\end{aligned} \tag{2.2}$$

with $0^0 = 1$.

One can show that for sufficiently large values l and m, the conditions $C(l)$ and $D(m)$ can be reduced to $B(p)$ with appropriate p. □

Remark 2.7. Interpretation of $B(p)$ and $C(l)$. Consider the initial value problem

$$y'(x) = f(x), \quad y(x_0) = 0.$$

With the fundamental theorem of differential calculus, one sees that this problem has the solution

$$y(x_0 + h) = \int_{x_0}^{x_0+h} f(\xi)\ d\xi = h \int_0^1 f\left(x_0 + h\theta\right)\ d\theta.$$

A Runge–Kutta method with s stages gives

$$y_1 = h \sum_{i=1}^{s} b_i f\left(x_0 + c_i h\right).$$

Consider in particular the case that $f(x)$ is a polynomial $f(x) = (x - x_0)^{k-1}$, $k \in \mathbb{N} \setminus \{0\}$. Then, the analytical solution has the form

$$y(x_0 + h) = h \int_0^1 (h\theta)^{k-1}\, d\theta = \frac{(h\theta)^k}{k}\bigg|_{\theta=0}^{\theta=1} = \frac{h^k}{k}. \tag{2.3}$$

The Runge–Kutta scheme yields

$$y_1 = h \sum_{i=1}^{s} b_i (c_i h)^{k-1} = h^k \sum_{i=1}^{s} b_i c_i^{k-1}. \tag{2.4}$$

Comparing (2.3) and (2.4), one can observe that condition $B(p)$ means that the quadrature rule that is the basis of the Runge–Kutta method is exact for polynomials of degree $(p-1)$.

Condition $C(1)$ is (1.13) with the upper limit s

$$c_i = \sum_{j=1}^{s} a_{ij}, \quad i = 1, \ldots, s. \tag{2.5}$$

□

Example 2.8. Classes of implicit Runge–Kutta schemes.

- *Gauss–Legendre schemes.* The nodes of the Gauss–Legendre quadrature are used. A method with s stages possesses the maximal possible order $2s$, where all nodes are in the interior of the intervals. To get the optimal order, one has to show that $B(2s)$, $C(s)$, $D(s)$ are satisfied, see (Strehmel *et al.*, 2012, Section 8.1.2), i.e.,

$$\begin{aligned} \sum_{i=1}^{s} b_i c_i^{k-1} &= \frac{1}{k}, \quad k = 1, \ldots, 2s, \\ \sum_{j=1}^{s} a_{ij} c_j^{k-1} &= \frac{1}{k} c_i^k, \quad i = 1, \ldots, s,\ k = 1, \ldots, s, \end{aligned} \tag{2.6}$$

$$\sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{1}{k} b_j \left(1 - c_j^k\right), \quad j = 1, \ldots, s,\ k = 1, \ldots, s.$$

An example is the implicit mid point rule, whose coefficients can be derived by setting $s = 1$ in (2.6). One obtains the following conditions

$$b_1 = 1,\ b_1 c_1 = \frac{1}{2},\ a_{11} = c_1,\ b_1 a_{11} = b_1 (1 - c_1).$$

Consequently, the implicit mid point rule is given by

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

- *Gauss–Radau*[1] *methods.* These methods are characterized by the feature that one of the end points of the interval $[x_k, x_{k+1}]$ belongs to the nodes. A method of this class with s stages has at most order $2s - 1$.
  Examples ($s = 1$):
  - $\begin{array}{c|c} 0 & 1 \\ \hline & 1 \end{array}$ $s = 1,\ p = 1,$
  - $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$ $s = 1,\ p = 1,$ implicit Euler scheme.

  The first scheme does not satisfy condition (2.5).
- *Gauss–Lobatto*[2] *methods.* In these methods, both end points of the interval $[x_k, x_{k+1}]$ are nodes. A method of this kind with s stages cannot be of higher order than $(2s - 2)$.
  Examples:
  - trapezoidal rule, Crank[3]–Nicolson[4] scheme

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

  - other scheme

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

  The second scheme does not satisfy condition (2.5).

□

[1] Rodolphe Radau (1835 – 1911)
[2] Rehuel Lobatto (1797 – 1866)
[3] John Crank (1916 – 2006)
[4] Phyllis Nicolson (1917 – 1968)

Remark 2.9. Diagonally implicit Runge–Kutta methods (DIRK methods). For an implicit Runge–Kutta method with s stages, one has to solve a coupled nonlinear system for the increments $K_1(x,y),\dots,K_s(x,y)$. This step is expensive for a large number of stages s. A compromise is the use of so-called diagonally implicit Runge–Kutta (DIRK) methods

$$\begin{array}{c|ccccc} c_1 & a_{11} & 0 & 0 & \cdots & 0 \\ c_2 & a_{21} & a_{22} & 0 & \cdots & 0 \\ c_3 & a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & a_{s2} & \cdots & & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array}.$$

In DIRK methods, one has to solve s independent nonlinear equations for the increments. In the equation for $K_i(x,y)$, only the stages $K_1(x,y),\dots,K_i(x,y)$ appear, where $K_1(x,y),\dots,K_{i-1}(x,y)$ were already computed. □

2.3 Stability Theory

Remark 2.10. On the stability theory. The stability theory studies numerical methods for solving the linear initial value problem

$$y'(x) = \lambda y(x), \quad y(0) = 1,\ \lambda \in \mathbb{C}. \tag{2.7}$$

It will turn out the even at the simple initial value problem (2.7) the most important stability properties of numerical methods can be explored. The solution of (2.7) is

$$y(x) = e^{\lambda x}.$$

If the initial condition will be slightly perturbed to be $1+\delta_0$, then the solution of the perturbed initial value problem is

$$\tilde{y}(x) = (1+\delta_0)e^{\lambda x} = e^{\lambda x} + \delta_0 e^{\lambda x}.$$

If $\lambda = a + ib$ with $a = \mathrm{Re}(\lambda) > 0$, then the difference

$$|y(x) - \tilde{y}(x)| = \left|\delta_0 e^{\lambda x}\right| = |\delta_0|\,|e^{ax}|\left|e^{ibx}\right| = |\delta_0|\,|e^{ax}|$$

becomes for each $\delta_0 \neq 0$ arbitrarily large if x is sufficiently large. That means, the initial value problem (2.7) is not stable in this case. In this situation, one cannot expect that any numerical method is stable. Hence, this situation is not of interest for numerical simulations.

In contrast, if $\mathrm{Re}(\lambda) < 0$, then the difference $|y(x) - \tilde{y}(x)|$ becomes arbitrarily small and the initial value problem is stable, i.e., small changes of the

data result only in small changes of the solution. This case is of interest for the stability theory of methods for solving ordinary differential equations.

This section considers one-step methods with equidistant meshes with step size h. The solution of (2.7) in the node $x_{k+1} = (k+1)h$ is

$$y(x_{k+1}) = e^{\lambda x_{k+1}} = e^{\lambda(x_k+h)} = e^{\lambda h} e^{\lambda x_k} = e^{\lambda h} y(x_k) =: e^z y(x_k),$$

with $z := \lambda h \in \mathbb{C}$, $\mathrm{Re}(z) \leq 0$. Now, it will be studied how the step from x_k to x_{k+1} looks like for different one-step methods. In particular, large steps are of interest, i.e., $|z| \to \infty$. □

Example 2.11. Behavior of different one-step methods for one step of the model problem (2.7).

1. *Explicit Euler method.* The general form of this method is

$$y_{k+1} = y_k + hf\left(x_k, y_k\right).$$

   In particular, one obtains for (2.7)

$$y_{k+1} = y_k + h\lambda y_k = (1+z)\, y_k =: R(z) y_k.$$

   It holds, independently of $\mathrm{Re}(z)$, that $\lim_{|z|\to\infty} |R(z)| = \infty$.
2. *Implicit Euler method.* This method has the form

$$y_{k+1} = y_k + hf\left(x_{k+1}, y_{k+1}\right).$$

   For applying it to (2.7), one can rewrite it as follows

$$\begin{aligned} y_{k+1} &= y_k + h\lambda y_{k+1} && \Longleftrightarrow \\ (1-z) y_{k+1} &= y_k && \Longleftrightarrow \\ y_{k+1} &= \frac{1}{1-z} y_k = \left(1 + \frac{z}{1-z}\right) y_k =: R(z) y_k. \end{aligned}$$

   For this method, one has, independently of $\mathrm{Re}(z)$, that $\lim_{|z|\to\infty} |R(z)| = 0$.
3. *Trapezoidal rule.* The general form of this method is

$$y_{k+1} = y_k + \frac{h}{2}\left(f(x_k, y_k) + f(x_{k+1}, y_{k+1})\right),$$

   which can be derived from the Butcher tableau given in Example 2.8. For the linear differential equation (2.7), one gets

$$
\begin{aligned}
y_{k+1} &= y_k + \frac{h}{2}\left(\lambda y_k + \lambda y_{k+1}\right) && \Longleftrightarrow \\
\left(1 - \frac{z}{2}\right) y_{k+1} &= \left(1 + \frac{z}{2}\right) y_k && \Longleftrightarrow \\
y_{k+1} &= \frac{1 + z/2}{1 - z/2} y_k = \left(1 + \frac{z}{1 - z/2}\right) y_k =: R(z) y_k.
\end{aligned}
$$

Let $z = 2r\left(\cos(\phi) + i\sin(\phi)\right)$. Inserting this expression gives

$$
\begin{aligned}
\lim_{|z|\to\infty} \left|\frac{1 + z/2}{1 - z/2}\right| &= \lim_{r\to\infty} \left|\frac{1 + r\left(\cos(\phi) + i\sin(\phi)\right)}{1 - r\left(\cos(\phi) + i\sin(\phi)\right)}\right| \\
&= \lim_{r\to\infty} \left|\frac{1/r + \left(\cos(\phi) + i\sin(\phi)\right)}{1/r - \left(\cos(\phi) + i\sin(\phi)\right)}\right| \\
&= \frac{\left|\left(\cos(\phi) + i\sin(\phi)\right)\right|}{\left|-\left(\cos(\phi) + i\sin(\phi)\right)\right|} = \frac{1}{1} = 1.
\end{aligned}
$$

Hence, one has that $\lim_{|z|\to\infty} |R(z)| = 1$ for the trapezoidal rule, independently of ϕ, and with that independently of $\mathrm{Re}(z)$.

The function $R(z)$ describes for each method the step from x_k to x_{k+1}. Thus, this function is an approximation of e^z, which has for different methods different properties, e.g., the limit for $|z| \to \infty$. □

Definition 2.12. Stability function. Let $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^s$, $\hat{\mathbb{C}} = \mathbb{C} \cup \infty$, where ∞ has to be understood as in function theory (Riemann sphere), and consider a Runge–Kutta method with s stages and with the parameters $(A, \boldsymbol{b}, \boldsymbol{c})$. Then, the function

$$
R \,:\, \hat{\mathbb{C}} \to \hat{\mathbb{C}}, \quad z \mapsto 1 + z\boldsymbol{b}^T (I - zA)^{-1}\mathbf{1} \tag{2.8}
$$

is called stability function of the Runge–Kutta method. □

Remark 2.13. Stability functions from Example 2.11. All stability functions from Example 2.11 can be written in the form (2.8). One obtains, e.g., for the trapezoidal rule

$$
\boldsymbol{b} = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \quad I - zA = \begin{pmatrix} 1 & 0 \\ -\frac{z}{2} & 1 - \frac{z}{2} \end{pmatrix}, \quad (I - zA)^{-1} = \frac{1}{1 - \frac{z}{2}} \begin{pmatrix} 1 - \frac{z}{2} & 0 \\ \frac{z}{2} & 1 \end{pmatrix},
$$

from what follows that

$$
1 + z\boldsymbol{b}^T (I - zA)^{-1}\mathbf{1} = 1 + \frac{z}{1 - z/2}\left(\frac{1}{2} - \frac{z}{4} + \frac{z}{4} + \frac{1}{2}\right) = 1 + \frac{z}{1 - z/2}.
$$

□

Theorem 2.14. Form of the stability function of Runge–Kutta methods. *Given a Runge–Kutta scheme with s stages and with the parameters*

$(A, \boldsymbol{b}, \boldsymbol{c})$, then the stability function $R(z)$ is a rational function defined on $\hat{\mathbb{C}}$, whose polynomial order in the numerator and in the denominator is at most s. The poles of this functions might be only at values that correspond to the inverse of an eigenvalue of A. For an explicit Runge–Kutta scheme, $R(z)$ is a polynomial.

Proof. Consider first an explicit Runge–Kutta scheme. In this case, the matrix A is a strictly lower triangular matrix. Hence, $I - zA$ is a triangular matrix with the values one at its main diagonal. This matrix is invertible and it is

$$(I - zA)^{-1} = I + zA + \ldots + z^{s-1}A^{s-1}, \tag{2.9}$$

which can be checked easily by multiplication with $(I - zA)$ and using that $A^s = 0$ since A is strictly lower triangular. It follows from (2.8) and (2.9) that $R(z)$ is a polynomial in z of degree at most s.

Now, the general case will be considered. The expression $(I-zA)^{-1}\mathbf{1}$ can be interpreted as the solution of the linear system of equations $(I - zA)\underline{\zeta} = \mathbf{1}$. Using the Cramer rule, one finds that the i-th component of the solution has the form

$$\zeta_i = \frac{\det A_i}{\det(I - zA)},$$

where A_i is the matrix that is obtained by replacing the i-th column of $(I - zA)$ by the right-hand side, i.e., by $\mathbf{1}$. The numerator of ζ_i is a polynomial in z of order at most $(s-1)$ since there is one column where z does not appear. The denominator is a polynomial of degree at most s. Multiplying with $z\boldsymbol{b}^T$ from the left-hand side gives just a rational function with polynomials of at most degree s both in the numerator and in the denominator.

There is only one case where this approach does not work, namely if

$$\det(I - zA) = \det(z(I/z - A)) = z^s \det(I/z - A) = 0,$$

i.e., if $1/z$ is an eigenvalue of A. ■

Theorem 2.15. Solution of the initial value problem (2.7) obtained with a Runge–Kutta scheme. *Consider a Runge–Kutta method with s stages and with the parameters $(A, \boldsymbol{b}, \boldsymbol{c})$. If $z^{-1} = (\lambda h)^{-1}$ is not an eigenvalue of A, then the Runge–Kutta scheme is well-defined for the initial value problem (2.7). In this case, it is*

$$y_k = (R(h\lambda))^k, \quad k = 0, 1, 2, \ldots.$$

Proof. The statement of the theorem follows directly if one writes the Runge–Kutta scheme for (2.7) and applies induction. *exercise* ■

Definition 2.16. Stability domain. The stability domain of a one-step method is the set

$$S := \{z \in \hat{\mathbb{C}} \ : \ |R(z)| \leq 1\}.$$

□

Remark 2.17. Desirable property for the stability domain. The stability domain of the initial value problem (2.7) is, see Remark 2.10,

$$S_{\text{anal}} = \mathbb{C}_0^- := \{z \in \mathbb{C} \ : \ \text{Re}(z) \leq 0\},$$

since $R(z) = e^z$. In this domain, the solution decreases (for $\text{Re}(z) < 0$) or its absolute value is constant (for $\text{Re}(z) = 0$). A desirable property of a numerical method is that it should be stable for all parameters where the initial value problem is stable, i.e., $\mathbb{C}_0^- \subseteq S$. □

Definition 2.18. A-stable method. If for the stability domain S of a one-step method, it holds that $\mathbb{C}_0^- \subseteq S$, then this one-step method is called A-stable. □

Lemma 2.19. Property of an A-stable method. *Consider an A-stable one-step method, then it is* $|R(\infty)| \leq 1$.

Proof. By the assumption $\mathbb{C}_0^- \subseteq S$, the absolute value of the stability function is bounded from above by 1 for all $|z| \to \infty$ with $\text{Re}(z) \leq 0$. From Theorem 2.14, it follows that the stability function has to be a rational function where the polynomial degree of the numerator is not larger than the polynomial degree of the denominator, since otherwise the function is unbounded for $|z| \to \infty$. It is known from function theory that such rational functions are continuous in ∞. Hence, it is $|R(\infty)| \leq 1$. ■

Remark 2.20. On A-stable methods. The behavior of the stability function for $|z| \to \infty$, $z \in \mathbb{C}_0^-$, is of utmost interest, since it describes the length of the steps that is admissible for given λ such that the method is still stable. However, from the property $|R(\infty)| \leq 1$, it does not follow that the step length can be chosen arbitrarily large without loosing the stability of the method. □

Definition 2.21. Strongly A-stable method, L-stable method. An A-stable one-step method is called strongly A-stable, if it satisfies in addition $|R(\infty)| < 1$. It is called L-stable (left stable), if even it holds that $|R(\infty)| = 0$. □

Example 2.22. Stability of some one-step methods. The types of stability defined in Definitions 2.18 and 2.21 are of utmost importance for the quality of a numerical method.

1. *Explicit Euler method.* It is $R(z) = 1 + z$, i.e., the stability domain is the closed circle with radius 1 and center $(-1, 0)$, see Figure 2.2. This method is not A-stable. For $|\lambda|$ large, one has to use very small steps in order to get stable simulations.
   The smallness of the step lengths for stable simulations of stiff problems is the basic problem of all explicit methods.
2. *Implicit Euler method.* One has for this method $R(z) = 1/(1 - z)$. The stability domain is the complete complex plane without the open circle with radius 1 and center $(1, 0)$, see Figure 2.2. Hence, the method is A-stable. From Example 2.11, it is already known that $|R(\infty)| = 0$ such

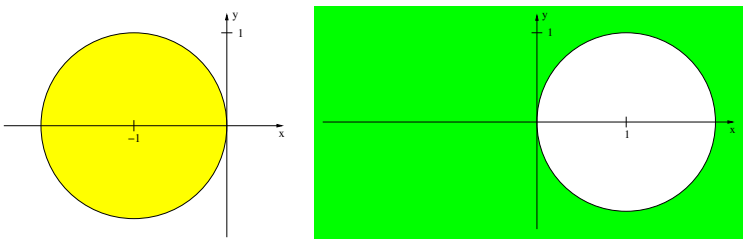

Fig. 2.2 Stability domain of the explicit Euler method (left) and the implicit Euler method (right).

that the implicit Euler method is even L-stable. A smallness condition on the step lengths does not arise for this method, at least for the model problem (2.7).
In general, one can apply with the implicit Euler method much larger steps than, e.g., with the explicit Euler method. Step size restrictions arise, e.g., from the physics of the problem and from the required accuracy of the simulations. However, one has to solve in general in each node a nonlinear equation, like for each implicit scheme. Thus, the numerical costs and the computing time per step are usually much larger than for explicit schemes.

3. *Trapezoidal rule.* For the trapezoidal rule, one gets with $z = a + ib$, $a, b \in \mathbb{R}$,

$$|R(z)|^2 = \left|\frac{1 + z/2}{1 - z/2}\right|^2 = \left|\frac{1 + a/2 + ib/2}{1 - a/2 - ib/2}\right|^2 = \frac{(2+a)^2 + b^2}{(2-a)^2 + b^2}.$$

Thus, $|R(z)| \leq 1$ if $|2 + a| \leq |2 - a|$, compare Figure 2.3, i.e.,

$$R(z) \begin{cases} < 1 \text{ for } a < 0 \iff \operatorname{Re}(z) < 0, \\ = 1 \text{ for } a = 0 \iff \operatorname{Re}(z) = 0, \\ = 1 \text{ for } z = \infty, \end{cases}$$

see also Example 2.11. Hence, one obtains $S = \mathbb{C}_0^-$. This method is A-stable but not L-stable. However, in contrast to the implicit Euler method, which is a first order method, the trapezoidal rule is a second order method.

□

Remark 2.23. Linear systems of ordinary differential equations. Consider the linear system of ordinary differential equations with constant coefficients

$$\boldsymbol{y}'(x) = A\boldsymbol{y}(x), \quad \boldsymbol{y}(0) = \boldsymbol{y}_0, \ A \in \mathbb{R}^{n\times n}, \ \boldsymbol{y}_0 \in \mathbb{R}^n. \tag{2.10}$$

The solution of (2.10) has the form

$$\boldsymbol{y}(x) = e^{Ax}\boldsymbol{y}_0,$$

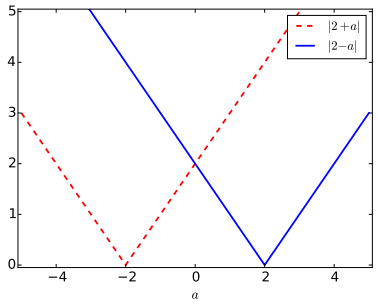

Fig. 2.3 Illustration to the trapezoidal rule, Example 2.22.

where Ax is defined component-wise, as a multiplication of a scalar with a matrix. The first factor on the right-hand side is the matrix exponential. □

Definition 2.24. Matrix exponential. Let $A \in \mathbb{R}^{n\times n}$ and

$$A^0 := I,\ A^1 := A,\ A^2 := AA,\ \ldots,\ A^k := A^{k-1}A.$$

The matrix exponential is defined by

$$e^A := \exp(A)\ :\ \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}, \quad A \mapsto \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

□

Lemma 2.25. Properties of the matrix exponential. *The matrix exponential has the following properties:*

i) The series

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

converges absolutely for all $A \in \mathbb{R}^{n\times n}$, like in the real case $n = 1$.

ii) If the matrices $A, B \in \mathbb{R}^{n\times n}$ are commuting, i.e., if $AB = BA$ holds, then it follows that

$$e^A e^B = e^{A+B}.$$

iii) The matrix $\left(e^A\right)^{-1} \in \mathbb{R}^{n\times n}$ exists for all $A \in \mathbb{R}^{n\times n}$ and it holds

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

This property corresponds to $e^x \neq 0$ for the scalar case.

iv) It holds $\operatorname{rank}\left(e^A\right) = n$, $\det\left(e^A\right) \neq 0$.

v) The matrix-valued function $\mathbb{R} \to \mathbb{R}^{n\times n}$, $x \mapsto e^{Ax}$, *where* Ax *is defined component-wise, is continuously differentiable with respect to* x *with*

$$\frac{d}{dx}e^{Ax} = Ae^{Ax}.$$

The derivative of the exponential is the first factor in this matrix product. The formula looks the same as in the scalar case.

Proof. i) with comparison test with a majorizing series, using that the corresponding series with real argument converges for all real numbers, see literature,
ii) follows from i), exercise,
iii) follows from ii), exercise,
iv) follows from iii),
v) direct calculation with difference quotient, exercise.

■

Example 2.26. Matrix exponential. There are only few classes of matrices that allow an easy computation of the matrix exponential: diagonal matrices and nilpotent matrices.

1. Consider

$$A = \begin{pmatrix} 1&0&0\\ 0&2&0\\ 0&0&3 \end{pmatrix} \quad\Longrightarrow\quad A^k = \begin{pmatrix} 1^k&0&0\\ 0&2^k&0\\ 0&0&3^k \end{pmatrix}.$$

It follows that

$$\begin{aligned} e^{Ax} &= \sum_{k=0}^{\infty}\frac{(Ax)^k}{k!} = \sum_{k=0}^{\infty}\frac{1}{k!}\begin{pmatrix} x^k&0&0\\ 0&(2x)^k&0\\ 0&0&(3x)^k \end{pmatrix}\\ &= \begin{pmatrix} \sum_{k=0}^{\infty}\frac{x^k}{k!}&0&0\\ 0&\sum_{k=0}^{\infty}\frac{(2x)^k}{k!}&0\\ 0&0&\sum_{k=0}^{\infty}\frac{(3x)^k}{k!} \end{pmatrix} = \begin{pmatrix} e^x&0&0\\ 0&e^{2x}&0\\ 0&0&e^{3x} \end{pmatrix}. \end{aligned}$$

2. This example illustrates property ii) of Lemma 2.25. For the matrices

$$A = \begin{pmatrix} 2&0\\ 0&3 \end{pmatrix}, \quad B = \begin{pmatrix} 0&1\\ 0&0 \end{pmatrix},$$

it is possible to calculate the corresponding series easily, since B is a nilpotent matrix ($B^2 = 0$). One obtains

$$e^A = \begin{pmatrix} e^2&0\\ 0&e^3 \end{pmatrix}, \quad e^B = I + B = \begin{pmatrix} 1&1\\ 0&1 \end{pmatrix}.$$

It holds $AB \neq BA$ and

$$e^A e^B = \begin{pmatrix} e^2 & e^2 \\ 0 & e^3 \end{pmatrix} \neq \begin{pmatrix} e^2 & e^3 \\ 0 & e^3 \end{pmatrix} = e^B e^A.$$

Assume that $e^A e^B = e^{A+B}$. Since $e^{A+B} = e^{B+A}$, it follows that then also $e^B e^A = e^{B+A} = e^{A+B} = e^A e^B$, which is a contradiction to the calculations from above.

□

Remark 2.27. Extension of the stability theory to linear systems. Consider system (2.10). It will be assumed that the matrix A possesses n eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$.

Further, it will be assumed that this matrix can be diagonalized, i.e., there exists a matrix $Q \in \mathbb{R}^{n \times n}$ such that

$$\Lambda = Q^{-1} A Q, \quad \text{with } \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

This property is given, e.g., if all eigenvalues are mutually different. The columns $\boldsymbol{q}_i$ of Q are the eigenvectors of A. With the substitution

$$\boldsymbol{y}(x) = Q\boldsymbol{z}(x) \quad \Longrightarrow \quad \boldsymbol{y}'(x) = Q\boldsymbol{z}'(x),$$

one obtains the differential equation

$$Q\boldsymbol{z}'(x) = AQ\boldsymbol{z}(x) \quad \Longleftrightarrow \quad \boldsymbol{z}'(x) = Q^{-1}AQ\boldsymbol{z}(x) = \Lambda \boldsymbol{z}(x).$$

The equations of this system are decoupled. Its general solution is given by

$$\boldsymbol{z}(x) = e^{\Lambda x}\boldsymbol{c} = \left(c_i e^{\lambda_i x}\right)_{i=1,\ldots,n}.$$

It follows that the general solution of (2.10) has the form

$$\boldsymbol{y}(x) = Q\boldsymbol{z}(x) = \sum_{i=1}^{n} c_i e^{\lambda_i x} \boldsymbol{q}_i.$$

Inserting this expression in the initial condition yields

$$\boldsymbol{y}(0) = \sum_{i=1}^{n} c_i \boldsymbol{q}_i = Q\boldsymbol{c} = \boldsymbol{y}_0 \quad \Longrightarrow \quad \boldsymbol{c} = Q^{-1}\boldsymbol{y}_0.$$

Hence, one obtains the following solution of the initial value problem

$$\boldsymbol{y}(x) = \sum_{i=1}^{n} \left(Q^{-1}\boldsymbol{y}_0\right)_i e^{\lambda_i x} \boldsymbol{q}_i, \tag{2.11}$$

where $\left(Q^{-1}\boldsymbol{y}_0\right)_i$ is the i-th component of $Q^{-1}\boldsymbol{y}_0$. Now, one can easily see that the solution is stable (small changes of the initial data lead to small changes of the solution) only if all eigenvalues have a negative real part.

The study of numerical methods makes sense only in the case that the problem is well posed, i.e., all eigenvalues have a negative real part. Then, the most important term in (2.11) with respect to stability is the term with the eigenvalue of A with the largest absolute value of its real part, since for the stability, the absolute values of the product of the real parts of the eigenvalues and the step length are important. □

Definition 2.28. Stiff system of ordinary differential equations. The linear system of ordinary differential equations

$$\boldsymbol{y}'(x) = \mathcal{A}\boldsymbol{y}(x), \quad \mathcal{A} \in \mathbb{R}^{n\times n},$$

is called stiff, if all eigenvalues λ_i of $\mathcal{A}$ possess a negative real part and if

$$q := \frac{\max\{|\mathrm{Re}(\lambda_i)|\,,\ i = 1,\ldots,n\}}{\min\{|\mathrm{Re}(\lambda_i)|\,,\ i = 1,\ldots,n\}} \gg 1.$$

Sometimes, the system is called weakly stiff if $q \approx 10$ and stiff if $q > 10$. □

Remark 2.29. On Definition 2.28. Definition 2.28 has a disadvantage. The ratio becomes large also in the case that the eigenvalue with the smallest absolute value of the real part is close to zero. However, this eigenvalue is not important for the stability of numerical methods, only the eigenvalue with the largest absolute value of the real part. □

Remark 2.30. Local stiffness for general ordinary differential equations. The concept of stiffness can be extended in some sense from linear differential equations to general differential equations. The differential equation

$$\boldsymbol{y}'(x) = \boldsymbol{f}(x, \boldsymbol{y}(x))$$

can be transformed, by introducing the functions

$$\overline{y}(x) := x \quad \text{and} \quad \tilde{\boldsymbol{y}}(x) := \begin{pmatrix} \boldsymbol{y}(x) \\ \overline{y}(x) \end{pmatrix},$$

to the autonomous form

$$\tilde{\boldsymbol{y}}'(x) = \tilde{\boldsymbol{f}}\left(\tilde{\boldsymbol{y}}(x)\right) = \begin{pmatrix} \boldsymbol{f}(x, \boldsymbol{y}(x)) \\ 1 \end{pmatrix}.$$

By linearizing at the initial value $\tilde{\boldsymbol{y}}_0$, one obtains a differential equation of the form $\tilde{\boldsymbol{y}}'(x) = \mathcal{A}\tilde{\boldsymbol{y}}(x)$. Applying some definition of stiffness to the linearized equation, it is possible to define a local stiffness for the general equation.

However, if one considers nonlinear problems, one has to be careful in the interpretation of the results. In general, the results are valid only locally, i.e., in a neighborhood of the point of linearization, and they do not describe the stability of a numerical method in the whole domain of definition of the nonlinear problem. □

2.4 Rosenbrock Methods

Remark 2.31. Goal. From the stability theory, it became obvious that one has to use implicit methods for stiff problems. However, implicit methods are computationally expensive, one has to solve in general nonlinear problems in each step. The goal consists in constructing implicit methods that have on the one hand a reduced computational complexity but on the other hand, they should be still accurate and stable. □

Remark 2.32. Linearly implicit Runge–Kutta methods. Consider, without loss of generality, the autonomous initial value problem in $\mathbb{R}^n$

$$\boldsymbol{y}'(x) = \boldsymbol{f}(\boldsymbol{y}), \quad \boldsymbol{y}(0) = \boldsymbol{y}_0,$$

compare Remark 1.30. DIRK methods, see Remark 2.9, have a Butcher tableau of the form

$$\begin{array}{c|ccccc} c_1 & a_{11} & 0 & 0 & \cdots & 0 \\ c_2 & a_{21} & a_{22} & 0 & \cdots & 0 \\ c_3 & a_{31} & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & & \\ c_s & a_{s1} & a_{s2} & \cdots & & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_{s-1} & b_s \end{array}.$$

One has to solve s decoupled nonlinear equations

$$\boldsymbol{K}_j = \boldsymbol{f}\left(\boldsymbol{y}_k + h\sum_{l=1}^{j-1} a_{jl}\boldsymbol{K}_l + ha_{jj}\boldsymbol{K}_j\right), \quad j = 1, \ldots, s. \tag{2.12}$$

This fixed point equation can be solved with a fixed point iteration. As a special fixed point iteration, the quasi Newton method for solving the j-th equation leads to an iterative scheme of the form

$$\begin{aligned} \boldsymbol{K}_j^{(m+1)} &= \boldsymbol{K}_j^{(m)} \\ &\quad - \left(I - a_{jj}hJ\right)^{-1} \underbrace{\left[\boldsymbol{K}_j^{(m)} - \boldsymbol{f}\left(\boldsymbol{y}_k + h\sum_{l=1}^{j-1} a_{jl}\boldsymbol{K}_l + ha_{jj}\boldsymbol{K}_j^{(m)}\right)\right]}_{\text{residual}}, \end{aligned} \tag{2.13}$$

$m = 0, 1, \ldots$. The derivative with respect to $\boldsymbol{K}_j$ of the corresponding nonlinear problem to (2.12) with right-hand side $\mathbf{0}$ is

$$I - \underbrace{a_{jj}h}_{\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{K}_j}} \partial_{\boldsymbol{y}} \boldsymbol{f} \left(\boldsymbol{y}_k + h \sum_{l=1}^{j-1} a_{jl} \boldsymbol{K}_l + h a_{jj} \boldsymbol{K}_j \right) \in \mathbb{R}^{n \times n}.$$

In (2.13), one uses usually the approximation of the derivative $J = \partial_{\boldsymbol{y}} \boldsymbol{f}(\boldsymbol{y}_k)$ instead of the derivative at the current iterate, hence it is a quasi Newton method. If the step length h is sufficiently small, then the matrix $\left(I - a_{jj}hJ\right)$ is non-singular, since then it is sufficiently close to the identity, and the linear systems of equations possess a unique solution.

Often, it turns out to be sufficient for reaching the required accuracy to perform just one step of the iteration. This statement holds in particular if the step length is sufficiently small and if a sufficiently accurate start value $\boldsymbol{K}_j^{(0)}$ is available. One utilizes the ansatz (linear combination of the already computed increments)

$$\boldsymbol{K}_j^{(0)} := \sum_{l=1}^{j-1} \frac{d_{jl}}{a_{jj}} \boldsymbol{K}_l,$$

where the coefficients d_{jl}, $l = 1, \ldots, j-1$, still need to be determined. Applying just one step in (2.13) with this ansatz, one obtains an implicit method with linear systems of equations of the form

$$\left(I - a_{jj}hJ\right) \boldsymbol{K}_j = \boldsymbol{f} \left(\boldsymbol{y}_k + h \sum_{l=1}^{j-1} \left(a_{jl} + d_{jl}\right) \boldsymbol{K}_l \right) - hJ \sum_{l=1}^{j-1} d_{jl} \boldsymbol{K}_l, \ j = 1, \ldots, s,$$

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + h \sum_{j=1}^{s} b_j \boldsymbol{K}_j. \tag{2.14}$$

This class of methods is called linearly implicit Runge–Kutta methods.

Linearly implicit Runge–Kutta methods are still implicit methods. One has to solve in each step only s linear systems of equations. That means, these methods are considerably less computationally complex than the original implicit methods and the first goal stated in Remark 2.31 is achieved. Now, one has to study which properties of the original methods are transferred to the linearly implicit methods. In particular, stability is of importance. If stability will be lost, then the linearly implicit methods are not suited for solving stiff differential equations. □

Theorem 2.33. Stability of linearly implicit Runge–Kutta methods. *Consider a Runge–Kutta method with the parameters $(A, \boldsymbol{b}, \boldsymbol{c})$, where $A \in \mathbb{R}^{s \times s}$ is a non-singular lower triangular matrix (which was used for the derivation of* (2.14)*). Then, the corresponding linearly implicit Runge–Kutta*

method (2.14) *with* $J = \partial_{\boldsymbol{y}} \boldsymbol{f}(\boldsymbol{y}_k)$ *has the same stability function* $R(z)$ *as the original method, independently of the choice of* $\{d_{jl}\}$.

Proof. The linearly implicit method will be applied to the one-dimensional (to simplify notations) test problem

$$y'(x) = \lambda y(x), \quad y(0) = 1,$$

with $\mathrm{Re}(\lambda) < 0$. Since $f(y) = \lambda y$, one obtains $J = \lambda$. The j-th equation of (2.14) has the form

$$\begin{aligned}\left(1 - a_{jj} h\lambda\right) K_j &= \lambda \left(y_k + h \sum_{l=1}^{j-1} \left(a_{jl} + d_{jl}\right) K_l \right) - h\lambda \sum_{l=1}^{j-1} d_{jl} K_l \\ &= \lambda y_k + h\lambda \sum_{l=1}^{j-1} a_{jl} K_l, \quad j = 1, \ldots, s.\end{aligned}$$

Multiplication with h gives with $z = \lambda h$

$$K_j h - z \sum_{l=1}^{j} a_{jl} K_l h = z y_k, \quad j = 1, \ldots, s.$$

This equation is equivalent, using matrix-vector notation, to

$$(I - zA)\, \boldsymbol{K} h = z y_k \mathbf{1}, \quad \boldsymbol{K} = (K_1, \ldots, K_s)^T.$$

Let h be chosen in such a way that z^{-1} is not an eigenvalue of A. Then, one obtains by inserting this equation in the second equation of (2.14)

$$y_{k+1} = y_k + h \boldsymbol{b}^T \boldsymbol{K} = y_k + h \boldsymbol{b}^T \left(I - zA\right)^{-1} \mathbf{1} \frac{z}{h} y_k = \left(1 + z \boldsymbol{b}^T \left(I - zA\right)^{-1} \mathbf{1}\right) y_k = R(z) y_k.$$

Now one can see that in the parentheses there is the stability function $R(z)$ of the original Runge–Kutta method, see (2.8). ■

Remark 2.34. On the stability and consistency. Since the most important stability properties of a numerical method for solving initial value problems with ordinary differential equations depend only on the stability function, these properties transfer from the original implicit Runge–Kutta method to the corresponding linearly implicit method.

The choice of the coefficients $\{d_{jl}\}$ will influence the order of the linearly implicit method. For an inappropriate choice of these coefficients, the order of the linearly implicit method might be lower than the order of the original method. □

Example 2.35. Linearly implicit Euler method. The implicit Euler method has the Butcher tableau

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

With (2.14), it follows that the linearly implicit Euler method has the form

$$\left(I - h \partial_{\boldsymbol{y}} \boldsymbol{f}(\boldsymbol{y}_k)\right) \boldsymbol{K}_1 = \boldsymbol{f}\left(\boldsymbol{y}_k\right), \quad \boldsymbol{y}_{k+1} = \boldsymbol{y}_k + h \boldsymbol{K}_1.$$

The linearly implicit Euler method is L-stable, like the implicit Euler method, and one has to solve in each step only one linear system of equations. There are no coefficients $\{d_{jl}\}$ to be chosen in this method. □

Remark 2.36. Rosenbrock[5] *methods.* Another possibility for simplifying the use of linearly implicit methods and decreasing the numerical costs consists in using for all increments the same coefficient $a_{jj} = a$. In this case, all linear systems of equations in (2.14) possess the same system matrix $(I - ahJ)$. Then, one needs only one LU decomposition of this matrix and can solve all systems in (2.14) with this decomposition. This approach is called Rosenbrock methods or Rosenbrock–Wanner[6] methods (ROW methods)

$$\begin{aligned}(I - ahJ)\,\boldsymbol{K}_j &= \boldsymbol{f}\left(\boldsymbol{y}_k + h\sum_{l=1}^{j-1}\left(a_{jl} + d_{jl}\right)\boldsymbol{K}_l\right) - hJ\sum_{l=1}^{j-1} d_{jl}\boldsymbol{K}_l,\ j = 1,\ldots,s,\\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + h\sum_{j=1}^{s} b_j\boldsymbol{K}_j.\end{aligned} \tag{2.15}$$

In practice, it is often even possible to use the same approximation J of the Jacobian for some subsequent steps. This is true in particular, if the solution changes only slowly. In this way, one can save additional computational costs. □

Example 2.37. The method `ode23s`. In MATLAB, one can find for solving stiff initial value problems with ordinary differential equations the Rosenbrock method `ode23s`, see Shampine & Reichelt (1997). This method has the form

$$\begin{aligned}(I - ahJ)\,\boldsymbol{K}_1 &= \boldsymbol{f}(\boldsymbol{y}_k), \quad a = \frac{1}{2+\sqrt{2}} \approx 0.2928932,\\ (I - ahJ)\,\boldsymbol{K}_2 &= \boldsymbol{f}\left(\boldsymbol{y}_k + \frac{1}{2}h\boldsymbol{K}_1\right) - ahJ\boldsymbol{K}_1,\\ \boldsymbol{y}_{k+1} &= \boldsymbol{y}_k + h\boldsymbol{K}_2.\end{aligned} \tag{2.16}$$

From the equation for the second increment, it follows that $d_{21} = a$. Then, one obtains with (2.15) $a_{21} = 1/2 - d_{21} = 1/2 - a$. Using the condition that the nodes are the sums of the rows of the matrix, it follows that the corresponding Butcher tableau looks like

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

□

[5] Howard H. Rosenbrock (1920 – 2010)

[6] Gerhard Wanner, born 1942

Theorem 2.38. Order of `ode23s`. *The Rosenbrock method* `ode23s` *is of second order if* $h \in (0, 1/(2a \|J\|_2))$.

Proof. Let $h \in (0, 1/(2a \|J\|_2))$, where $\|\cdot\|_2$ denotes the spectral norm of J, which is induced by the Euclidean vector norm $\|\cdot\|_2$. It can be shown, see class Computer Mathematics, that the matrix $(I - ahJ)$ is invertible if $\|ahJ\|_2 < 1$. This condition is satisfied for the choice of h from above.

Let $\boldsymbol{K}$ be the solution of

$$(I - ahJ)\boldsymbol{K} = \boldsymbol{f}. \tag{2.17}$$

Then, one obtains with the triangle inequality, with the compatibility of the Euclidean vector norm and the spectral matrix norm, and with the choice of h that

$$\begin{aligned}\|(I - ahJ)\boldsymbol{K}\|_2 &\geq \|\boldsymbol{K}\|_2 - ah\,\|J\boldsymbol{K}\|_2 \geq \|\boldsymbol{K}\|_2 - ah\,\|J\|_2\,\|\boldsymbol{K}\|_2 \\ &\geq \|\boldsymbol{K}\|_2 - \frac{a\,\|J\|_2}{2a\,\|J\|_2}\,\|\boldsymbol{K}\|_2 = \frac{1}{2}\,\|\boldsymbol{K}\|_2\,.\end{aligned}$$

It follows with (2.17) that

$$\frac{1}{2}\,\|\boldsymbol{K}\|_2 \leq \|(I - ahJ)\boldsymbol{K}\|_2 = \|\boldsymbol{f}\|_2 \quad \Longrightarrow \quad \|\boldsymbol{K}\|_2 \leq 2\,\|\boldsymbol{f}\|_2\,. \tag{2.18}$$

Thus, the solution of the linear system of equations is bounded by the right-hand side.

One obtains for the first increment of `ode23s` by recursive insertion, using (2.16),

$$\begin{aligned}\boldsymbol{K}_1 &= \boldsymbol{f}(\boldsymbol{y}_k) + ahJ\boldsymbol{K}_1 = \boldsymbol{f}(\boldsymbol{y}_k) + ahJ\left(\boldsymbol{f}(\boldsymbol{y}_k) + ahJ\boldsymbol{K}_1\right) \\ &= \boldsymbol{f}(\boldsymbol{y}_k) + ahJ\boldsymbol{f}(\boldsymbol{y}_k) + h^2a^2J^2\boldsymbol{K}_1 \\ &= \boldsymbol{f}(\boldsymbol{y}_k) + ahJ\boldsymbol{f}(\boldsymbol{y}_k) + \mathcal{O}\left(h^2\right).\end{aligned} \tag{2.19}$$

The last step is allowed since $\boldsymbol{K}_1$ is bounded by the data of the problem (the right-hand side $\boldsymbol{f}(\boldsymbol{y}_k)$) independently of h, see (2.18) where the constant in the estimate is 2. Using a Taylor series expansion and considering only first order terms explicitly, one obtains in a similar way for the second increment of `ode23s`

$$\begin{aligned}\boldsymbol{K}_2 &= \boldsymbol{f}\left(\boldsymbol{y}_k + \frac{1}{2}h\boldsymbol{K}_1\right) - ahJ\boldsymbol{K}_1 + ahJ\boldsymbol{K}_2 \\ &= \boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{1}{2}h\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)\boldsymbol{K}_1 - ahJ\boldsymbol{K}_1 + ahJ\boldsymbol{K}_2 + \mathcal{O}\left(h^2\right) \\ &\overset{(2.19)}{=} \boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{1}{2}h\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)\boldsymbol{f}\left(\boldsymbol{y}_k\right) - ahJ\boldsymbol{f}(\boldsymbol{y}_k) + ahJ\boldsymbol{K}_2 + \mathcal{O}\left(h^2\right) \\ &\overset{(2.20)}{=} \boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{1}{2}h\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)\boldsymbol{f}\left(\boldsymbol{y}_k\right) - ahJ\boldsymbol{f}(\boldsymbol{y}_k) + ahJ\boldsymbol{f}(\boldsymbol{y}_k) + \mathcal{O}\left(h^2\right) \\ &= \boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{1}{2}h\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)\boldsymbol{f}\left(\boldsymbol{y}_k\right) + \mathcal{O}\left(h^2\right).\end{aligned} \tag{2.20}$$

Inserting these results in (2.16) gives for one step of `ode23s`

$$\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + h\boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{1}{2}h^2\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)\boldsymbol{f}\left(\boldsymbol{y}_k\right) + \mathcal{O}\left(h^3\right). \tag{2.21}$$

The Taylor series expansion of the solution $\boldsymbol{y}(x)$ of the system of differential equations in x_k has the form, using the differential equation and the chain rule,

$$\boldsymbol{y}(x_{k+1}) = \boldsymbol{y}(x_k) + h\boldsymbol{y}'(x_k) + \frac{h^2}{2}\boldsymbol{y}''(x_k) + \mathcal{O}\left(h^3\right)$$

$$
\begin{aligned}
&= \boldsymbol{y}(x_k) + h\boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{h^2}{2}\frac{\partial \boldsymbol{f}\left(\boldsymbol{y}\right)}{\partial x}(x_k) + \mathcal{O}\left(h^3\right) \\
&= \boldsymbol{y}(x_k) + h\boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{h^2}{2}\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)\boldsymbol{y}'(x_k) + \mathcal{O}\left(h^3\right) \\
&= \boldsymbol{y}(x_k) + h\boldsymbol{f}\left(\boldsymbol{y}_k\right) + \frac{h^2}{2}\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)\boldsymbol{f}\left(\boldsymbol{y}_k\right) + \mathcal{O}\left(h^3\right).
\end{aligned}
$$

Starting with the exact value at x_k, then the first three terms of (2.21) correspond to the Taylor series expansion of the solution $\boldsymbol{y}(x)$ of the system of differential equations in x_k. Thus, it follows that the local error is of order $\mathcal{O}\left(h^3\right)$, from what follows that the consistency order of `ode23s` is two, see Definition 1.14. ■

Remark 2.39. To the proof of Theorem 2.38. Note that it is not needed in the proof of Theorem 2.38 that J is the exact derivative $\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)$. The method `ode23s` remains a second order method if J is only an approximation of $\partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)$ and even if J is an arbitrary matrix. However, the transfer of the stability properties from the original method to `ode23s` is only guaranteed for the choice $J = \partial_{\boldsymbol{y}}\boldsymbol{f}\left(\boldsymbol{y}_k\right)$, see Theorem 2.33. □

Theorem 2.40. Stability function of `ode23s`. *Assume that $J = \partial_{\boldsymbol{y}}\boldsymbol{f}(\boldsymbol{y}_k)$, then the stability function of the Rosenbrock method `ode23s` has the form*

$$
R(z) = \frac{1 + (1-2a)z}{(1-az)^2}. \tag{2.22}
$$

Proof. The statement of the theorem follows from applying the method to the usual test equation, *exercise*. ■

Corollary 2.41. Stability of `ode23s`. *If $J = \partial_{\boldsymbol{y}}\boldsymbol{f}(\boldsymbol{y}_k)$, then the Rosenbrock method `ode23s` is L-stable.*

Proof. The statement is obtained by applying the definition of L-stability to the stability function (2.22). ■

Remark 2.42. On the order of `ode23s`. It remains the question whether an appropriate choice of J might even increase the order of the method. However, for the model problem of the stability analysis, a series expansion of the stability function shows that the exponential function is reproduced exactly only to the quadratic term. From this observation, it follows that one does not obtain a third order method even with exact Jacobian. In practice, there is no important reason from the point of view of accuracy to compute a new Jacobian in each step. Often, it is sufficient to update J every now and then. □

Chapter 3
Multi-Step Methods

3.1 Definition

Remark 3.1. Multi-step methods. The characteristic feature of one-step methods is that they need for computing y_{k+1} only the value from the previous approximation y_k of the solution. A straightforward extension consists in constructing methods that use for computing y_{k+1} more than one of the previous approximations $y_k, y_{k-1}, \ldots$. Such methods are called multi-step methods. □

Definition 3.2. q-step method, linear q-step method. A q-step method with $q \geq 1$ is a numerical method for approximately solving

$$y'(x) = f\left(x, y(x)\right), \quad y(x_0) = y_0, \tag{3.1}$$

where y_{k+1} depends on y_{k+1-q} but not on y_i with $i < k + 1 - q$.

A q-step method is called linear, if it has the form

$$y_{k+1} = \sum_{j=0}^{q-1} a_j y_{k-j} + h \sum_{j=0}^{q-1} b_j f\left(x_{k-j}, y_{k-j}\right) + h b_{-1} f\left(x_{k+1}, y_{k+1}\right), \tag{3.2}$$

$k = q - 1, q, \ldots$, with $q \geq 1$, $a_0, \ldots, a_{q-1}, b_{-1}, \ldots, b_{q-1} \in \mathbb{R}$, $a_{q-1} \neq 0$ or $b_{q-1} \neq 0$. For $q = 1$, the method is called a one-step method. If $b_{-1} \neq 0$, then the linear q-step method is an implicit method, otherwise it is an explicit method. □

Remark 3.3. Initial values for a q-step method. A q-step method needs q initial values. However, the initial value problem (3.1) provides only the initial value y_0. The second initial value y_1 can be computed with a one-step method, the next initial value y_2 with a one-step method or with a two-step method and so on. It follows that all initial values y_i, $i > 0$, are already numerical approximations. This aspect has to be taken into account in the error analysis of multi-step methods, see Remark 3.23. □

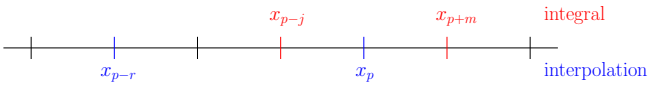

Fig. 3.1 Parameters in the derivation of predictor-corrector schemes.

3.2 Predictor-Corrector Methods

Remark 3.4. Construction. Starting point of the construction of predictor-corrector methods is the equivalent integral form of the initial value problem (3.1)

$$y(x) = y_0 + \int_{x_0}^{x} f\left(t, y(t)\right)\ dt. \tag{3.3}$$

Denote the solution at $\tilde{x}$ by $y(\tilde{x})$, then it holds that

$$y(x) = y(\tilde{x}) + \int_{\tilde{x}}^{x} f\left(t, y(t)\right)\ dt. \tag{3.4}$$

The main idea of predictor-corrector methods consists in approximating the integral on the right-hand side of (3.4) in an appropriate way. There are two principal difficulties:

- The dependency of the term in the integral on t is generally not known since the function $y(t)$ is unknown.
- Even if the dependency of the function in the integral on t is known, generally it will be impossible to find an analytic expression of the solution.

Consider an equidistant grid with nodes

$$x_i = x_0 + ih, \quad i = 0, 1, \ldots.$$

For the derivation of the methods, assume that the term in the integral is known. Then, the derivation is similar to the derivation of the Newton[1]–Cotes[2] formulas for numerical quadrature. In this approach, the term in the integral of (3.4) is replaced by a polynomial interpolant. Let the boundaries of the integral be the nodes

$$\begin{aligned} \tilde{x} &= x_{p-j}, \quad \text{starting point with parameter } j, \\ x &= x_{p+m} \quad \text{end point with parameter } m, \end{aligned} \tag{3.5}$$

with parameters $j, m \in \mathbb{N}_0$ that need yet to be determined. It will be required that the interpolation polynomial $p_r(x)$ satisfies the following properties:

- the degree of $p_r(x)$ is lower than or equal to r,
- $p_r(x_i) = f(x_i, y(x_i))$ for $i = p, p-1, \ldots, p-r$.

[1] Isaac Newton (1642 – 1727)
[2] Roger Cotes (1682 – 1716)

Thus, x_p is the most right-hand side node for computing the interpolation polynomial. The value r is a third parameter, compare Figure 3.1. Note that two sets of nodes are involved in the construction, namely the nodes that determine the boundaries of the integral and the nodes that are used to define the interpolation polynomial. The solution of this interpolation problem is given by the Lagrange[3] interpolation polynomial

$$p_r(x) = \sum_{i=0}^{r} f\left(x_{p-i}, y(x_{p-i})\right) L_i(x)$$

with

$$L_i(x) = \prod_{l=0, l\neq i}^{r} \frac{x - x_{p-l}}{x_{p-i} - x_{p-l}},\ i = 0, 1, \ldots, r. \tag{3.6}$$

It follows by using (3.4), (3.5), and (3.6) that

$$\begin{aligned} y_{p+m} &\approx y_{p-j} + \sum_{i=0}^{r} f\left(x_{p-i}, y(x_{p-i})\right) \int_{\tilde{x}}^{x} L_i(t)\ dt \\ &= y_{p-j} + h \sum_{i=0}^{r} \beta_i f\left(x_{p-i}, y(x_{p-i})\right) \end{aligned} \tag{3.7}$$

with

$$\beta_i = \frac{1}{h} \int_{\tilde{x}}^{x} L_i(t)\ dt = \frac{1}{h} \int_{\tilde{x}}^{x} \left(\prod_{l=0, l\neq i}^{r} \frac{t - x_{p-l}}{x_{p-i} - x_{p-l}} \right) dt.$$

The constructed method is in particular linear. Note that so far the assumption of having an equidistant grid was not used.

Finally, the formula for β_i should be simplified. To this end, note that all fixed values from the interval are nodes of the equidistant grid, such that, e.g., $x_p = x_0 + ph$. Replacing these values and using the substitution

$$t = x_p + sh \quad \Longrightarrow \quad dt = h ds,$$

yields

$$\begin{aligned} \beta_i &= \frac{1}{h} \int_{-j}^{m} \left(\prod_{l=0, l\neq i}^{r} \frac{x_p + sh - x_{p-l}}{x_{p-i} - x_{p-l}} \right) h\ ds \\ &= \int_{-j}^{m} \left(\prod_{l=0, l\neq i}^{r} \frac{x_0 + ph + sh - x_0 - ph + lh}{x_0 + ph - ih - x_0 - ph + lh} \right)\ ds \end{aligned}$$

[3] Joseph Louis Lagrange (1736 – 1813)

$$= \int_{-j}^{m} \left(\prod_{l=0, l\neq i}^{r} \frac{s+l}{-i+l} \right) \, ds. \tag{3.8}$$

Now, different methods can be obtained depending on the choice of m, j, and r and by replacing the generally unknown value $y(x_{p-i})$ in (3.7) with y_{p-i}. There are four important classes of methods. □

Example 3.5. Adams[4]*–Bashforth*[5] *methods.* The class of q-step Adams–Bashforth methods is given by $m = 1$, $j = 0$, and $r = q-1$. It follows that the q-step Adams–Bashforth method uses the nodes $x_{k+1-q}, \ldots, x_k$ for computing the Lagrangian interpolation polynomial. These are q nodes and $p_q(x)$ is at most of degree $q-1$. Adams–Bashforth methods are explicit methods. They have the general form

$$y_{k+1} = y_k + h \sum_{i=0}^{q-1} \beta_i f\left(x_{k-i}, y_{k-i}\right), \tag{3.9}$$

see (3.7), with

$$\beta_i = \int_0^1 \left(\prod_{l=0, l\neq i}^{q-1} \frac{s+l}{-i+l} \right) \, ds, \tag{3.10}$$

compare (3.8).

In the case $q = 1$, the term in the integral in (3.4) is replaced by a constant interpolation polynomial with the node $(x_k, f(x_k, y_k))$. Using the convention that the product is 1 if there is formally no factor in (3.10), this approach yields

$$y_{k+1} = y_k + h \left(\int_0^1 \, ds \right) f\left(x_k, y_k\right) = y_k + h f\left(x_k, y_k\right),$$

i.e., one obtains the explicit Euler method.

If $q = 2$, then the term in the integral is approximated by a linear interpolation polynomial with the nodes $(x_{k-1}, f(x_{k-1}, y_{k-1}))$ and $(x_k, f(x_k, y_k))$. Using (3.9) and (3.10), one obtains

$$\begin{aligned} y_{k+1} &= y_k + h \left[\left(\int_0^1 \frac{s+1}{1} \, ds \right) f\left(x_k, y_k\right) + \left(\int_0^1 \frac{s}{-1} \, ds \right) f\left(x_{k-1}, y_{k-1}\right) \right) \right] \\ &= y_k + h \left[\frac{3}{2} f\left(x_k, y_k\right) - \frac{1}{2} f\left(x_{k-1}, y_{k-1}\right) \right] \\ &= y_k + \frac{h}{2} \left[3 f\left(x_k, y_k\right) - f\left(x_{k-1}, y_{k-1}\right) \right]. \end{aligned}$$

$q \geq 3$, *exercise* □

[4] John Couch Adams (1819 – 1892)

[5] Francis Bashforth (1819 – 1912)

Example 3.6. Adams–Moulton[6] *methods.* Adams–Moulton methods are defined by $m = 0$, $j = 1$, and $r = q$. Hence, it follows that

$$\beta_i = \int_{-1}^{0} \left(\prod_{l=0, l\neq i}^{q} \frac{s+l}{-i+l} \right) \, ds$$

and from (3.7) that

$$y_k = y_{k-1} + h \sum_{i=0}^{q} \beta_i f\left(x_{k-i}, y_{k-i}\right)$$

or, by transforming the index,

$$y_{k+1} = y_k + h \sum_{i=0}^{q} \beta_i f\left(x_{k+1-i}, y_{k+1-i}\right).$$

The $q+1$ nodes of these methods are given by $x_{k+1-q}, \ldots, x_k, x_{k+1}$. That means, Adams–Moulton methods are implicit methods.

This class contains two one-step methods that are obtained for $q = 0$ (which can be used in contrast to the requirement in Definition 3.2) and $q = 1$. Note that the parameter q in (3.2) determines both the previous approximations to be used and the previous arguments of the function f. But in the construction of the methods, three independent parameters were introduced to determine these values. This construction introduces some freedom which allows here to set $q = 0$.

Considering the case $q = 0$, then the term in the integral is replaced by a constant interpolation polynomial with the node at $(x_{k+1}, f(x_{k+1}, y_{k+1}))$. This approach results in the method

$$y_{k+1} = y_k + h \left(\int_{-1}^{0} ds \right) f\left(x_{k+1}, y_{k+1}\right) = y_k + h f\left(x_{k+1}, y_{k+1}\right),$$

which is the implicit Euler method.

For $q = 1$, one uses a linear interpolation polynomial with the points $(x_k, f(x_k, y_k))$ and $(x_{k+1}, f(x_{k+1}, y_{k+1}))$. One gets

$$\begin{aligned} y_{k+1} &= y_k + h \left[\left(\int_{-1}^{0} \frac{s+1}{1} \, ds \right) f\left(x_{k+1}, y_{k+1}\right) + \left(\int_{-1}^{0} \frac{s}{-1} \, ds \right) f\left(x_k, y_k\right) \right] \\ &= y_k + h \left[\frac{1}{2} f\left(x_{k+1}, y_{k+1}\right) + \frac{1}{2} f\left(x_k, y_k\right) \right] \\ &= y_k + \frac{h}{2} \left[f\left(x_{k+1}, y_{k+1}\right) + f\left(x_k, y_k\right) \right]. \end{aligned}$$

[6] Forest Ray Moulton (1872 – 1952)

This method is the trapezoidal rule. □

Example 3.7. Nyström[7] methods. The class of Nyström methods is obtained by using $m = 1$, $j = 1$, and $r = q - 1$. They have the form

$$y_{k+1} = y_{k-1} + h \sum_{i=0}^{q-1} \beta_i f\left(x_{k-i}, y_{k-i}\right)$$

with

$$\beta_i = \int_{-1}^{1} \left(\prod_{l=0, l\neq i}^{q-1} \frac{s+l}{-i+l} \right) \, ds.$$

These methods are explicit and one uses the q nodes $x_{k+1-q}, \ldots, x_k$.

One gets, e.g., for $q = 1$, the method

$$y_{k+1} = y_{k-1} + h \left(\int_{-1}^{1} ds \right) f\left(x_k, y_k\right) = y_{k-1} + 2hf\left(x_k, y_k\right).$$

□

Example 3.8. Milne[8] method. Milne methods are defined by $m = 0$, $j = 2$, and $r = q$. Using a transform of the index, one finds that they have the form

$$y_{k+1} = y_{k-1} + h \sum_{i=0}^{q} \beta_i f\left(x_{k+1-i}, y_{k+1-i}\right)$$

with

$$\beta_i = \int_{-2}^{0} \left(\prod_{l=0, l\neq i}^{q} \frac{s+l}{-i+l} \right) \, ds.$$

Thus, these are implicit methods. □

Remark 3.9. On the coefficients of multi-step methods. One can find tables with the coefficients for multi-step methods in the literature. □

Remark 3.10. Using implicit methods in practice, predictor-corrector methods. If implicit methods are used, then one has to solve in each node x_{k+1} an equation that is generally nonlinear. This step can be performed with some kind of fixed point iteration, e.g., with a method of Newton-type. To achieve a good efficiency of the method, a good initial iterate is of importance. To obtain a good initial iterate, one can use an explicit (multi-step) method. For

[7] Evert J. Nyström (1895 – 1960)

[8] William Edwin Milne (1890 – 1971)

this reason, explicit multi-step methods are called predictor methods and implicit multi-step methods are called corrector methods. The combination of a predictor method with a corrector method is called predictor-corrector method.

Often, it is sufficient for computing the next iterate to perform the predictor step and one or two corrector steps. □

Remark 3.11. Nordsieck[9] *form.* It is possible to transform multi-step methods in a one-step form, the so-called Nordsieck form. This form uses instead of

$$y_k, \ldots, y_{k-q+1}, f\left(x_k, y_k\right), \ldots, f\left(x_{k-q+1}, y_{k-q+1}\right),$$

the values

$$y_k, y'(x_k), y''(x_k), \ldots, y^{(q)}(x_k),$$

see, e.g., (Strehmel *et al.*, 2012, Section 4.4.3). The advantage of the Nordsieck form consists in the possibility of applying a step length control as it is known from one-step methods, Section 1.3. Otherwise, a step length control for form (3.2) of multi-step methods becomes rather complicated. On the other hand, using the Nordsieck form requires that the solution of the initial value problem is q times continuously differentiable. □

3.3 Convergence of Multi-Step Methods

Remark 3.12. Generalities. In this section, linear multi-step methods of the form (3.2) will be considered. Similarly to one-step methods, notations like local error, consistency, or order of convergence will be introduced. The extension of these notations to nonlinear multi-step methods is straightforward. □

Definition 3.13. Local error. Let y_{k+1} be the results of (3.2), $k \geq q$, where the initial values are exactly the values of the solution

$$y_{k+1-q} = y(x_{k+1-q}), \ldots, y_k = y(x_k).$$

Then, the local error is defined by

$$\mathrm{le}(x_{k+1}) = \mathrm{le}_{k+1} = y(x_{k+1}) - \left[\sum_{j=0}^{q-1} a_j y\left(x_{k-j}\right) + h \sum_{j=-1}^{q-1} b_j f\left(x_{k-j}, y(x_{k-j})\right)\right]. \tag{3.11}$$

□

[9] Arnold Nordsieck (1911 – 1971)

Definition 3.14. Consistent method, consistency order. Let $y(x)$ be the solution of the initial value problem (3.1), $S = \{(x, y) \ : \ x \in I = [x_0, x_e], \ y \in \mathbb{R}\}$, and I_N an equidistant mesh on I with N intervals. The multi-step method (3.2) is called consistent if for all $f \in C(S)$, which satisfy in S a Lipschitz condition with respect to y, it holds

$$\lim_{h \to 0} \left(\max_{x_k \in I_N} \frac{\mathrm{le}(x_k + h)}{h} \right) = 0, \quad \text{with} \quad h = \frac{x_e - x_0}{N}. \tag{3.12}$$

If the expression on the left-hand side converges like h^p for $p \geq 1$, then the multi-step scheme has the consistency order p. □

Example 3.15. Consistency order for a Nyström method. The consistency order of a multi-step method can be computed in the same way as for a one-step method by expanding the local error in a Taylor series with respect to h. After having then divided by h, the order of the first non-vanishing term gives the consistency order.

Consider the Nyström method for $q = 3$

$$\begin{aligned} y_{k+1} = y_{k-1} + h \Bigg[& \left(\int_{-1}^{1} \prod_{l=1}^{2} \frac{s+l}{l} \, ds \right) f\left(x_k, y_k\right) \\ & + \left(\int_{-1}^{1} \prod_{l=0, l \neq 1}^{2} \frac{s+l}{-1+l} \, ds \right) f\left(x_{k-1}, y_{k-1}\right) \\ & + \left(\int_{-1}^{1} \prod_{l=0}^{1} \frac{s+l}{-2+l} \, ds \right) f\left(x_{k-2}, y_{k-2}\right) \Bigg] \\ = y_{k-1} + h \Bigg[& \frac{7}{3} f\left(x_k, y_k\right) - \frac{2}{3} f\left(x_{k-1}, y_{k-1}\right) + \frac{1}{3} f\left(x_{k-2}, y_{k-2}\right) \Bigg]. \end{aligned}$$

It follows with (3.11) and (3.1) that

$$\begin{aligned} & \mathrm{le}(x_{k+1}) \\ & = y(x_{k+1}) - y(x_{k-1}) \\ & \quad - h \left[\frac{7}{3} f\left(x_k, y(x_k)\right) - \frac{2}{3} f\left(x_{k-1}, y(x_{k-1})\right) + \frac{1}{3} f\left(x_{k-2}, y(x_{k-2})\right) \right] \\ & = y(x_{k+1}) - y(x_{k-1}) - h \left[\frac{7}{3} y'(x_k) - \frac{2}{3} y'(x_{k-1}) + \frac{1}{3} y'(x_{k-2}) \right]. \end{aligned} \tag{3.13}$$

Now, the the individual terms will be expanded

$$\begin{aligned} y(x_{k+1}) = y(x_k + h) = y(x_k) + h y'(x_k) + \frac{h^2}{2} y''(x_k) + \frac{h^3}{6} y'''(x_k) \\ + \frac{h^4}{24} y^{(4)}(x_k) + \mathcal{O}(h^5), \end{aligned}$$

$$
\begin{aligned}
y(x_{k-1}) &= y(x_k - h) = y(x_k) - hy'(x_k) + \frac{h^2}{2}y''(x_k) - \frac{h^3}{6}y'''(x_k) \\
&\quad + \frac{h^4}{24}y^{(4)}(x_k) + \mathcal{O}(h^5), \\
y'(x_{k-1}) &= y'(x_k - h) = y'(x_k) - hy''(x_k) + \frac{h^2}{2}y'''(x_k) \\
&\quad - \frac{h^3}{6}y^{(4)}(x_k) + \mathcal{O}(h^4), \\
y'(x_{k-2}) &= y'(x_k - 2h) = y'(x_k) - 2hy''(x_k) + 2h^2y'''(x_k) \\
&\quad - \frac{4h^3}{3}y^{(4)}(x_k) + \mathcal{O}(h^4).
\end{aligned}
$$

Inserting these expressions in formula (3.13) for the local error gives

$$
\begin{aligned}
\mathrm{le}(x_{k+1}) &= y(x_k) + hy'(x_k) + \frac{h^2}{2}y''(x_k) + \frac{h^3}{6}y'''(x_k) + \frac{h^4}{24}y^{(4)}(x_k) \\
&\quad - y(x_k) + hy'(x_k) - \frac{h^2}{2}y''(x_k) + \frac{h^3}{6}y'''(x_k) - \frac{h^4}{24}y^{(4)}(x_k) \\
&\quad - \frac{7h}{3}y'(x_k) + \frac{2}{3}\left[hy'(x_k) - h^2y''(x_k) + \frac{h^3}{2}y'''(x_k) - \frac{h^4}{6}y^{(4)}(x_k)\right] \\
&\quad - \frac{1}{3}\left[hy'(x_k) - 2h^2y''(x_k) + 2h^3y'''(x_k) - \frac{4h^4}{3}y^{(4)}(x_k)\right] + \mathcal{O}(h^5) \\
&= \frac{h^4}{3}y^{(4)}(x_k) + \mathcal{O}(h^5).
\end{aligned}
$$

With (3.12), one obtains that this method has consistency order 3. □

Remark 3.16. Linear multi-step methods with a high order of convergence. The goal in constructing multi-step methods consists of course in obtaining convergent methods of high order. A high order of convergence can be expected only if the consistency order is high, i.e., if the local error is small. Using the Taylor series expansion of the terms in the local error and requiring that as many leading terms as possible vanish, one gets a linear system of equations for determining the coefficients a_j, b_j, $j = 0, \ldots, q-1$ and b_{-1} in (3.11). In this way, one obtains a method of the form

$$
y_{k+1} - \sum_{j=0}^{q-1} a_j y_{k-j} = h \sum_{j=-1}^{q-1} b_j f\left(x_{k-j}, y_{k-j}\right). \tag{3.14}
$$

Constructing one-step methods in this way, one always obtains a convergent one-step method, e.g., compare Example 1.29. However, the situation might be different for multi-step methods. □

Example 3.17. Non-convergent multi-step method. Consider the idea from Remark 3.16 for the construction of an explicit linear multi-step method with $q = 2$ and with maximal order of consistency. That means, the ansatz for the method is as follows, compare (3.14),

$$y_{k+1} - a_0 y_k - a_1 y_{k-1} = h\left[b_0 f\left(x_k, y_k\right) + b_1 f\left(x_{k-1}, y_{k-1}\right)\right].$$

The local error has the form

$$\begin{aligned}\mathrm{le}(x_{k+1}) &= y\left(x_{k+1}\right) - a_0 y\left(x_k\right) - a_1 y\left(x_{k-1}\right) - h b_0 f\left(x_k, y(x_k)\right) \\ &\quad - h b_1 f\left(x_{k-1}, y(x_{k-1})\right) \\ &= y\left(x_{k+1}\right) - a_0 y\left(x_k\right) - a_1 y\left(x_{k-1}\right) - h b_0 y'(x_k) - h b_1 y'(x_{k-1}).\end{aligned}$$

Now, the individual terms are expanded in powers of h:

$$\begin{aligned}y\left(x_{k+1}\right) &= y\left(x_k + h\right) = y\left(x_k\right) + h y'\left(x_k\right) + \frac{h^2}{2} y''\left(x_k\right) + \frac{h^3}{6} y'''\left(x_k\right) + \mathcal{O}(h^4), \\ y\left(x_{k-1}\right) &= y\left(x_k - h\right) = y\left(x_k\right) - h y'\left(x_k\right) + \frac{h^2}{2} y''\left(x_k\right) - \frac{h^3}{6} y'''\left(x_k\right) + \mathcal{O}(h^4), \\ y'(x_{k-1}) &= y'\left(x_k - h\right) = y'\left(x_k\right) - h y''\left(x_k\right) + \frac{h^2}{2} y'''\left(x_k\right) + \mathcal{O}(h^3).\end{aligned}$$

Inserting the expansions gives

$$\begin{aligned}\mathrm{le}(x_{k+1}) &= y\left(x_k\right) + h y'\left(x_k\right) + \frac{h^2}{2} y''\left(x_k\right) + \frac{h^3}{6} y'''\left(x_k\right) - a_0 y\left(x_k\right) \\ &\quad - a_1 \left[y\left(x_k\right) - h y'\left(x_k\right) + \frac{h^2}{2} y''\left(x_k\right) - \frac{h^3}{6} y'''\left(x_k\right)\right] - h b_0 y'(x_k) \\ &\quad - h b_1 \left[y'\left(x_k\right) - h y''\left(x_k\right) + \frac{h^2}{2} y'''\left(x_k\right)\right] + \mathcal{O}(h^4) \\ &= \left[1 - a_1 - a_0\right] y\left(x_k\right) + \left[1 + a_1 - b_1 - b_0\right] h y'\left(x_k\right) \\ &\quad + \left[\frac{1}{2} - \frac{a_1}{2} + b_1\right] h^2 y''\left(x_k\right) + \left[\frac{1}{6} + \frac{a_1}{6} - \frac{b_1}{2}\right] h^3 y'''\left(x_k\right) + \mathcal{O}(h^4).\end{aligned}$$

Requiring that the first four terms should vanish leads to the following linear system of equations

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 1 \\ 1/2 & 0 & -1 & 0 \\ -1/6 & 0 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} a_1 \\ a_0 \\ b_1 \\ b_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1/2 \\ 1/6 \end{pmatrix}.$$

The unique solution of this system is $a_1 = 5$, $a_0 = -4$, $b_1 = 2$, $b_0 = 4$. Consequently, one obtains the method

$$y_{k+1} = -4y_k + 5y_{k-1} + h\left[4f\left(x_k, y_k\right) + 2f\left(x_{k-1}, y_{k-1}\right)\right] \tag{3.15}$$

with third order of consistency.

Next, the convergence of the method will be studied at the model initial value problem

$$y'(x) = -y(x), \quad y(0) = 1,$$

with the solution $y(x) = \exp(-x)$. As second initial condition, one takes the value of the solution in the mesh point $x_1 = h$, i.e., $y_1 = \exp(-h)$. Inserting the special form of the right-hand side of the model problem, $f(x_k, y_k) = -y_k$, in (3.15), one can represent the computed solution explicitly. This solution satisfies the homogeneous linear difference equation

$$y_{k+1} + (4 + 4h)y_k + (-5 + 2h)y_{k-1} = 0.$$

The solution of this difference equation can be obtained with the ansatz $y_k = \xi^k$. Inserting this ansatz in the difference equation leads to

$$\xi^{k+1} + (4 + 4h)\xi^k + (-5 + 2h)\xi^{k-1} = 0.$$

This equation is satisfied for $\xi = 0$. Other solutions can be obtained after division by ξ^{k-1} from

$$\xi^2 + (4 + 4h)\xi + (-5 + 2h) = 0. \tag{3.16}$$

One gets the solutions

$$\xi_1(h) = -2 - 2h + 3\sqrt{1 + \frac{2}{3}h + \frac{4}{9}h^2}, \quad \xi_2(h) = -2 - 2h - 3\sqrt{1 + \frac{2}{3}h + \frac{4}{9}h^2}.$$

For simplicity, the dependency on h will be neglected in the notation. The general solution of the difference equations can be represented as a linear combination of the special solutions (superposition principle)

$$y_k = C_1\xi_1^k + C_2\xi_2^k.$$

Now, the constants can be determined from the initial conditions. It holds

$$y_0 = C_1 + C_2 = 1, \quad y_1 = e^{-h} = C_1\xi_1 + C_2\xi_2,$$

from what follows that

$$C_1(h) = \frac{e^{-h} - \xi_2}{\xi_1 - \xi_2}, \quad C_2(h) = -\frac{e^{-h} - \xi_1}{\xi_1 - \xi_2}.$$

Expanding $\xi_1(h)$, $\xi_2(h)$, $C_1(h)$ and $C_2(h)$ in powers of h and inserting these expansions in the solution *(exercise)*, gives for fixed $x > 0$ and $h_N := x/N$

$$y_N = \left[1 + \mathcal{O}\left(\frac{x}{N}\right)\right]\left[1 - \frac{x}{N} + \mathcal{O}\left(\left(\frac{x}{N}\right)^2\right)\right]^N \\ -\frac{1}{216}\left(\frac{x}{N}\right)^4\left[1 + \mathcal{O}\left(\frac{x}{N}\right)\right]\left[-5 - 3\frac{x}{N} + \mathcal{O}\left(\left(\frac{x}{N}\right)^2\right)\right]^N.$$

Considering now the convergence of the method, i.e., $h_N \to 0 \iff N \to \infty$. Then, one obtains for the first term, using well known properties of the exponential, that

$$\lim_{N\to\infty}\left[1 + \mathcal{O}\left(\frac{x}{N}\right)\right]\left[1 - \frac{x}{N} + \mathcal{O}\left(\left(\frac{x}{N}\right)^2\right)\right]^N = e^{-x}.$$

This part approximates the solution of the model problem. For the second term, it holds that

$$-\frac{1}{216}\left(\frac{x}{N}\right)^4\left[1 + \mathcal{O}\left(\frac{x}{N}\right)\right]\left[-5 - 3\frac{x}{N} + \mathcal{O}\left(\left(\frac{x}{N}\right)^2\right)\right]^N \\ = -\frac{(-5)^N}{216}\left(\frac{x}{N}\right)^4\left[1 + \mathcal{O}\left(\frac{x}{N}\right)\right]\left[1 + \frac{3}{5}\frac{x}{N} + \mathcal{O}\left(\left(\frac{x}{N}\right)^2\right)\right]^N.$$

Since

$$\lim_{N\to\infty}\left[1 + \frac{3}{5}\frac{x}{N} + \mathcal{O}\left(\left(\frac{x}{N}\right)^2\right)\right]^N = e^{3x/5},$$

one finds that the second term behaves for large N as follows

$$-\frac{(-5)^N}{216}\left(\frac{x}{N}\right)^4 e^{3x/5}. \tag{3.17}$$

This expression oscillates with increasing N and the modulus is increasing for finer grids ('exponential $(-5)^N$ is stronger than polynomial $(x/N)^4$'), compare the values for $x = 1$ in the following table

N	value of (3.17)
1	0.0421787
2	- 0.1054467
3	0.3514890
4	- 1.3180836
5	5.2723345
6	- 21.96806
7	94.14883
8	- 411.90113
9	1830.6717
10	- 8238.0226

It follows that the method does not converge.

Such an oscillatory behavior can be observed also if the method is applied for solving other initial value problems. For the considered example, the reason for this behavior is that the general solution of the difference equation contains a term that becomes arbitrarily large for large k and for small h (or large N). For the considered method it is

$$\lim_{h\to 0} \xi_2(h) = -5 \quad \Longrightarrow \quad \lim_{k\to\infty} \left|\xi_2^k(h)\right| = \infty.$$

The solution of the difference equation was obtained from the roots of the polynomial (3.16). It can be guessed that the roots of this polynomial will be of importance for the convergence of multi-step methods. □

Definition 3.18. Null stable linear multi-step method. A linear q-step method is called null stable if the first characteristic polynomial

$$\Psi(\xi) = \xi^q - a_0\xi^{q-1} - \ldots - a_{q-1} \tag{3.18}$$

possesses only roots ξ_q with $\left|\xi_q\right| \le 1$ that are simple in the case that $\left|\xi_q\right| = 1$. For the notation 'null stable', compare Remark 3.37 below. □

Example 3.19. Null stability for predictor-corrector methods. The methods from the four most important classes of predictor-corrector methods are null stable.

- *Adams–Bashforth methods, Adams–Moulton methods.* The first characteristic polynomial has the form

  $$\Psi(\xi) = \xi^q - \xi^{q-1} = (\xi - 1)\,\xi^{q-1}.$$

  The only non-trivial root is $\xi_q = 1$ and this root is simple.
- *Nyström methods, Milne methods.* For these methods, the first characteristic polynomial is

  $$\Psi(\xi) = \xi^q - \xi^{q-2} = (\xi + 1)\,(\xi - 1)\,\xi^{q-2}.$$

Hence, the only non-trivial roots are $\xi_q = 1$ and $\xi_q = -1$. They are simple. Null stability does not mean stable in the sense that the method can be applied for the numerical solution of stiff problems, see Example 3.22. □

Theorem 3.20. First Dahlquist[10] barrier. *The maximal order of consistency of a null stable linear q-step method is*

$$p = \begin{cases} q+1 & \text{for } q \text{ odd,} \\ q+2 & \text{for } q \text{ even,} \\ q & \text{if } b_{-1} \le 0, \text{ in particular, if the method is explicit.} \end{cases}$$

[10] Germund Dahlquist (1925 – 2005)

Proof. Only a sketch of the proof is given here, for details see the literature, e.g., (Strehmel *et al.*, 2012, Section 4.2.3) or (Hairer *et al.*, 1993, Section III.3).

First, one sets for $\xi \in \mathbb{C}$, $|\xi| < 1$,

$$z = \frac{\xi - 1}{\xi + 1}.$$

Then, one defines the polynomials

$$R(z) = \left(\frac{1-z}{2}\right)^q \Psi(\xi) = \sum_{l=0}^{q} \alpha_l z^l,$$
$$S(z) = \left(\frac{1-z}{2}\right)^q \sigma(\xi) = \sum_{l=0}^{q} \beta_l z^l,$$

with

$$\sigma(\xi) = b_{-1}\xi^q + b_0 \xi^{q-1} + \ldots + b_{q-1}. \tag{3.19}$$

As next step, one can prove that a linear multi-step method has consistency order p if and only if

$$R(z) \left(\ln \frac{1+z}{1-z}\right)^{-1} - S(z) = \mathcal{O}\left(z^p\right) \quad \text{for } z \to 0.$$

Using a Taylor series expansion of the term with the logarithm, one has on the left-hand side of this statement a polynomial. Now, one studies which coefficients of this polynomial might vanish such that the method is null stable in the individual cases given in the theorem. ■

Example 3.21. Consistency order of some predictor-corrector methods.

- Adams–Bashforth methods with q steps have the consistency order q and Adams–Moulton methods with q steps possess the consistency order $q+1$. Thus, Adams–Moulton methods where q is even have an order that is less than the maximal possible order according to Theorem 3.20.
- The 2-step Milne method (also Milne–Simpson method)

$$y_{k+1} = y_{k-1} + h\left(\frac{1}{3}f(x_{k+1}, y_{k+1}) + \frac{4}{3}f(x_k, y_k) + \frac{1}{3}f(x_{k-1}, y_{k-1})\right) \tag{3.20}$$

  has the consistency order 4. This method achieves the maximal order of consistency for a null stable 2-step method.

□

Example 3.22. Stability of the 2*-step Milne method.* This implicit method is null stable, see Example 3.19, and it possesses the maximal possible order of consistency for a null stable method, see Example 3.21. Thus, so far it shows favorable properties. But having a closer look on its stability reveals that this method has a severe drawback.

Consider again the model initial value problem

$$y'(x) = \lambda y(x), \quad y(0) = 1,$$

with the solution $y(x) = \exp(\lambda x)$. Applying the 2-step Milne method for the solution of this problem, then the method (3.20) has the form

$$y_{k+1} = y_{k-1} + h\lambda \left(\frac{1}{3} y_{k+1} + \frac{4}{3} y_k + \frac{1}{3} y_{k-1} \right).$$

This equation can be rewritten as a linear difference equation

$$\left(1 - \frac{h\lambda}{3}\right) y_{k+1} - \frac{4h\lambda}{3} y_k - \left(1 + \frac{h\lambda}{3}\right) y_{k-1} = 0.$$

The general solution of this difference equation can be represented in the form

$$y_k = C_1 \xi_1^k + C_2 \xi_2^k, \tag{3.21}$$

where $\xi_1(h)$ and $\xi_2(h)$ are the solutions of the quadratic equation

$$\left(1 - \frac{h\lambda}{3}\right) \xi^2 - \frac{4h\lambda}{3} \xi - \left(1 + \frac{h\lambda}{3}\right) = 0.$$

One obtains

$$\xi_1(h) = \frac{3}{3 - h\lambda} \left(\frac{2h\lambda}{3} + \sqrt{1 + \frac{(h\lambda)^2}{3}} \right),$$

$$\xi_2(h) = \frac{3}{3 - h\lambda} \left(\frac{2h\lambda}{3} - \sqrt{1 + \frac{(h\lambda)^2}{3}} \right).$$

Now, the constants C_1, C_2 can be determined from the initial condition and from the value of the first step. It is for $x = 0$

$$C_1 + C_2 = 1 \tag{3.22}$$

and for $x = h$, taking the exact value,

$$e^{\lambda h} = C_1 \xi_1 + (1 - C_1) \xi_2 = (1 - C_2)\, \xi_1 + C_2 \xi_2. \tag{3.23}$$

Expanding $\xi_1(h)$ and $\xi_2(h)$ in powers of h at $h = 0$, one obtains as first order approximation by computing the derivatives of $\xi_1(h)$ and $\xi_2(h)$, respectively, and inserting zero

$$\xi_1(h) = 1 + \lambda h + \mathcal{O}\left(h^2\right), \quad \xi_2(h) = -1 + \frac{\lambda}{3} h + \mathcal{O}\left(h^2\right). \tag{3.24}$$

In the interesting case, $\lambda < 0$, $\xi_1(h)$ approaches 1 from left, with values smaller than 1, and $\xi_2(h)$ approaches -1 also from left, but here the modulus of $\xi_2(h)$ is larger than 1. The last property leads to undesired effects.

For the approximate solution in the node $x_k = kh$, $k = 0, 1, \ldots$, one gets with (3.21)

$$y_k = C_1 \left(1 + \lambda h + \mathcal{O}\left(h^2\right)\right)^{x_k/h} + C_2 \left(-1 + \frac{\lambda}{3}h + \mathcal{O}\left(h^2\right)\right)^{x_k/h}. \quad (3.25)$$

The first term converges to $\exp(\lambda x_k)$ for $h \to 0$, since

$$\lim_{h\to 0} (1 + \lambda h)^{x_k/h} = \lim_{h\to 0} \left(1 + \lambda x_k \frac{h}{x_k}\right)^{x_k/h} = \exp(\lambda x_k).$$

It behaves like the solution of the model initial value problem. The second term behaves for small h like

$$(-1)^{x_k/h} \left(1 - \frac{\lambda}{3}h\right)^{x_k/h}.$$

Here, the second factor converges to $\exp(-\lambda x_k/3)$, but the first factor oscillates for $x_k/h \in \mathbb{N}$. That means, for the stable initial value problem with $\lambda < 0$, this term gives an oscillatory, bounded (for fixed x_k), but exponentially large perturbation.

The behavior of the solution depends on the constants C_1 and C_2. Inserting the expansion (3.24) in condition (3.23) gives

$$\begin{aligned} e^{\lambda h} &= (1 - C_2)\left(1 + \lambda h + \mathcal{O}\left(h^2\right)\right) + C_2 \left(-1 + \frac{\lambda}{3}h + \mathcal{O}\left(h^2\right)\right) \\ &= 1 + \lambda h - 2C_2 - \frac{2\lambda h}{3} C_2 + \mathcal{O}\left(h^2\right). \end{aligned}$$

An expansion of the exponential yields

$$\begin{aligned} 1 + \lambda h + \mathcal{O}\left(h^2\right) = 1 + \lambda h - 2C_2 - \frac{2\lambda h}{3} C_2 + \mathcal{O}\left(h^2\right) \quad &\Longrightarrow \\ \mathcal{O}\left(h^2\right) = -2C_2 - \frac{2\lambda h}{3} C_2. \end{aligned}$$

It follows that $C_2(h) = \mathcal{O}\left(h^2\right)$ and from (3.22), it follows that $C_1 = \mathcal{O}(1)$. In summary, it is for the second term of (3.25)

$$\lim_{h\to 0} C_2(h) \underbrace{\left(-1 + \frac{\lambda}{3}h + \mathcal{O}\left(h^2\right)\right)^{x_k/h}}_{\text{bounded}} = 0.$$

The method converges.

However, the term

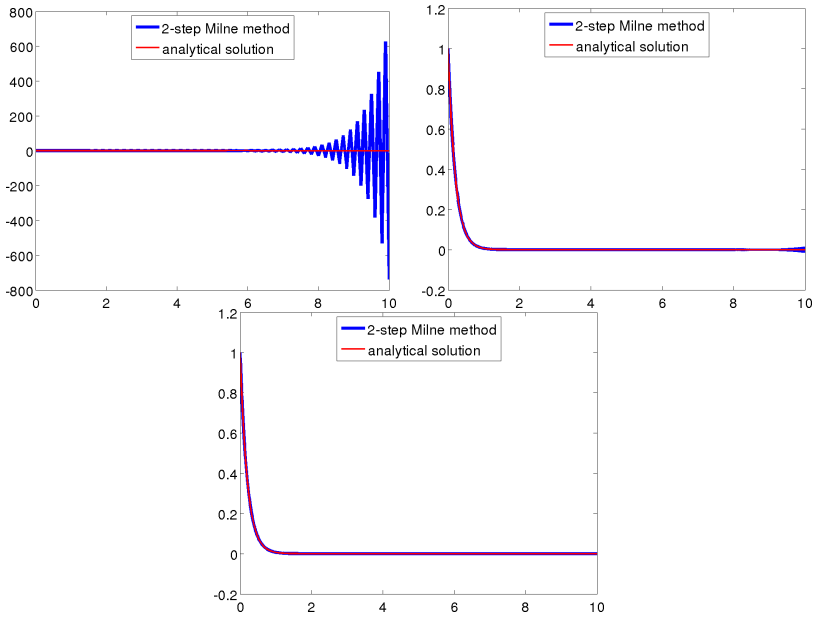

Fig. 3.2 Example 3.22: application of the 2-step Milne method to the model problem with $\lambda = -5$ and $h \in \{0.1, 0.01, 0.001\}$ (left to right, top to bottom).

$$C_2(h)\left(-1+\frac{\lambda}{3}h+\mathcal{O}\left(h^2\right)\right)^{x_k/h} \approx \pm h^2 \exp\left(-\frac{\lambda x_k}{3}\right)$$

becomes small in the case $\lambda \ll -1$ and large x_k only if the step size h is very small, see Figure 3.2.

The behavior found for this method can be observed in practice for all q-step methods of consistency order $q+2$ if these methods are applied to initial value problems with exponentially decaying solution. This kind of instability is a strong restriction of the usefulness of these methods. □

Remark 3.23. Start of multi-step methods and convergence. Apart of the consistency of multi-step methods, one is above all interested in their convergence. For one-step methods, convergence follows from consistency under rather general assumptions and the order of consistency and convergence are the same, see Theorem 1.19. The situation becomes more complicated for multi-step methods.

First of all, one needs for starting a q-step method besides the known initial value $y_0 = y(x_0)$ still $(q-1)$ further approximations $y_1, \ldots, y_{q-1}$ for $y(x_1), \ldots, y(x_{q-1})$. These values can be computed, for instance by a one-step method. The accuracy of these approximations has a strong influence on the accuracy of the q-step method that uses these values. Assume that the

approximations behave as follows

$$y_0 = y(x_0), \quad y_1 = y(x_1) + \varepsilon_1(h), \quad \ldots, \quad y_{q-1} = y(x_{q-1}) + \varepsilon_{q-1}(h).$$

Then, the values that are computed with the q-step method depend also on the perturbations $\varepsilon_1(h), \ldots, \varepsilon_{q-1}(h)$ and one should write for the computed solution in the node x_k more exactly $y_k(\varepsilon, h)$, where $\varepsilon(x, h)$ is a function for which $\varepsilon_i(h) = \varepsilon(x_i, h)$, $i = 1, \ldots, q-1$, holds. □

Definition 3.24. Global error. Let $y(x)$ be the solution of the initial value problem (3.1). Denote the approximations of $y(x)$ that are computed with a multi-step method with step length h by $y_k(\varepsilon, h)$, where the accuracy of the initial approximations is given by the function $\varepsilon(x, h)$. Then, the quantity

$$\mathrm{e}(x_k, \varepsilon, h) := y_k(\varepsilon, h) - y(x_k)$$

is called global error or global discretization error at the node x_k with respect to the step length h and the perturbations $\varepsilon(x, h)$. □

Definition 3.25. Convergence of a multi-step method. Consider the ordinary differential equation of the initial value problem (3.1) in $[a, b]$ and let $x_0 \in [a, b]$. A multi-step method for solving initial value problems of form (3.1) is called convergent if

$$\lim_{n\to\infty} \mathrm{e}(x, \varepsilon, h_n) = 0, \quad \text{with } h_n = \frac{x - x_0}{n},$$

for all $x \in [a, b]$, for all $f \in C^1([a, b] \times \mathbb{R})$, and for all functions $\varepsilon(x, h)$ with

$$\lim_{n\to\infty} |\varepsilon(x, h_n)| = 0, \quad \text{for } x = x_0 + ih_n,\ i = 0, \ldots, q-1.$$

□

Lemma 3.26. A convergent linear multi-step methods is null stable. *A convergent linear multi-step method* (3.2) *is null stable.*

Proof. The proof is performed by contradiction. Assume that the linear multi-step method is not null stable. Consider the initial value problem

$$y'(x) = 0, \quad y(0) = 0,$$

whose solution is $y(x) = 0$. Applying a linear multi-step method of form (3.2) to this problem yields the homogeneous linear difference equation

$$y_{k+1} - \sum_{j=0}^{q-1} a_j y_{k-j} = 0. \tag{3.26}$$

Since the method is assumed to be not null stable, the corresponding first characteristic polynomial $\Psi(\xi)$ has a root with $|\xi_1| > 1$ or a root $|\xi_2| = 1$ that is not simple. Without loss

of generality, let the multiplicity of ξ_1 be one and of ξ_2 be two. Similarly to Example 3.17, one finds that the solution of (3.26) in the node $x_k = kh$ is given by

$$y_k = C_1\xi_1^k + C_2 k\xi_2^k, \quad C_1, C_2 \in \mathbb{R},$$

where one of these coefficients is not zero.

Consider a fixed $\overline{x}$ with $\overline{x} = mh$, $m \in \mathbb{N}$. Choosing $C_1 = C_2 = \sqrt{h}$, where it will be discussed below that this is an admissible choice, the solution in $\overline{x}$ is given by

$$\sqrt{h}\xi_1^{\overline{x}/h} + \frac{\overline{x}}{\sqrt{h}}\xi_2^{\overline{x}/h}. \tag{3.27}$$

For the initial value $\overline{x} = 0$, the value of (3.27) is $\sqrt{h}$ and for the initial value $\overline{x} = h$, it is $\sqrt{h}\xi_1 + \sqrt{h}\xi_2$. Thus, for the initial values, (3.27) converges to the analytic solution as $h \to 0$, so that the choices of C_1 and C_2 were admissible. However, for other values of $\overline{x}$, which is assumed to be fix, both terms in (3.27) diverge. This observation contradicts the assumed convergence of the linear multi-step method. Hence, it is null stable. ■

Theorem 3.27. Connection of convergence and null stability. *Let*

$$y_{k+1} = \sum_{j=0}^{q-1} a_j y_{k-j} + h\Phi(x_{k+1}, \ldots, x_{k+1-q}, y_{k+1}, \ldots, y_{k+1-q}, h) \tag{3.28}$$

be a consistent multi-step method for the solution of initial value problems of form (3.1)*, which is more general than a linear multi-step method. Assume that the incremental function satisfies the following conditions:*

i) $\Phi(x_{k+1}, \ldots, x_{k+1-q}, y_{k+1}, \ldots, y_{k+1-q}, h) \equiv 0$ *for all* $x \in [a, b]$*, all* $y_k \in \mathbb{R}$*, and all* $h \in \mathbb{R}$ *if* $f(x, y) \equiv 0$*.*

ii) Lipschitz continuity with respect to the y-components, i.e., there are constants $h_0 > 0$ *and* M *such that*

$$\begin{aligned}\big|\Phi(x_q, \ldots, x_0, v_q, \ldots, v_0, h) - \Phi(x_q, \ldots, x_0, w_q, \ldots, w_0, h)\big| \\ \leq M \sum_{i=0}^{q} |v_i - w_i|\end{aligned}$$

for all $x_q, \ldots, x_0 \in [a, b]$*, all* $v_i, w_i \in \mathbb{R}$*,* $i = 0, \ldots, q$*, and all step sizes* h *with* $h < h_0$*.*

Then, the multi-step method converges if and only if it is null stable.

Proof. For the proof, it is referred to the literature, e.g., (Strehmel *et al.*, 2012, Section 4.2.5). ■

Remark 3.28. To Theorem 3.27.

- The first assumption and the null stability guarantee that the multi-step method solves the trivial initial value problem

$$y'(x) = 0, \quad y(x_0) = 0,$$

exactly if $\varepsilon_0 = \varepsilon_1 = \ldots = \varepsilon_{q-1} = 0$.

- A linear multi-step method is a special case of (3.28). For linear multi-step methods, the first assumption is always satisfied, since the incremental function is a linear combination of values of the right-hand side $f(x, y)$ of the ordinary differential equation. Due to the same reason, the incremental function of these methods satisfies the second assumption if the right-hand side $f(x, y)$ is Lipschitz continuous with respect to the second argument. Altogether, if the right-hand side of the initial value problem is sufficiently smooth, then a consistent linear multi-step method is convergent if and only if it is null stable.

□

Theorem 3.29. Order of convergence. *Consider a multi-step method of the form* (3.28) *that satisfies the assumptions stated in Theorem 3.27 and which possesses the order of consistency p. Then, it holds for all $f \in C^p([a, b] \times \mathbb{R})$ and for all $x \in [a, b]$ that*

$$|\mathrm{e}(x, \varepsilon, h)| = \mathcal{O}\left(h^p\right),$$

if for the accuracy of the initial values it holds

$$|\varepsilon_i(h)| = \mathcal{O}\left(h^p\right) \quad \textit{for} \quad i = 0, \ldots, q-1.$$

Proof. See literature, e.g., (Strehmel *et al.*, 2012, Section 4.2.5) or (Hairer *et al.*, 1993, Chapter III.4). ■

Remark 3.30. Interpretation of Theorem 3.29. If a multi-step method with consistency order p should also have convergence order p, then it is necessary to compute the initial approximations sufficiently accurately, e.g., with a one-step method of order p. Considering the complete method, which consists of the starting method for computing the approximations $y_1, \ldots, y_{q-1}$ and a predictor-corrector method for computing the other values, then the order of the complete method is determined by the partial method with the lowest order. □

3.4 Backward Difference Formula (BDF) Methods

Remark 3.31. Construction. The construction of Backward Difference Formula (BDF) methods is based on the original initial value problem (3.1) and not on the integral form (3.3) as it is the case for predictor-corrector methods.

Given $q + 1$ nodes $x_{k+1-q}, \ldots, x_{k+1}$ and $q \geq 1$ known approximations of the solution $y_{k+1-q}, \ldots, y_k$. Then, the idea of BDF methods consists in approximating the solution by an interpolation polynomial $p_q(x)$ of degree q with the nodes $\left(x_{k+1-q}, y_{k+1-q}\right), \ldots, (x_k, y_k)$. Now, another condition is needed in order to define a polynomial of degree q and this condition shall also allow to compute y_{k+1}. For BDF methods, one uses the requirement that

this polynomial should satisfy the differential equation (3.1) in x_{k+1}, i.e.,

$$p_q'(x_{k+1}) = f(x_{k+1}, y_{k+1}), \tag{3.29}$$

which leads to an interpolation of Hermite type. It follows from this requirement that BDF methods are implicit methods. □

Example 3.32. BDF methods. Consider an equidistant grid with grid size h.

- $q = 1$. The linear interpolation polynomial through the points (x_k, y_k) and (x_{k+1}, y_{k+1}) is given by the Newton representation (using divided differences)

$$p_1(x) = y_{k+1} + (x - x_{k+1}) \frac{y_k - y_{k+1}}{x_k - x_{k+1}}.$$

It is

$$p_1'(x) = \frac{y_k - y_{k+1}}{x_k - x_{k+1}}$$

such that requirement (3.29) and $x_k - x_{k+1} = -h$ leads to

$$\frac{y_k - y_{k+1}}{-h} = f(x_{k+1}, y_{k+1}) \quad \Longleftrightarrow \quad y_{k+1} = y_k + hf(x_{k+1}, y_{k+1}).$$

Hence, BDF(1) is just the implicit Euler method.

- $q = 2$. The Newton representation of the quadratic interpolation polynomial through $(x_{k-1}, y_{k-1}), (x_k, y_k), (x_{k+1}, y_{k+1})$ is given by

$$\begin{aligned} p_2(x) = y_{k+1} &+ (x - x_{k+1}) \frac{y_k - y_{k+1}}{x_k - x_{k+1}} \\ &+ \frac{(x - x_{k+1})(x - x_k)}{x_{k-1} - x_{k+1}} \left(\frac{y_{k-1} - y_k}{x_{k-1} - x_k} - \frac{y_k - y_{k+1}}{x_k - x_{k+1}} \right). \end{aligned} \tag{3.30}$$

Computing the derivative of this polynomial and using that the grid is equidistant yields

$$p_2'(x) = \frac{y_k - y_{k+1}}{-h} + \frac{(x - x_{k+1}) + (x - x_k)}{-2h} \left(\frac{y_{k-1} - y_k}{-h} - \frac{y_k - y_{k+1}}{-h} \right),$$

such that requirement (3.29) leads to

$$p_2'(x_{k+1}) = \frac{y_{k+1} - y_k}{h} + \frac{h}{2h} \left(\frac{y_{k+1} - 2y_k + y_{k-1}}{h} \right) = f(x_{k+1}, y_{k+1}).$$

Collecting terms gives the BDF(2) method

$$\frac{3}{2} y_{k+1} - 2y_k + \frac{1}{2} y_{k-1} = hf(x_{k+1}, y_{k+1}). \tag{3.31}$$

BDF(2) is the most popular multi-step method for stiff problems.

- $q \geq 3$. The derivation of higher order methods proceeds in the same way. One obtains, e.g., as BDF(3) method

$$\frac{11}{6}y_{k+1} - \frac{18}{6}y_k + \frac{9}{6}y_{k-1} - \frac{2}{6}y_{k-2} = hf\left(x_{k+1}, y_{k+1}\right). \tag{3.32}$$

It should be emphasized that in BDF methods the right-hand side of the initial value problem appears only in one term, namely $f\left(x_{k+1}, y_{k+1}\right)$. This situation is in contrast to the predictor-corrector methods from Section 3.2. This property of BDF methods is of advantage if the computation of the right-hand side is complicated or numerically expensive, like for special discretizations of partial differential equations. □

Lemma 3.33. Null stability of BDF(1), BDF(2), and BDF(3). *The methods BDF(1), BDF(2), and BDF(3) are null stable.*

Proof. The statement of the lemma is obtained by computing the roots of the first characteristic polynomial.

- $q = 1$. The characteristic polynomial is $\lambda - 1$ with the root $\lambda_1 = 1$.
- $q = 2$. The characteristic polynomial of BDF(2) (3.31) is

$$\lambda^2 - \frac{4}{3}\lambda + \frac{1}{3}.$$

A straightforward calculation gives

$$\lambda_1 = \frac{2}{3} + \frac{1}{3} = 1, \quad \lambda_2 = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}.$$

- $q = 3$. For BDF(3), see (3.32), one obtains the characteristic polynomial

$$\lambda^3 - \frac{18}{11}\lambda^2 + \frac{9}{11}\lambda - \frac{2}{11}.$$

By inserting, one checks that $\lambda_1 = 1$ is a root of this polynomial. Extracting the linear factor with this root yields

$$\frac{\lambda^3 - \frac{18}{11}\lambda^2 + \frac{9}{11}\lambda - \frac{2}{11}}{\lambda - 1} = \lambda^2 - \frac{7}{11}\lambda + \frac{2}{11}.$$

The remaining roots are given by the roots of the quadratic polynomial, which are

$$\lambda_2 = \frac{7 + i\sqrt{39}}{22}, \quad \lambda_3 = \frac{7 - i\sqrt{39}}{22},$$

such that $|\lambda_2| = |\lambda_3| = \sqrt{22}/11 \approx 0.4264$.

■

Remark 3.34. Null stability of BDF(q) methods. It can be shown that BDF(q) methods are null stable only for $q \leq 6$, e.g., see Cryer (1972). Hence, BDF(q) methods for $q > 6$ are not of interest. □

Lemma 3.35. Consistency of BDF(q) methods. *BDF(q) methods with $q \leq 6$ are consistent of order q.*

Proof. The proof is obtained by a Taylor series expansion (*exercise*). ■

Theorem 3.36. Convergence of BDF(q) methods. *Let $f \in C^q([a,b]\times\mathbb{R})$ and Lipschitz continuous with respect to the second argument and let the initial values be computed sufficiently accurately, then the BDF(q) methods with $q \leq 6$ are convergent of order q.*

Proof. The incremental function of BDF(q) methods is

$$\Phi(x_{k+1},\ldots,x_{k+1-q},y_{k+1},\ldots,y_{k+1-q},h) = f\left(x_{k+1},y_{k+1}\right),$$

so that the assumptions of Theorem 3.27 are satisfied. Because BDF(q) methods with $q \leq 6$ are null stable and consistent of order q, the other assumptions of Theorem 3.29 are also satisfied and the statement of the theorem follows now from Theorem 3.29. ■

Remark 3.37. On the stability. Stability of multi-step methods is studied at the same initial value problem (2.7) as it was used for one-step methods. In the same way as in Example 3.17, one obtains a homogeneous difference equation

$$y_{k+1} - \sum_{j=0}^{q-1} a_j y_{k-j} = z \sum_{j=-1}^{q-1} b_j y_{k-j}$$

with $z = \lambda h$. With the ansatz $y_k = \xi^k$ and after division by ξ^{k+1-q}, one obtains a characteristic equation

$$\Psi(\xi) - z\sigma(\xi) = 0, \tag{3.33}$$

compare (3.16), where $\Psi(\xi)$ is the first characteristic polynomial (3.18). The polynomial σ has the coefficients b_j, compare (3.19).

Note that for $z = 0$, only the roots of the first characteristic polynomial are considered, which are important for the null stability of the method. This relation might be the reason for the notion 'null' stable. □

Definition 3.38. Stability domain. The set

$$S = \big\{ z \in \mathbb{C} \ : \ \text{for all roots } \xi_l \text{ of (3.33) it holds } |\xi_l| \leq 1; \\ \text{if } \xi_l \text{ is a multiple root, then it holds } |\xi_l| < 1 \ \big\}$$

is called stability domain of a linear multi-step method. □

Definition 3.39. A-stability, $A(\alpha)$-stability. A linear multi-step method is called A-stable if $\mathbb{C}^- \subset S$. It is called $A(\alpha)$-stable with $\alpha \in (0,\pi/2)$ if

$$\big\{ z \in \mathbb{C}^- \text{ with } |\arg(z) - \pi| \leq \alpha \big\} \subset S,$$

with $\arg(z) \in [0, 2\pi)$. □

Theorem 3.40 (Second Dahlquist barrier). *An A-stable linear multi-step method is at most of second order.*

Table 3.1 Values of α (in degree) for the A(α)-stability of BDF(q) methods.

q	1	2	3	4	5	6
α	90	90	86.03	73.35	51.84	17.84

Proof. See literature, e.g., (Strehmel *et al.*, 2012, Section 9.1). ■

Remark 3.41. A(α)-stability of BDF(q) methods. BDF(q) methods are $A(\alpha)$-stable for $q \leq 6$ and even A-stable for $q \leq 2$. The values of α for BDF(q) methods are given in Table 3.1. Because of the small value of α for $q = 6$, the method BDF(6) is not used in practice. □

Remark 3.42. Variable step size for BDF(q) methods. BDF(q) methods can be used on non-equidistant grids, e.g., a formula for BDF(2) with variable step size can be derived on the basis of the quadratic interpolation polynomial (3.30). For $q > 1$ there is some restriction on the admissible change of the mesh size from one mesh cell to its neighbor, e.g., for BDF(2) stability is guaranteed as long as $h_{k+1}/h_k \leq 2.41421$, see (Strehmel *et al.*, 2012, p. 328) for more details. □

Chapter 4
Summary and Outlook

4.1 Comparison of Numerical Methods

Remark 4.1. Motivation. Given an initial value problem in practice, one has to choose a method for its numerical solution. It is desirable to use a method that is appropriate for the given problem. This section discusses some criteria for making the choice. □

Remark 4.2. Criteria for comparing numerical methods for solving initial value problems.

- *Computing time.* Computing time is important in many applications. If the evaluation of the right-hand side of the initial value problem is time-consuming, the number of evaluations is important. For implicit methods, the number of calculations of the Jacobian and the number of LU factorizations is of importance.
- *Accuracy.* Computing an accurate numerical solution is of course desirable. However, aiming for high accuracy is often in conflict with having short computing times. An easy step length control should be possible.
- *Memory.* On modern computers, the amount of memory is usually not a big issue. However, if the given initial value problem has special structures, like a sparse Jacobian, such structures should be supported by the numerical method. It should be kept into consideration that on modern computers the access to the memory determines essentially the computing time (and not the number of floating point operations).
- *Reliability.* The step length control (or order control) should be sensitive with respect to local changes of the right-hand side and act in a correct way.
- *Robustness.* The method should work also for complicated examples. It should be flexible with the step length control, e.g., reduce the step length appropriately if the right-hand side has steep gradients.
- *Simplicity.* In complex applications, often the simplicity of the method is of importance.

□

Remark 4.3. Some experience. There are much more numerical methods for solving initial value problems than presented in this course. Here, only some remarks to the presented methods are given.

- *Non-stiff problems.* For such problems, explicit one-step and linear multi-step methods were presented. Often, one-step methods need less steps than multi-step methods and they require fewer evaluations of the right-hand side. A few popular explicit one-step methods are given in Remark 1.45.
- *Complex problems.* For complex problems, e.g., from fluid dynamics, where one has an initial value problem with respect to time, very often only the simplest methods are used, like the explicit or implicit Euler method, the trapezoidal rule (Crank–Nicolson scheme), or sometimes BDF2. An efficient and theoretically supported step length control is not possible with these methods. Other methods, like Rosenbrock schemes, have been used only for academic problems so far, e.g., in John & Rang (2010).

□

4.2 Boundary Value Problems

Remark 4.4. A one-dimensional boundary value problem. Boundary value problems prescribe, in contrast to initial value problems, values at (some part of) the boundary of the domain. A typical example in one dimension is

$$-u'' = f \quad \text{in } (0,1), \quad u(0) = a,\ u(1) = b, \tag{4.1}$$

with given values $a, b \in \mathbb{R}$. Solving (4.1) can be performed in principal by integrating the right-hand side of the differential equation twice. Whether or not this analytic calculation can be performed depends on f. Each integration gives a constant, such that the general solution of the differential equation has two constants. The values of these constants can be determined with the given boundary conditions. □

Remark 4.5. Boundary value problems in higher dimensions. Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded domain. Then, a typical boundary value problem is

$$-\Delta u = -\sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2} = f \quad \text{in } \Omega, \quad u = g \quad \text{on } \partial\Omega, \tag{4.2}$$

where Δ is the Laplacian and $\partial\Omega$ is the boundary of Ω. Solutions of problems of type (4.2) can be hardly found analytically.

The topic of Numerical Mathematics III will be the introduction of numerical methods for solving problems of type (4.2). Such methods include Finite Difference Methods, Finite Element Methods, and Finite Volume Methods. □

4.3 Differential-Algebraic Equations

Remark 4.6. Differential-Algebraic Equations (DAEs). In many applications, the modeling of processes leads to a coupled system of equations of different type. A typical example are systems of the form

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)), \\ 0 &= g(t, y(t), z(t)), \end{aligned} \tag{4.3}$$

with given functions f and g. In (4.3), the derivative of y with respect to t occurs, but not the derivative of z with respect to t. Problem (4.3) is called semi-explicit differential-algebraic equation (DAE), the variable y is the differential variable, and z is the algebraic variable. In this context, the notion 'algebraic' means that there are no derivatives. □

Example 4.7. Equations for incompressible fluids. Equations for the behavior of incompressible fluids are derived on the basis of two conservation laws. The first one is Newton's second law of motion (net force equals mass times acceleration, conservation of linear momentum) and the second one is the conservation of mass. The unknown variables in these equations are the velocity $\boldsymbol{u}(t, x, y, z)$ and the pressure $p(t, x, y, z)$, where t is the time, (x, y, z) the spatial variable, and

$$\boldsymbol{u}(t, x, y, z) = \begin{pmatrix} u_1(t, x, y, z) \\ u_2(t, x, y, z) \\ u_3(t, x, y, z) \end{pmatrix}.$$

Whereas the conservation of linear momentum contains the temporal derivative $\partial_t \boldsymbol{u}$ of the velocity, i.e., it is a differential equation with respect to time, the conservation of mass reads as follows

$$\nabla \cdot \boldsymbol{u} = \partial_x u_1 + \partial_y u_2 + \partial_z u_3 = 0. \tag{4.4}$$

Thus, one obtains a coupled model of the differential equation (with respect to time) and the algebraic equation (4.4). □

Remark 4.8. Theory of DAEs. There is not sufficient time for presenting the theory of DAEs. One can find it, e.g., in (Strehmel *et al.*, 2012, Chapter 13) or (Kunkel & Mehrmann, 2006, Part I). □

Remark 4.9. Direct approach for the discretization of DAEs. In the so-called direct approach, the DAE (4.3) is embedded in the so-called singularly perturbed problem

$$\begin{aligned} y'(t) &= f(t, y(t), z(t)), \\ \varepsilon z'(t) &= g(t, y(t), z(t)), \end{aligned} \tag{4.5}$$

with $0 < \varepsilon \ll 1$. Problem (4.5) is an ODE, for which the methods presented in this course can be applied. Formulating these methods for problem (4.5), then the so-called singular perturbation parameter ε appears. Setting then $\varepsilon = 0$ in these formulations leads formally to methods for the DAE (4.3). Now, one has to study the properties of these methods. □

Appendix A
Topics on the Theory of Ordinary Differential Equations

A.1 Ordinary Differential Equations of Higher Order

Remark A.1. Motivation. The notation of stiffness comes from the consideration of first order systems of ordinary differential equations. There are some connections of such systems to ordinary differential equations of higher order, e.g. a solution method for linear first order systems requires the solution of a higher order linear differential equation, see Remark A.36. □

A.1.1 Definition, Connection to First Order Systems

Definition A.2. General and explicit n-th order ordinary differential equation. The general ordinary differential equation of order n has the form

$$F\left(x, y(x), y'(x), \ldots, y^{(n)}(x)\right) = 0. \tag{A.1}$$

This equation is called explicit, if one can write it in the form

$$y^{(n)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right). \tag{A.2}$$

The function $y(x)$ is a solution of (A.1) in an interval I if $y(x)$ is n times continuously differentiable in I and if $y(x)$ satisfies (A.1).

Let $x_0 \in I$ be given. Then, (A.1) together with the conditions

$$y(x_0) = y_0,\ y'(x_0) = y_1, \ldots, y^{(n-1)}(x_0) = y_{n-1}$$

is called initial value problem for (A.1). □

Example A.3. Special cases. The general resp. explicit ordinary differential equation of higher order can be solved analytically only in special cases. Two special cases, that will not be considered here, are as follows:

- Consider the second order differential equation

$$y''(x) = f(x, y'(x)).$$

Substituting $y'(x) = z(x)$, one obtains a first order differential equation for $z(x)$

$$z'(x) = f(x, z(x)).$$

If one can solve this equation analytically, one gets $y'(x)$. If it is then possible to find a primitive of $y'(x)$, one has computed an analytical solution of the differential equation of second order. In the case of an initial value problem with

$$y(x_0) = y_0,\ y'(x_0) = y_1,$$

the initial value for the first order differential equation is

$$z(x_0) = y_1.$$

The second initial value is needed for determining the constant of the primitive of $y'(x)$.

- Consider the differential equation of second order

$$y''(x) = f(y, y').$$

Let a solution $y(x)$ of this differential equation be known and let $y^{-1}(y)$ its inverse function, i.e. $y^{-1}(y(x)) = x$. Then, one can use the ansatz

$$p(y) := y'\left(y^{-1}(y)\right).$$

With the rule for differentiating the inverse function $((f^{-1})'(y_0) = 1/f'(x_0))$, one obtains

$$\begin{aligned}\frac{dp}{dy}(y) &= y''\left(y^{-1}(y)\right)\frac{d}{dy}\left(y^{-1}(y(x))\right) = \frac{y''\left(y^{-1}(y)\right)}{y'(x)} = \frac{y''\left(y^{-1}(y)\right)}{y'\left(y^{-1}(y)\right)} \\ &= \frac{y''\left(y^{-1}(y)\right)}{p(y)} = \frac{y''(x)}{p(y)}.\end{aligned}$$

This approach leads then to the first order differential equation

$$p'(y) = \frac{f(y, p(y))}{p(y)}.$$

□

Theorem A.4. Connection of explicit ordinary differential equations of higher order and systems of differential equations of first order. *Every explicit differential equation of n-th order* (A.2) *can be transformed equivalently to a system of n differential equations of first order*

$$\begin{aligned} y_k'(x) &= y_{k+1}(x), \quad k = 1, \dots, n-1, \\ y_n'(x) &= f(x, y_1(x), \dots, y_n(x)) \end{aligned} \tag{A.3}$$

or (note that the system is generally nonlinear, since the unknown functions appear also in $f(\cdot, \dots, \cdot)$)

$$\boldsymbol{y}'(x) = \begin{pmatrix} y_1'(x) \\ y_2'(x) \\ \vdots \\ y_n'(x) \end{pmatrix} = \begin{pmatrix} 0\,1\,0 \cdots 0 \\ 0\,0\,1 \cdots 0 \\ \vdots\,\vdots\,\vdots\,\vdots\,\vdots \\ 0\,0\,0 \cdots 0 \end{pmatrix} \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ f(x, y_1, \dots, y_n) \end{pmatrix}$$

for the n functions $y_1(x), \dots, y_n(x)$. The solution of (A.2) *is $y(x) = y_1(x)$.*

Proof. Insert in (A.2)

$$\begin{aligned} &y_1(x) := y(x), \quad y_2(x) := y_1'(x) = y'(x), \quad y_3(x) := y_2'(x) = y''(x), \quad \dots \\ &y_n(x) := y_{n-1}'(x) = y^{(n-1)}(x). \end{aligned}$$

If $y \in C^n(I)$ is a solution of (A.2), then $y_1(x), \dots, y_n(x)$ is obviously a solution of (A.3) in I.

Conversely, if $y_1(x), \dots, y_n(x) \in C^1(I)$ is a solution of (A.3), then it holds

$$\begin{aligned} &y_2(x) = y_1'(x), \quad y_3(x) = y_2'(x) = y_1''(x), \dots, y_n(x) = y_1^{(n-1)}(x) \\ &y_n'(x) = y_1^{(n)}(x) = f(x, y_1, \dots, y_n). \end{aligned}$$

Hence, the function $y_1(x)$ is n times continuously differentiable and it is the solution of (A.2) in I. ■

Example A.5. Transform of a higher order differential equation into a system of first order equations. The third order differential equation

$$y'''(x) + 2y''(x) - 5y'(x) = f(x, y(x))$$

can be transformed into the form

$$\begin{aligned} y_1(x) &= y(x) \\ y_1'(x) &= y_2(x) (= y'(x)) \\ y_2'(x) &= y_3(x) (= y''(x)) \\ y_3'(x) &= y'''(x) = -2y''(x) + 5y'(x) + f(x, y(x)) \\ &= -2y_3(x) + 5y_2(x) + f(x, y_1(x)). \end{aligned}$$

□

A.1.2 Linear Differential Equations of n-th Order

Definition A.6. Linear n-th order differential equations. A linear differential equation of n-th order has the form

$$a_n(x)y^{(n)}(x)+a_{n-1}(x)y^{(n-1)}(x)+\ldots+a_1(x)y'(x)+a_0(x)y(x)=f(x), \quad \text{(A.4)}$$

where the functions $a_0(x),\ldots,a_n(x)$ are continuous in the interval I, in which a solution of (A.4) is searched, and it holds $a_n(x)\neq 0$ in I. The linear n-th order differential equation is called homogeneous if $f(x)=0$ for all $x\in I$

$$a_n(x)y^{(n)}(x)+a_{n-1}(x)y^{(n-1)}(x)+\ldots+a_1(x)y'(x)+a_0(x)y(x)=0. \quad \text{(A.5)}$$

□

Theorem A.7. Superposition principle for linear differential equations of higher order. *Consider the linear differential equation of n-th order* (A.4)*, then the superposition principle holds:*

i) If $y_1(x)$ and $y_2(x)$ are two solutions of the homogeneous equation (A.5)*, then $c_1y_1(x)+c_2y_2(x)$, $c_1,c_2\in\mathbb{R}$, is a solution of the homogeneous equation, too.*

ii) If $y_0(x)$ is a solution of the inhomogeneous equation and $y_1(x)$ is a solution of the homogeneous equation, then $y_0(x)+y_1(x)$ is a solution of the inhomogeneous equation.

iii) If $y_1(x)$ and $y_2(x)$ are two solutions of the inhomogeneous equation, then $y_1(x)-y_2(x)$ is a solution of the homogeneous equation.

Proof. Direct calculations, exercise. ■

Corollary A.8. General solution of the inhomogeneous differential equation. *The general solution of* (A.4) *is the sum of the general solution of the homogeneous linear differential equation of n-th order* (A.5) *and one special solution of the inhomogeneous n-th order differential equation* (A.4)*.*

Remark A.9. Transform in a linear system of ordinary differential equations of first order. A linear differential equation of n-th order can be transformed equivalently into a linear $n\times n$ system

$$\begin{aligned} y_k'(x) &= y_{k+1}(x), \quad k=1,\ldots,n-1,\\ y_n'(x) &= -\sum_{i=0}^{n-1}\frac{a_i(x)}{a_n(x)}y_{i+1}(x)+\frac{f(x)}{a_n(x)} \end{aligned}$$

or

$$\boldsymbol{y}'(x) = \begin{pmatrix} y_1'(x) \\ y_2'(x) \\ \vdots \\ y_n'(x) \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{a_0(x)}{a_n(x)} & -\frac{a_1(x)}{a_n(x)} & -\frac{a_2(x)}{a_n(x)} & \cdots & -\frac{a_{n-1}(x)}{a_n(x)} \end{pmatrix} \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{f(x)}{a_n(x)} \end{pmatrix}$$

$$=: A(x)\boldsymbol{y}(x) + \boldsymbol{f}(x). \tag{A.6}$$

□

Theorem A.10. Existence and uniqueness of a solution of the initial value problem. *Let $I = [x_0 - a, x_0 + a]$ and $a_i \in C(I)$, $i = 0, \ldots, n$, $f \in C(I)$. Then, the linear differential equation of n-th order* (A.4) *has exactly one solution $y \in C^n(I)$ for given initial value*

$$y(x_0) = y_0,\ y'(x_0) = y_1, \ldots, y^{(n-1)}(x_0) = y_{n-1}.$$

Proof. Since (A.4) is equivalent to the system (A.6), one can apply the theorem on global existence and uniqueness of a solution of an initial value problem from Picard–Lindelöf, see lecture notes Numerical Mathematics I or the literature. To this end, one has to show the Lipschitz continuity of the right-hand side of (A.6) with respect to $y_1, \ldots, y_n$. Denoting the right-hand side by $F(x, \boldsymbol{y})$ gives

$$\|\mathbf{F}(x, \boldsymbol{y}) - \mathbf{F}(x, \tilde{\boldsymbol{y}})\|_{[C(I)]^n} = \|A(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{[C(I)]^n} \leq \|A\|_{[C(I)]^n, \infty} \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_{[C(I)]^n},$$

where one uses the triangle inequality to get

$$\begin{aligned} \|A_{i\cdot}\boldsymbol{y}\|_{C(I)} &= \max_{x \in I} \left| \sum_{j=1}^{n} a_{ij}(x) y_j(x) \right| \leq \max_{x \in I} \sum_{j=1}^{n} |a_{ij}(x)| \max_{j=1,\ldots,n} \left\{ \max_{x \in I} |y_j(x)| \right\} \\ &= \|A_{i\cdot}\|_{C(I)} \|\boldsymbol{y}\|_{[C(I)]^n} \end{aligned}$$

for $i = 1, \ldots, n$. Now, one can choose

$$L = \|A\|_{[C(I)]^n, \infty} = \max_{x \in I} \left\{ \max \left\{ 1, \left| \frac{a_1(x)}{a_n(x)} \right| + \ldots + \left| \frac{a_{n-1}(x)}{a_n(x)} \right| \right\} \right\}.$$

All terms are bounded since I is closed (compact) and continuous functions are bounded on compact sets. ■

Definition A.11. Linearly independent solutions, fundamental system. The solutions $y_i(x) : I \to \mathbb{R}$, $i = 1, \ldots, k$, of (A.5) are called linearly independent if from

$$\sum_{i=1}^{k} c_i y_i(x) = 0, \quad \text{for all } x \in I,\ c_i \in \mathbb{R},$$

it follows that $c_i = 0$ for $i = 1, \ldots, k$. A set of n linearly independent solutions is called a fundamental system of (A.5). □

Definition A.12. Wronski[1] matrix, Wronski determinant. Let $y_i(x)$, $i = 1, \ldots, k$, be solutions of (A.5). The matrix

$$\mathcal{W}(x) = \begin{pmatrix} y_1(x) & \ldots & y_k(x) \\ y_1'(x) & \ldots & y_k'(x) \\ & \vdots & \\ y_1^{(n-1)}(x) & \ldots & y_k^{(n-1)}(x) \end{pmatrix}$$

is called Wronski matrix. For $k = n$ the Wronski determinant is given by $\det(\mathcal{W})(x) =: W(x)$. □

Lemma A.13. Properties of the Wronski matrix and Wronski determinant. *Let $I = [a, b]$ and let $y_1(x), \ldots, y_n(x)$ be solutions of* (A.5).

i) The Wronski determinant fulfills the linear first order differential equation

$$W'(x) = -\frac{a_{n-1}(x)}{a_n(x)} W(x).$$

ii) It holds for all $x \in I$

$$W(x) = W(x_0) \exp\left(- \int_{x_0}^{x} \frac{a_{n-1}(t)}{a_n(t)} \, dt \right)$$

with arbitrary $x_0 \in I$.

iii) If there exists a $x_0 \in I$ with $W(x_0) \neq 0$, then it holds $W(x) \neq 0$ for all $x \in I$.

iv) If there exists a $x_0 \in I$ with rank$(\mathcal{W}(x_0)) = k$, then there are at least k solutions of (A.5)*, e.g. $y_1(x), \ldots, y_k(x)$, linearly independent.*

Proof. i) Let S_n be the set of all permutations of $\{1, \ldots, n\}$ and let $\sigma \in S_n$. Denote the entries of the Wronski matrix by $\mathcal{W}(x) = \big(y_{jk}(x)\big)_{j,k=1}^{n}$. If $\sigma = (\sigma_1, \ldots, \sigma_n)$, then let

$$\prod_{j=1}^{n} y_{j,\sigma_j}(x) = \left(y_{1,\sigma_1} y_{2,\sigma_2} \ldots y_{n,\sigma_n}\right)(x).$$

Applying the Laplace[2] formula for determinants and the product rule yields

[1] Joseph Marie Wronski (1758 – 1853)

[2] Pierre–Simon (Marquis de) Laplace (1749 – 1827)

$$\begin{aligned}
\frac{d}{dx}\det(\mathcal{W}(x)) &= \frac{d}{dx}\left(\sum_{\sigma\in S_n}\left(\operatorname{sgn}(\sigma)\prod_{j=1}^{n} y_{j,\sigma_j}(x)\right)\right)\\
&= \sum_{\sigma\in S_n}\left(\operatorname{sgn}(\sigma)\sum_{i=1}^{n}\left(\prod_{j=1,j\neq i}^{n} y_{j,\sigma_j}(x)\right) y'_{i,\sigma_i}(x)\right)\\
&= \sum_{i=1}^{n}\left(\sum_{\sigma\in S_n}\left(\operatorname{sgn}(\sigma)\prod_{j=1,j\neq i}^{n} y_{j,\sigma_j}(x) y'_{i,\sigma_i}(x)\right)\right)\\
&= \sum_{i=1}^{n}\det\begin{pmatrix} \cdots & \cdots & \cdots\\ \left(y_1^{(i-1)}(x)\right)' & \cdots & \left(y_n^{(i-1)}(x)\right)'\\ \cdots & \cdots & \cdots \end{pmatrix}.
\end{aligned}$$

exercise for $n = 2, 3$. In the last step, again the Laplace formula for determinants was applied. In the i-th row of the last matrix is the first derivative of the corresponding row of the Wronski matrix, i.e. there is the i-th order derivative of $(y_1(x), \ldots, y_n(x))$. The rows with dots in this matrix coincide with the respective rows of $\mathcal{W}(x)$. For $i = 1, \ldots, n-1$, the determinants vanish, since in these cases there are two identical rows, namely row i and $i+1$. Thus, it is

$$\frac{d}{dx}\det(\mathcal{W}(x)) = \det\begin{pmatrix} y_1(x) & \ldots & y_n(x)\\ y'_1(x) & \ldots & y'_n(x)\\ & \vdots & \\ y_1^{(n-2)}(x) & \ldots & y_n^{(n-2)}(x)\\ y_1^{(n)}(x) & \ldots & y_n^{(n)}(x) \end{pmatrix}.$$

Now, one uses that $y_1(x), \ldots, y_n(x)$ are solutions of (A.5) and one replaces the n-th derivative in the last row by (A.5). Using rules for the evaluation of determinants, one obtains

$$\frac{d}{dx}\det(\mathcal{W}(x)) = \sum_{i=1}^{n} -\frac{a_{i-1}(x)}{a_n(x)}\det\begin{pmatrix} y_1(x) & \ldots & y_n(x)\\ y'_1(x) & \ldots & y'_n(x)\\ & \vdots & \\ y_1^{(i-1)}(x) & \ldots & y_n^{(i-1)}(x) \end{pmatrix}.$$

Apart of the last term, all other determinants vanish, since all other terms have two identical rows, namely the i-th row and the last row.

ii) This term is the solution of the initial value problem for the Wronski determinant and the initial value $W(x_0)$, see the respective theorem in the lecture notes of Numerical Mathematics I.

iii) This statement follows directly from ii) since the exponential does not vanish.

iv) *exercise*

■

Theorem A.14. Existence of a fundamental system, representation of the solution of a homogeneous linear differential equation of n-th order by the fundamental system. *Let $I = [a, b]$ with $x_0 \in I$. The homogeneous equation* (A.5) *has a fundamental system in I. Each solution of* (A.5) *can be written as a linear combination of the solutions of an arbitrary fundamental system.*

Proof. Consider n homogeneous initial value problems with the initial values

$$y_j^{(i-1)}(x_0) = \delta_{ij}, \quad i,j = 1,\dots,n.$$

Each of these initial value problems has a unique solution $y_j(x)$, see Theorem A.10. It is $W(x_0) = 1$ for these solutions. From Lemma A.13, iii), it follows that $\{y_1(x),\dots,y_n(x)\}$ is a fundamental system.

Let $y(x)$ be an arbitrary solution of (A.5) with the initial values $y^{(i-1)}(x_0) = \tilde{y}_{i-1}$, $i = 1,\dots,n$, and $\{y_1(x),\dots,y_n(x)\}$ an arbitrary fundamental system. The system

$$\begin{pmatrix} y_1(x_0) & \dots & y_n(x_0) \\ y_1'(x_0) & \dots & y_n'(x_0) \\ & \vdots & \\ y_1^{(n-1)}(x_0) & \dots & y_n^{(n-1)}(x_0) \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \end{pmatrix} = \begin{pmatrix} \tilde{y}_0 \\ \tilde{y}_1 \\ \vdots \\ \tilde{y}_{n-1} \end{pmatrix}$$

has a unique solution since the matrix spanned by a fundamental system is not singular. The function $\sum_{i=1}^n c_{i-1} y_i(x)$ satisfies the initial conditions (these are just the equations of the system) and, because of the superposition principle, it is a solution of (A.5). Since the solution of the initial value problem to (A.5) is unique, Theorem A.10, it follows that $y(x) = \sum_{i=1}^n c_{i-1} y_i(x)$. ■

Theorem A.15. Special solution of the inhomogeneous equation. *Let $\{y_1(x),\dots,y_n(x)\}$ be a fundamental system of the homogeneous equation* (A.5) *in $I = [a,b]$. In addition, let $W_l(x)$ be the determinant, which is obtained from the Wronski determinant $W(x)$ with respect to $\{y_1(x),\dots,y_n(x)\}$ by replacing the l-th column by $(0,0,\dots,f(x)/a_n(x))^T$. Then,*

$$y(x) = \sum_{l=1}^n y_l(x) \int_{x_0}^x \frac{W_l(t)}{W(t)}\, dt, \quad x_0, x \in I,$$

is a solution of the inhomogeneous equation (A.4).

Proof. The proof uses the principle of the variation of the constants. This principle will be explained in a simpler setting in Remark A.27. For details of the proof, see the literature. ■

A.1.3 Linear n-th Order Differential Equations with Constant Coefficients

Definition A.16. Linear differential equation of n-th order with constant coefficients. A linear n-th order differential equation with constant coefficients has the form

$$a_n y^{(n)}(x) + a_{n-1} y^{(n-1)}(x) + \dots + a_1 y'(x) + a_0 y(x) = f(x), \tag{A.7}$$

with $a_i \in \mathbb{R}$, $i = 0,\dots,n$, $a_n \neq 0$. □

A.1.3.1 The Homogeneous Equation

Remark A.17. Basic approach for solving the homogeneous linear differential equation of n-th order with constant coefficients. Because of the superposition principle, one needs the general solution of the homogeneous differential equation. That means, one has to find a fundamental system, i.e. n linearly independent solutions.

Consider

$$\sum_{i=0}^{n} a_i y_h^{(i)}(x) = 0. \tag{A.8}$$

In the case of a differential equation of first order, i.e. $n = 1$,

$$a_1 y_h'(x) + a_0 y_h(x) = 0,$$

one can get the solution by the method of separating the variables (unknowns), see lecture notes of Numerical Mathematics I. One obtains

$$y_h(x) = c \exp\left(-\frac{a_0}{a_1} x\right), \quad c \in \mathbb{R}.$$

One uses the same structural ansatz for computing the solution of (A.8)

$$y_h(x) = e^{\lambda x}, \quad \lambda \in \mathbb{C}. \tag{A.9}$$

It follows that

$$y_h'(x) = \lambda e^{\lambda x}, \ldots, y_h^{(n)}(x) = \lambda^n e^{\lambda x}.$$

Inserting into (A.8) gives

$$\left(a_n \lambda^n + a_{n-1} \lambda^{n-1} + \ldots + a_1 \lambda + a_0\right) e^{\lambda x} = 0. \tag{A.10}$$

It is $e^{\lambda x} \neq 0$, also for complex λ. Because, using Euler's formula, it holds for $\lambda = a + ib$, $a, b \in \mathbb{R}$, that

$$e^{\lambda x} = e^{ax} \left(\cos(bx) + i \sin(bx)\right) = e^{ax} \cos(bx) + i e^{ax} \sin(bx).$$

A complex number is zero iff its real part and its imaginary part are vanish. It is $e^{ax} > 0$ and there does not exist a $(bx) \in \mathbb{R}$ such that at the same time $\sin(bx)$ and $\cos(bx)$ vanish. Hence, $e^{\lambda x} \neq 0$.

The equation (A.10) is satisfied iff one of the factors is equal to zero. Since the second factor cannot vanish, it must hold

$$p(\lambda) := a_n \lambda^n + a_{n-1} \lambda^{n-1} + \ldots + a_1 \lambda + a_0 = 0.$$

The function $p(\lambda)$ is called characteristic polynomial of (A.8). The roots of the characteristic polynomial are the values of λ in the ansatz of $y_h(x)$.

From the fundamental theorem of algebra it holds that $p(\lambda)$ has exactly n roots, which do not need to be mutually different. Since the coefficients of $p(\lambda)$ are real numbers, it follows that with each complex root $\lambda_1 = a + ib$, $a, b \in \mathbb{R}$, $b \neq 0$, also its conjugate $\lambda_2 = a - ib$ is a root of $p(\lambda)$.

It will be shown that the basic ansatz (A.9) is not sufficient in the case of multiple roots. □

Theorem A.18. Linearly independent solutions in the case of real roots with multiplicity k. *Let $\lambda_0 \in \mathbb{R}$ be a real root of the characteristic polynomial $p(\lambda)$ with multiplicity k, $1 \leq k \leq n$. Then, one can obtain with λ_0 the k linearly independent solutions of* (A.8)

$$y_{h,1}(x) = e^{\lambda_0 x},\ y_{h,2}(x) = xe^{\lambda_0 x},\ \ldots,\ y_{h,k}(x) = x^{k-1}e^{\lambda_0 x}. \tag{A.11}$$

Proof. For $k = 2$.

$y_{h,1}(x), y_{h,2}(x)$ solve (A.8). This statement is already clear for $y_{h,1}(x)$ since this function has the form of the ansatz (A.9). For $y_{h,2}(x)$ it holds

$$\begin{aligned} y'_{h,2}(x) &= (1 + \lambda_0 x)\, e^{\lambda_0 x}, \\ y''_{h,2}(x) &= \left(2\lambda_0 + \lambda_0^2 x\right) e^{\lambda_0 x}, \\ &\vdots \\ y^{(n)}_{h,2}(x) &= \left(n\lambda_0^{n-1} + \lambda_0^n x\right) e^{\lambda_0 x}. \end{aligned}$$

Inserting into the left-hand side of (A.8) yields

$$e^{\lambda_0 x} \sum_{i=0}^{n} a_i (i\lambda_0^{i-1} + \lambda_0^i x) = e^{\lambda_0 x} \left(x \underbrace{\sum_{i=0}^{n} a_i \lambda_0^i}_{p(\lambda_0)} + \underbrace{\sum_{i=0}^{n} a_i i \lambda_0^i}_{p'(\lambda_0)} \right). \tag{A.12}$$

It is $p(\lambda_0) = 0$, since λ_0 is a root of $p(\lambda)$. The second term is the derivative $p'(\lambda)$ of $p(\lambda)$ at λ_0. Since the multiplicity of λ_0 is two, one can write $p(\lambda)$ in the form

$$p(\lambda) = (\lambda - \lambda_0)^2\, p_0(\lambda),$$

where $p_0(\lambda)$ is a polynomial of degree $n - 2$. It follows that

$$p'(\lambda) = 2\,(\lambda - \lambda_0)\, p_0(\lambda) + (\lambda - \lambda_0)^2\, p'_0(\lambda).$$

Hence, it holds $p'(\lambda_0) = 0$, (A.12) vanishes, and $y_{h,2}(x)$ is a solution of (A.8).

$y_{h,1}(x), y_{h,2}(x)$ are linearly independent. One has to show, Lemma A.13, that the Wronski determinant does not vanish. It holds

$$\begin{aligned} W(x) &= \det \begin{pmatrix} y_{h,1}(x) & y_{h,2}(x) \\ y'_{h,1}(x) & y'_{h,2}(x) \end{pmatrix} = \det \begin{pmatrix} e^{\lambda_0 x} & xe^{\lambda_0 x} \\ \lambda_0 e^{\lambda_0 x} & (1 + \lambda_0 x)e^{\lambda_0 x} \end{pmatrix} \\ &= e^{2\lambda_0 x} \det \begin{pmatrix} 1 & x \\ \lambda_0 & 1 + \lambda_0 x \end{pmatrix} = e^{2\lambda_0 x}\,(1 + \lambda_0 x - \lambda_0 x) = e^{2\lambda_0 x} > 0 \end{aligned}$$

for all $x \in I$.

Roots of multiplicity $k > 2$. The principle proof is analogous to the case $k = 2$, where one uses the factorization $p(\lambda) = (\lambda - \lambda_0)^k \, p_0(\lambda)$. The computation of the Wronski determinant becomes more involved. ■

Remark A.19. Complex roots. The statement of Theorem A.18 is true also for complex roots of $p(\lambda)$. The Wronski determinant is $e^{2\lambda_1 x} \neq 0$. However, the corresponding solutions, e.g.

$$\tilde{y}_{1,h}(x) = e^{\lambda_1 x} = e^{(a+ib)x}$$

are complex-valued. Since one has real coefficients in (A.8), one likes to obtain also real-valued solutions. Such solutions can be constructed from the complex-valued solutions.

Let $\lambda_1 = a + ib$, $\overline{\lambda_1} = a - ib$, $a, b \in \mathbb{R}$, $b \neq 0$, be a conjugate complex roots of $p(\lambda)$, then one obtains with Euler's formula

$$\begin{aligned} e^{\lambda_1 x} &= e^{(a+ib)x} = e^{ax} \left(\cos(bx) + i \sin(bx)\right), \\ e^{\overline{\lambda_1} x} &= e^{(a-ib)x} = e^{ax} \left(\cos(bx) - i \sin(bx)\right). \end{aligned}$$

Because of the superposition principle, each linear combination is also solution of (A.8). □

Theorem A.20. Linearly independent solution for simple conjugate complex roots. *Let $\lambda_1 \in \mathbb{C}$, $\lambda_1 = a+ib$, $b \neq 0$, be a simple conjugate complex root of the characteristic polynomial $p(\lambda)$ with real coefficients. Then,*

$$y_{h,1}(x) = Re\left(e^{\lambda_1 x}\right) = e^{ax}\cos(bx), \quad y_{h,2}(x) = Im\left(e^{\lambda_1 x}\right) = e^{ax}\sin(bx),$$

are real-valued, linearly independent solutions of (A.8).

Proof. Use the superposition principle for proving that the functions are solutions and the Wronski determinant for proving that they are linearly independent, exercise. ■

Theorem A.21. Linearly independent solution for conjugate complex roots with multiplicity greater than one. *Let $\lambda_1 \in \mathbb{C}$, $\lambda_1 = a + ib$, $b \neq 0$, be a conjugate complex root with multiplicity k of the characteristic polynomial $p(\lambda)$ with real coefficients. Then,*

$$\begin{aligned} y_{h,1}(x) = e^{ax}\cos(bx), \ldots, y_{h,k}(x) = x^{k-1}e^{ax}\cos(bx), \\ y_{h,k+1}(x) = e^{ax}\sin(bx), \ldots, y_{h,2k}(x) = x^{k-1}e^{ax}\sin(bx) \end{aligned} \tag{A.13}$$

are real-valued, linearly independent solutions of (A.8).

Proof. The proof is similarly to the previous theorems. ■

Theorem A.22. Fundamental system for (A.8)**.** *Let $p(\lambda)$ be the characteristic polynomial of* (A.8) *with the roots $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$, where the roots are counted in correspondence to their multiplicity. Then, the set of solutions of form* (A.11) *and* (A.13) *form a fundamental system of* (A.8).

Proof. A real root with multiplicity k gives k linearly independent solutions and a conjugate complex root with multiplicity k gives $2k$ linearly independent solutions. Thus, the total number of solutions of form (A.11) and (A.13) is equal to the number of roots of $p(\lambda)$. This number is equal to n, because of the fundamental theorem of algebra. It is known from Theorem A.14 that a fundamental system has exactly n functions. Altogether, the correct number of functions is there.

One can show that solutions that correspond to different roots are linearly independent, e.g., (Günther *et al.*, 1974, p. 75). The linearly independence of the solutions that belong to the same root, was already proved. ■

Example A.23. Homogeneous second order linear differential equation with constant coefficients.

1. Consider
$$y''(x) + 6y'(x) + 9y(x) = 0.$$
The characteristic polynomial is
$$p(\lambda) = \lambda^2 + 6\lambda + 9$$
with the roots $\lambda_1 = \lambda_2 = -3$. One obtains the fundamental system
$$y_{h,1}(x) = e^{-3x}, \quad y_{h,2}(x) = xe^{-3x}.$$
The general solution of the homogeneous equation has the form
$$y_h(x) = c_1 y_{h,1}(x) + c_2 y_{h,2}(x) = c_1 e^{-3x} + c_2 x e^{-3x}, \quad c_1, c_2 \in \mathbb{R}.$$

2. Consider
$$y''(x) + 4y(x) = 0 \quad \Longrightarrow \quad p(\lambda) = \lambda^2 + 4 \quad \Longrightarrow \quad \lambda_{1,2} = \pm 2i.$$
It follows that
$$\begin{aligned} y_{h,1}(x) &= \cos(2x), \quad y_{h,2}(x) = \sin(2x) \\ y_h(x) &= c_1 \cos(2x) + c_2 \sin(2x), \quad c_1, c_2 \in \mathbb{R}. \end{aligned}$$

□

A.1.3.2 The Inhomogeneous Equation

Remark A.24. Goal. Because of the superposition principle, a special solution of (A.7) has to be found. This section sketches several possibilities to obtain such a solution. □

Remark A.25. Appropriate ansatz (Störgliedansätze). If the right-hand side $f(x)$ possesses a special form, it is possible to obtain a solution of the inhomogeneous equation (A.7) with an appropriate ansatz. From (A.7) it becomes clear, that this way works only if on the left-hand side and the right-hand

side of the equation are the same types of functions. In particular, one needs the same types of functions for $y_i(x)$ and all derivatives up to order n. This approach works, e.g., for the following classes of right-hand sides:

- $f(x)$ is a polynomial

$$f(x) = b_0 + b_1 x + \ldots + b_m x^m, \quad b_m \neq 0.$$

  The appropriate ansatz is also a polynomial

$$y_i(x) = x^k \left(c_0 + c_1 x + \ldots + c_m x^m\right),$$

  where 0 is a root of $p(\lambda)$ with multiplicity k.
- If the right-hand side is

$$f(x) = \left(b_0 + b_1 x + \ldots + b_m x^m\right) e^{ax},$$

  then one can use the following ansatz

$$y_i(x) = x^k \left(c_0 + c_1 x + \ldots + c_m x^m\right) e^{ax},$$

  where a is a root of $p(\lambda)$ with multiplicity k. The first class of functions is just a special case for $a = 0$.
- For right-hand sides of the form

$$f(x) = \left(b_0 + b_1 x + \ldots + b_m x^m\right) \cos(bx),$$
$$f(x) = \left(b_0 + b_1 x + \ldots + b_m x^m\right) \sin(bx),$$

  one can use the ansatz

$$\begin{aligned} y_i(x) = {} & x^k \left(c_0 + c_1 x + \ldots + c_m x^m\right) \cos(bx) \\ & + x^k \left(d_0 + d_1 x + \ldots + d_m x^m\right) \sin(bx), \end{aligned}$$

  if ib is a root of $p(\lambda)$ with multiplicity k.

One can find the ansatz for more right-hand sides in the literature, e.g. in Heuser (2006). □

Example A.26. Appropriate ansatz (Störgliedansatz). Consider

$$y''(x) - y'(x) + 2y(x) = \cos x.$$

The appropriate ansatz is given by

$$\begin{aligned} y_i(x) &= a \cos x + b \sin x \quad \Longrightarrow \\ y_i'(x) &= -a \sin x + b \cos x \quad \Longrightarrow \\ y_i''(x) &= -a \cos x - b \sin x. \end{aligned}$$

Inserting into the equation gives

$$-a\cos x - b\sin x + a\sin x - b\cos x + 2a\cos x + 2b\sin x = \cos x \implies$$
$$(-a - b + 2a)\cos x + (-b + a + 2b)\sin x = \cos x.$$

The last equation is satisfied if the numbers a, b solve the following linear system of equations

$$a - b = 1,\ a + b = 0 \quad \Longrightarrow \quad a = \frac{1}{2},\ b = -\frac{1}{2}.$$

One obtains the special solution

$$y_i(x) = \frac{1}{2}\left(\cos x - \sin x\right).$$

□

Remark A.27. Variation of the constants. If one cannot find an appropriate ansatz, then one can try the variation of the constants. This approach will be demonstrated for the second order differential equation

$$y''(x) + a_1 y'(x) + a_0 y(x) = f(x). \tag{A.14}$$

Let $y_{h,1}(x), y_{h,2}(x)$ be two linearly independent solutions of the homogeneous differential equation such that

$$y_h(x) = c_1 y_{h,1}(x) + c_2 y_{h,2}(x)$$

is the general solution of the homogeneous equation. Now, one makes the ansatz

$$y_i(x) = c_1(x) y_{h,1}(x) + c_2(x) y_{h,2}(x)$$

with two unknown functions $c_1(x), c_2(x)$. The determination of these functions requires two conditions. One has

$$\begin{aligned} y_i'(x) &= c_1'(x) y_{h,1}(x) + c_1(x) y_{h,1}'(x) + c_2'(x) y_{h,2}(x) + c_2(x) y_{h,2}'(x) \\ &= (c_1'(x) y_{h,1}(x) + c_2'(x) y_{h,2}(x)) + c_1(x) y_{h,1}'(x) + c_2(x) y_{h,2}'(x). \end{aligned}$$

Now, one sets the term in the parentheses zero. This is the first condition. It follows that

$$y_i''(x) = c_1'(x) y_{h,1}'(x) + c_1(x) y_{h,1}''(x) + c_2'(x) y_{h,2}'(x) + c_2(x) y_{h,2}''(x).$$

Inserting this expression into (A.14) gives

$$\begin{aligned} f(x) &= c_1'(x)y_{h,1}'(x) + c_1(x)y_{h,1}''(x) + c_2'(x)y_{h,2}'(x) + c_2(x)y_{h,2}''(x) \\ &\quad + a_1\left(c_1(x)y_{h,1}'(x) + c_2(x)y_{h,2}'(x)\right) + a_0\left(c_1(x)y_{h,1}(x) + c_2(x)y_{h,2}(x)\right) \\ &= c_1(x)\underbrace{\left(y_{h,1}''(x) + a_1y_{h,1}'(x) + a_0y_{h,1}(x)\right)}_{=0} \\ &\quad + c_2(x)\underbrace{\left(y_{h,2}''(x) + a_1y_{h,2}'(x) + a_0y_{h,2}(x)\right)}_{=0} \\ &\quad + c_1'(x)y_{h,1}'(x) + c_2'(x)y_{h,2}'(x). \end{aligned}$$

This is the second condition. Summarizing both conditions gives the following system of equations

$$\begin{pmatrix} y_{h,1}(x) & y_{h,2}(x) \\ y_{h,1}'(x) & y_{h,2}'(x) \end{pmatrix} \begin{pmatrix} c_1'(x) \\ c_2'(x) \end{pmatrix} = \begin{pmatrix} 0 \\ f(x) \end{pmatrix}.$$

This system possesses a unique solution since $y_{h,1}(x), y_{h,2}(x)$ are linearly independent from what follows that the determinant of the system matrix, which is just the Wronksi matrix, is not equal to zero. The solution is

$$c_1'(x) = -\frac{f(x)y_{h,2}(x)}{y_{h,1}(x)y_{h,2}'(x) - y_{h,1}'(x)y_{h,2}(x)},\ c_2'(x) = \frac{f(x)y_{h,1}(x)}{y_{h,1}(x)y_{h,2}'(x) - y_{h,1}'(x)y_{h,2}(x)}.$$

The success of the method of the variation of the constants depends only on the difficulty to find the primitives of $c_1'(x)$ and $c_2'(x)$.

For equations of order higher than two, one has the goal to get a linear system of equations for $c_1'(x), \ldots, c_n'(x)$. To this end, one sets for each derivative of the ansatz the terms with $c_1'(x), \ldots, c_n'(x)$ equal to zero. The obtained linear system of equations has as matrix the Wronski matrix and as right-hand side a vector, whose first $(n-1)$ components are equal to zero and whose last component is $f(x)$. □

Example A.28. Variation of the constants. Find the general solution of

$$y''(x) + 6y'(x) + 9y(x) = \frac{e^{-3x}}{1+x}.$$

The general solution of the homogeneous equation is

$$y_h(x) = c_1e^{-3x} + c_2xe^{-3x},$$

see Example A.23. The variation of the constants leads to the following system of linear equations

$$\begin{pmatrix} e^{-3x} & xe^{-3x} \\ -3e^{-3x} & (1-3x)e^{-3x} \end{pmatrix} \begin{pmatrix} c_1'(x) \\ c_2'(x) \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{e^{-3x}}{1+x} \end{pmatrix}.$$

Using, e.g., the Cramer rule, gives

$$c_1'(x) = -\frac{e^{-6x}\left(\frac{x}{1+x}\right)}{(1-3x+3x)e^{-6x}} = -\frac{x}{1+x},$$
$$c_2'(x) = \frac{e^{-6x}\left(\frac{1}{1+x}\right)}{(1-3x+3x)e^{-6x}} = \frac{1}{1+x}.$$

One obtains

$$c_1(x) = -\int \frac{x}{1+x}\,dx = -\int \frac{1+x}{1+x}\,dx + \int \frac{1}{1+x}\,dx = -x + \ln|1+x|\,,$$
$$c_2(x) = \int \frac{1}{1+x}\,dx = \ln|1+x|\,.$$

Thus, one gets

$$y_i(x) = (-x + \ln|1+x|)\,e^{-3x} + \ln|1+x|\,xe^{-3x}$$

and one obtains for the general solution

$$y(x) = (-x + \ln|1+x| + c_1)\,e^{-3x} + (\ln|1+x| + c_2)\,xe^{-3x}.$$

Inserting this function into the equation proves the correctness of the result. □

A.2 Linear Systems of Ordinary Differential Equations of First Order

A.2.1 Definition, Existence and Uniqueness of a Solution

Definition A.29. Linear system of first order differential equations. In a linear system of ordinary differential equations of first order one tries to find functions $y_1(x), \ldots, y_n(x) \;:\; I \to \mathbb{R}$, $I = [a,b] \subset \mathbb{R}$, that satisfy the system

$$y_i'(x) = \sum_{j=1}^{n} a_{ij}(x)y_j(x) + f_i(x), i = 1, \ldots, n,$$

or in matrix-vector notation

$$\boldsymbol{y}'(x) = A(x)\boldsymbol{y}(x) + \boldsymbol{f}(x) \tag{A.15}$$

with

$$\boldsymbol{y}(x) = \begin{pmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{pmatrix}, \ \boldsymbol{y}'(x) = \begin{pmatrix} y_1'(x) \\ \vdots \\ y_n'(x) \end{pmatrix},$$
$$A(x) = \begin{pmatrix} a_{11}(x) & \cdots & a_{1n}(x) \\ \vdots & \ddots & \vdots \\ a_{n1}(x) & \cdots & a_{nn}(x) \end{pmatrix}, \ \boldsymbol{f}(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix},$$

where $a_{ij}(x), f_i(x) \in C(I)$. If $\boldsymbol{f}(x) \equiv \boldsymbol{0}$, then the system is called homogeneous. □

Theorem A.30. Superposition principle for linear systems. *Consider the linear system of ordinary differential equations* (A.15)*, then the superposition principle holds:*

i) If $\boldsymbol{y}_1(x)$ and $\boldsymbol{y}_2(x)$ are two solutions of the homogeneous systems, then $c_1\boldsymbol{y}_1(x) + c_2\boldsymbol{y}_2(x)$, $c_1, c_2 \in \mathbb{R}$, is a solution of the homogeneous system, too.

ii) If $\boldsymbol{y}_0(x)$ is a solution of the inhomogeneous system and $\boldsymbol{y}_1(x)$ is a solution of the homogeneous system, then $\boldsymbol{y}_0(x) + \boldsymbol{y}_1(x)$ is a solution of the inhomogeneous system.

iii) If $\boldsymbol{y}_1(x)$ and $\boldsymbol{y}_2(x)$ are two solutions of the inhomogeneous system, then $\boldsymbol{y}_1(x) - \boldsymbol{y}_2(x)$ is a solution of the homogeneous system.

Proof. Direct calculations, exercise. ■

Corollary A.31. General solution of the inhomogeneous system.

i) If $\boldsymbol{y}_1(x), \boldsymbol{y}_2(x), \ldots, \boldsymbol{y}_k(x)$ are solutions of the homogeneous system, then any linear combination $\sum_{i=1}^{k} c_i\boldsymbol{y}_i(x)$, $c_1, \ldots, c_k \in \mathbb{R}$, is also a solution of the homogeneous system.

ii) The general solution of the inhomogeneous system is the sum of a special solution of the inhomogeneous system and the general solution of the homogeneous system.

Theorem A.32. Existence and uniqueness of a solution of the initial value problem. *Let $I = [x_0 - a, x_0 + a]$ and $a_{ij} \in C(I)$, $f_i \in C(I)$, $i, j = 1, \ldots, n$. Then, there is exactly one solution $\boldsymbol{y}(x) : I \to \mathbb{R}^n$ of the initial value problem to* (A.15) *with the initial value $\boldsymbol{y}(x_0) = \boldsymbol{y}_0 \in \mathbb{R}^n$.*

Proof. The statement of the theorem follows from the theorem on global existence and uniqueness of a solution of an initial value problem from Picard–Lindelöf, see lecture notes Numerical Mathematics I or the literature.

Since the functions $a_{ij}(x)$ are continuous on the closed (compact) interval I, they are also bounded due to the Weierstrass theorem. That means, there is a constant M with

$$|a_{ij}(x)| \leq M, \quad x \in I, \ i, j = 1, \ldots, n.$$

Denoting the right hand side of (A.15) by $\boldsymbol{f}(x, \boldsymbol{y})$, it follows that

$$
\begin{aligned}
\|\boldsymbol{f}(x,\boldsymbol{y}_1)-\boldsymbol{f}(x,\boldsymbol{y}_2)\|_\infty &= \max_{i=1,\dots,n} |f_i(x,\boldsymbol{y}_1)-f_i(x,\boldsymbol{y}_2)| \\
&= \max_{i=1,\dots,n} \left| \sum_{j=1}^n a_{ij}(x)y_{1,j}(x)+f_i(x)-\sum_{j=1}^n a_{ij}(x)y_{2,j}(x)-f_i(x) \right| \\
&= \max_{i=1,\dots,n} \left| \sum_{j=1}^n a_{ij}(x)\left(y_{1,j}(x)-y_{2,j}(x)\right) \right| \\
&\le n \max_{i,j=1,\dots,n} |a_{ij}(x)| \max_{i=1,\dots,n} |y_{1,i}(x)-y_{2,i}(x)| \\
&\le nM \|\boldsymbol{y}_1-\boldsymbol{y}_2\|_\infty ,
\end{aligned}
$$

i.e. the right hand side satisfies a uniform Lipschitz condition with respect to $\boldsymbol{y}$ with the Lipschitz constant nM. Hence, the assumptions of the theorem on global existence and uniqueness of a solution of an initial value problem from Picard–Lindelöf are satisfied. ■

A.2.2 Solution of the Homogeneous System

Remark A.33. Scalar case. Because of the superposition principle, one needs the general solution of the homogeneous system

$$\boldsymbol{y}'(x) = A(x)\boldsymbol{y}(x) \tag{A.16}$$

for finding the general solution of (A.15). The homogeneous system has always the trivial solution $\boldsymbol{y}(x)=\boldsymbol{0}$.

In the scalar case $y'(x)=a(x)y(x)$, the general solution has the form

$$y(x) = c\exp\left(\int_{x_0}^x a(t)\ dt\right), \quad c\in\mathbb{R}, x_0\in(a,b),$$

see lecture notes Numerical Mathematics I or the literature. Also for the system (A.16), it is possible to specify the general solution with the help of the exponential. □

Theorem A.34. General solution of the homogeneous linear system of first order. *The general solution of* (A.16) *is*

$$\boldsymbol{y}_h(x) = e^{\int_{x_0}^x A(t)\ dt}\boldsymbol{c}, \quad \boldsymbol{c}\in\mathbb{R}^n, x_0\in(a,b). \tag{A.17}$$

The integral is defined component-wise.

Proof. i) (A.17) *is a solution of* (A.16). This statement follows from the derivative of the matrix exponential and the rule on the differentiation of an integral with respect to the upper limit

$$\boldsymbol{y}_h'(x) = \frac{d}{dx}\left(e^{\int_{x_0}^x A(t)\ dt}\boldsymbol{c}\right) = \frac{d}{dx}\left(\int_{x_0}^x A(t)\ dt\right) e^{\int_{x_0}^x A(t)\ dt}\boldsymbol{c} = A(x)e^{\int_{x_0}^x A(t)\ dt}\boldsymbol{c}.$$

ii) every solution of (A.16) *is of form* (A.17). Consider an arbitrary solution $\tilde{\boldsymbol{y}}_h(x)$ of (A.16) with $\tilde{\boldsymbol{y}}_h(x_0) \in \mathbb{R}^n$. Take in (A.17) $\boldsymbol{c} = \tilde{\boldsymbol{y}}_h(x_0)$. Then, it follows that

$$\boldsymbol{y}_h(x_0) = e^{\int_{x_0}^{x_0} A(t)\ dt} \tilde{\boldsymbol{y}}_h(x_0) = \underbrace{e^0}_{=I} \tilde{\boldsymbol{y}}_h(x_0) = \tilde{\boldsymbol{y}}_h(x_0).$$

That means, $e^{\int_{x_0}^{x} A(t)\ dt} \tilde{\boldsymbol{y}}_h(x_0)$ is a solution of (A.16) which has in x_0 the same initial value as $\tilde{\boldsymbol{y}}_h(x)$. Since the solution of the initial value problem is unique, Theorem A.32, it follows that $\tilde{\boldsymbol{y}}_h(x) = e^{\int_{x_0}^{x} A(t)\ dt} \tilde{\boldsymbol{y}}_h(x_0)$. ■

A.2.3 Linear Systems of First Order with Constant Coefficients

Remark A.35. Linear system of first order differential equations with constant coefficients. A linear system of first order differential equations with constant coefficients has the form

$$\boldsymbol{y}'(x) = A\boldsymbol{y}(x) + \boldsymbol{f}(x), \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n \times n}. \tag{A.18}$$

Thus, the homogeneous system has the form

$$\boldsymbol{y}'(x) = A\boldsymbol{y}(x). \tag{A.19}$$

Its general solution is given by

$$\boldsymbol{y}_h(x) = e^{Ax}\boldsymbol{c}, \quad \boldsymbol{c} \in \mathbb{R}^n, \tag{A.20}$$

see Theorem A.34. □

Remark A.36. Elimination method, substitution method for the homogeneous system. One needs, due to the superposition principle, the general solution of the homogeneous system. In practice, it is generally hard to compute $\exp(Ax)$ because it is defined by an infinity series. For small systems, i.e. $n \leq 3, 4$, one can use the elimination or substitution method for computing the general solution of (A.19). This method is already known from the numerical solution of linear systems of equations. One solves one equation for a certain unknown function $y_i(x)$ and inserts the result into the other equations. For differential equations, the equation has to be differentiated, see Example A.37. This step reduces the dimension of the system by one. One continues with this method until one reaches an equation with only one unknown function. For this function, a homogeneous linear differential equation of order n has to be solved, see Section A.1.3. The other components of the solution vector of (A.19) can be obtained by back substitution. □

Example A.37. Elimination method, substitution method. Find the solution of

$$\boldsymbol{y}'(x) = \begin{pmatrix} -3 & -1 \\ 1 & -1 \end{pmatrix} \boldsymbol{y}(x) \iff y_1'(x) = -3y_1(x) - y_2(x),\ y_2'(x) = y_1(x) - y_2(x).$$

Solving the second equation for $y_1(x)$ and differentiating gives

$$y_1(x) = y_2'(x) + y_2(x), \quad y_1'(x) = y_2''(x) + y_2'(x).$$

Inserting into the first equation yields

$$y_2''(x) + y_2'(x) = -3\,(y_2'(x) + y_2(x)) - y_2(x) \iff y_2''(x) + 4y_2'(x) + 4y_2(x) = 0.$$

The general solution of this equation is

$$y_2(x) = c_1 e^{-2x} + c_2 x e^{-2x}, \quad c_1, c_2 \in \mathbb{R}.$$

One obtains from the second equation

$$y_1(x) = y_2'(x) + y_2(x) = (-c_1 + c_2)\, e^{-2x} - c_2 x e^{-2x}.$$

Thus, the general solution of the given linear system of differential equations with constant coefficients is computed by

$$\boldsymbol{y} = \begin{pmatrix} -c_1 + c_2 \\ c_1 \end{pmatrix} e^{-2x} + \begin{pmatrix} -c_2 \\ c_2 \end{pmatrix} x e^{-2x}.$$

Note that one can choose the constants in $y_2(x)$, but the constants in $y_1(x)$ are determined by the back substitution. If the constants should be choosen by $y_1(x)$, one obtains

$$\boldsymbol{y} = \begin{pmatrix} C_1 \\ C_2 - C_1 \end{pmatrix} e^{-2x} + \begin{pmatrix} C_2 \\ -C_2 \end{pmatrix} x e^{-2x}.$$

If an initial condition is given, then corresponding constants can be determined. □

Remark A.38. Other methods for computing the general solution of the homogeneous system. There are also other methods for computing the general solution of (A.19).

- The idea of the method of main-vectors and eigenvectors consists in transforming the system to a triangular system. Then it is possible to solve the equations successively. To this end, one constructs with the so-called main-vectors and eigenvectors an invertible matrix $C \in \mathbb{R}^{n\times n}$ such that $C^{-1}AC$ is a triangular matrix. One can show that such a matrix C exists for each $A \in \mathbb{R}^{n\times n}$. Then, one sets

$$\boldsymbol{y}(x) = C\boldsymbol{z}(x) \quad \Longrightarrow \quad \boldsymbol{y}'(x) = C\boldsymbol{z}'(x).$$

Inserting into (A.19) yields

$$C\boldsymbol{z}'(x) = AC\boldsymbol{z}(x) \quad \Longleftrightarrow \quad \boldsymbol{z}'(x) = C^{-1}AC\boldsymbol{z}(x).$$

This is a triangular system for $\boldsymbol{z}(x)$, which is solved successively for the components of $\boldsymbol{z}(x)$. The solution of (A.19) is obtained by computing $C\boldsymbol{z}(x)$.

- The method of matrix functions is based on an appropriate ansatz for the solution.

However, the application of both methods becomes very time-consuming for larger n, see the literature. □

Remark A.39. Methods for determining a special solution of the inhomogeneous system. For computing the general solution of the inhomogeneous system of linear differential equations of first order with constant coefficients, one needs also a special solution of the inhomogeneous system. There are several possibilities for obtaining this solution:

- *Method of the variation of constants.* One replaces $\boldsymbol{c}$ in (A.20) by $\boldsymbol{c}(x)$, inserts this expression into (A.18), obtains conditions for $\boldsymbol{c}'(x)$, and tries to compute $\boldsymbol{c}(x)$ from these conditions.
- *Appropriate ansatz (Störgliedansätze).* If each component of the right hand side $\boldsymbol{f}(x)$ has a special form, e.g., a polynomial, sine, cosine, or exponential, then it is often possible to find the special solution with an appropriate ansatz.
- *Method of elimination.* If the right hand side of $\boldsymbol{f}(x)$ of (A.18) is $(n-1)$ times continuously differentiable, then one can proceed exactly as in the elimination method. One obtains for one component of $\boldsymbol{y}(x)$ an inhomogeneous ordinary differential equation of order n with constant coefficients, for which one has to find a special solution. A special solution for (A.18) is obtained by back substitution.

□

References

Butcher, J. C. (1964) Implicit Runge-Kutta processes. *Math. Comp.*, **18**, 50–64.

Cryer, C. W. (1972) On the instability of high order backward-difference multistep methods. *Nordisk Tidskr. Informations behandling (BIT)*, **12**, 17–25.

Curtiss, C. F. & Hirschfelder, J. O. (1952) Integration of stiff equations. *Proc. Nat. Acad. Sci. U. S. A.*, **38**, 235–243.

Deuflhard, P. & Bornemann, F. (2002) *Scientific computing with ordinary differential equations*. Texts in Applied Mathematics, vol. 42. Springer-Verlag, New York, pp. xx+485. Translated from the 1994 German original by Werner C. Rheinboldt.

Deuflhard, P. & Bornemann, F. (2008) *Numerische Mathematik 2*. de Gruyter Lehrbuch. [de Gruyter Textbook], revised edn. Walter de Gruyter & Co., Berlin, pp. xii+499. Gewöhnliche Differentialgleichungen. [Ordinary differential equations].

Dormand, J. R. & Prince, P. J. (1980) A family of embedded Runge-Kutta formulae. *J. Comput. Appl. Math.*, **6**, 19–26.

Fehlberg, E. (1964) New high-order Runge-Kutta formulas with step size control for systems of first- and second-order differential equations. *Z. Angew. Math. Mech.*, **44**, T17–T29.

Fehlberg, E. (1969) Klassische Runge-Kutta-Formeln fünfter und siebenter Ordnung mit Schrittweiten-Kontrolle. *Computing (Arch. Elektron. Rechnen)*, **4**, 93–106.

Günther, P., Beyer, K., Gottwald, S. & Wünsch, V. (1974) *Grundkurs Analysis. Teil 4*. Leipzig: BSB B. G. Teubner Verlagsgesellschaft, p. 308. Erste Auflage, Mathematisch-Naturwissenschaftliche Bibliothek, Band 56.

Hairer, E., Nørsett, S. P. & Wanner, G. (1993) *Solving ordinary differential equations. I*. Springer Series in Computational Mathematics, vol. 8, second edn. Berlin: Springer-Verlag, pp. xvi+528. Nonstiff problems.

Heuser, H. (2006) *Gewöhnliche Differentialgleichungen*. Mathematische Leitfäden. [Mathematical Textbooks], fifth edn. Stuttgart: B. G. Teubner, p. 628. Einführung in Lehre und Gebrauch. [Introduction to theory and application].

John, V. & Rang, J. (2010) Adaptive time step control for the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, **199**, 514–524.

Kunkel, P. & Mehrmann, V. (2006) *Differential-algebraic equations*. EMS Textbooks in Mathematics. European Mathematical Society (EMS), Zürich, pp. viii+377. Analysis and numerical solution.

Prince, P. J. & Dormand, J. R. (1981) High order embedded Runge-Kutta formulae. *J. Comput. Appl. Math.*, **7**, 67–75.

Shampine, L. F. & Reichelt, M. W. (1997) The MATLAB ODE suite. *SIAM J. Sci. Comput.*, **18**, 1–22. Dedicated to C. William Gear on the occasion of his 60th birthday.

Söderlind, G. (2002) Automatic control and adaptive time-stepping. *Numer. Algorithms*, **31**, 281–310. Numerical methods for ordinary differential equations (Auckland, 2001).

Strehmel, K., Weiner, R. & Podhaisky, H. (2012) *Numerik gewöhnlicher Differentialgleichungen*, second edn. Springer Spektrum, p. 505.

Strehmel, K. & Weiner, R. (1995) *Numerik gewöhnlicher Differentialgleichungen*. Teubner Studienbücher Mathematik. [Teubner Mathematical Textbooks]. Stuttgart: B. G. Teubner, p. 462.