Chapter 2

Numerical Methods for Stiff Ordinary Differential Equations

2.1 Stiff Ordinary Differential Equations

Remark 2.1 *Stiffness.* It was observed in Curtiss and Hirschfelder (1952) that explicit methods failed for the numerical solution of ordinary differential equations that model certain chemical reactions. They introduced the notation stiffness for such chemical reactions where the fastly 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 2.5. 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 and 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 continous 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 2.66. □

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

$$\begin{aligned} y_1' &= -80.6y_1 + 119.4y_2 \\ y_2' &= 79.6y_1 - 120.4y_2 \end{aligned}$$

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

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

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

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

It 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 □

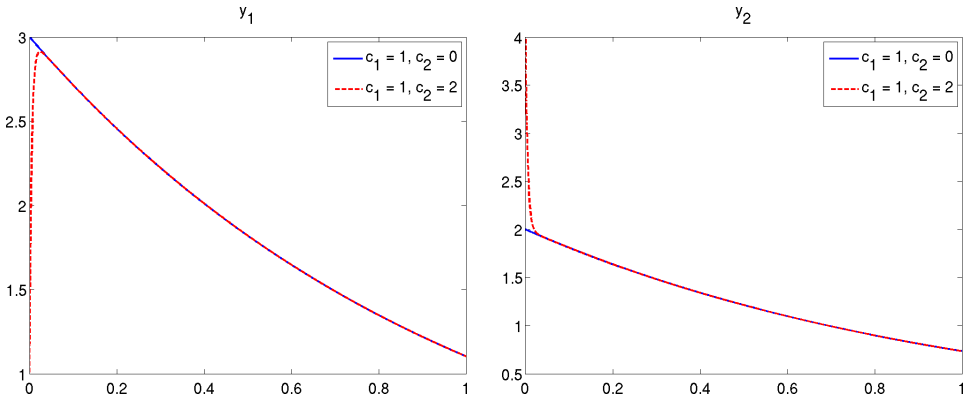

Figure 2.1: Solutions of Example 2.2, left: first component, right: second component.

2.2 Ordinary Differential Equations of Higher Order

Remark 2.3 *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 2.41. □

2.2.1 Definition, Connection to First Order Systems

Definition 2.4 General and explicit n-th order ordinary differential equation. The general ordinary differential equation of order n has the form

$$F\left(x, y(x), y'(x), \ldots, y^{(n)}(x)\right) = 0. \tag{2.2}$$

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

$$y^{(n)}(x) = f\left(x, y(x), y'(x), \ldots, y^{(n-1)}(x)\right). \tag{2.3}$$

The function $y(x)$ is a solution of (2.2) in an interval I if $y(x)$ is n times continuously differentiable in I and if $y(x)$ satisfies (2.2).

Let $x_0 \in I$ be given. Then, (2.2) 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 (2.2). □

Example 2.5 *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 2.6 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* (2.3) *can be transformed equivalently to a system of n differential equations of first order*

$$\begin{aligned}y_k'(x) &= y_{k+1}(x), \quad k = 1, \ldots, n-1, \\ y_n'(x) &= f(x, y_1(x), \ldots, y_n(x))\end{aligned} \tag{2.4}$$

or (note that the system is generally nonlinear, since the unknown functions appear also in $f(\cdot,\ldots,\cdot)$)

$$\mathbf{y}'(x) = \begin{pmatrix} y_1'(x) \\ y_2'(x) \\ \vdots \\ y_n'(x) \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ f(x, y_1, \ldots, y_n) \end{pmatrix}$$

for the n functions $y_1(x),\ldots,y_n(x)$. The solution of (2.3) is $y(x) = y_1(x)$.

Proof: Insert in (2.3)

$$\begin{aligned} y_1(x) &:= y(x), \quad y_2(x) := y_1'(x) = y'(x), \quad y_3(x) := y_2'(x) = y''(x), \quad \ldots \\ y_n(x) &:= y_{n-1}'(x) = y^{(n-1)}(x). \end{aligned}$$

If $y \in C^n(I)$ is a solution of (2.3), then $y_1(x),\ldots,y_n(x)$ is obviously a solution of (2.4) in I.

Conversely, if $y_1(x),\ldots,y_n(x) \in C^1(I)$ is a solution of (2.4), then it holds

$$\begin{aligned} y_2(x) &= y_1'(x), \quad y_3(x) = y_2'(x) = y_1''(x), \ldots, y_n(x) = y_1^{(n-1)}(x) \\ y_n'(x) &= y_1^{(n)}(x) = f(x, y_1, \ldots, y_n). \end{aligned}$$

Hence, the function $y_1(x)$ is n times continuously differentiable and it is the solution of (2.3) in I. ■

Example 2.7 *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

$$\begin{aligned} y_1(x) &= y(x) \\ y_1'(x) &= y_2(x) (= y'(x)) \\ y_2'(x) &= y_3(x) (= y''(x)) \\ y_3'(x) &= y'''(x) = -2y''(x) + 5y'(x) + f(x, y(x)) \\ &= -2y_3(x) + 5y_2(x) + f(x, y_1(x)). \end{aligned}$$

□

2.2.2 Linear Differential Equations of n-th Order

Definition 2.8 Linear n-th order differential equations. A linear differential equation of n-th order has the form

$$a_n(x)y^{(n)}(x) + a_{n-1}(x)y^{(n-1)}(x) + \ldots + a_1(x)y'(x) + a_0(x)y(x) = f(x), \qquad (2.5)$$

where the functions $a_0(x),\ldots,a_n(x)$ are continuous in the interval I, in which a solution of (2.5) 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 $x \in I$

$$a_n(x)y^{(n)}(x) + a_{n-1}(x)y^{(n-1)}(x) + \ldots + a_1(x)y'(x) + a_0(x)y(x) = 0. \qquad (2.6)$$

□

Theorem 2.9 Superposition principle for linear differential equations of higher order. *Consider the linear differential equation of n-th order* (2.5)*, then the superposition principle holds:*

i) If $y_1(x)$ *and* $y_2(x)$ *are two solutions of the homogeneous equation* (2.6)*, 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. ■

Cororllary 2.10 General solution of the inhomogeneous differential equation. *The general solution of* (2.5) *is the sum of the general solution of the homogeneous linear differential equation of n-th order* (2.6) *and one special solution of the inhomogeneous n-th order differential equation* (2.5)*.*

Remark 2.11 *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

$$\begin{aligned} \mathbf{y}'(x) &= \begin{pmatrix} y_1'(x) \\ y_2'(x) \\ \vdots \\ y_n'(x) \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{a_0(x)}{a_n(x)} & -\frac{a_1(x)}{a_n(x)} & -\frac{a_2(x)}{a_n(x)} & \cdots & -\frac{a_{n-1}(x)}{a_n(x)} \end{pmatrix} \begin{pmatrix} y_1(x) \\ y_2(x) \\ \vdots \\ y_n(x) \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \frac{f(x)}{a_n(x)} \end{pmatrix} \\ &=: A(x)\mathbf{y}(x) + \mathbf{f}(x). \end{aligned} \tag{2.7}$$

□

Theorem 2.12 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* (2.5) *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 (2.5) is equivalent to the system (2.7), 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 (2.7) with respect to $y_1, \ldots, y_n$. Denoting the right-hand side by $F(x, \mathbf{y})$ gives

$$\|\mathbf{F}(x, \mathbf{y}) - \mathbf{F}(x, \tilde{\mathbf{y}})\|_{[C(I)]^n} = \|A(\mathbf{y} - \tilde{\mathbf{y}})\|_{[C(I)]^n} \leq \|A\|_{[C(I)]^n, \infty} \|\mathbf{y} - \tilde{\mathbf{y}}\|_{[C(I)]^n},$$

where one uses the triangle inequality to get

$$\begin{aligned}\|A_{i\cdot}\mathbf{y}\|_{C(I)} &= \max_{x\in I}\left|\sum_{j=1}^{n} a_{ij}(x)y_j(x)\right| \le \max_{x\in I}\sum_{j=1}^{n}|a_{ij}(x)|\max_{j=1,\ldots,n}\left\{\max_{x\in I}|y_j(x)|\right\}\\ &= \|A_{i\cdot}\|_{C(I)}\,\|\mathbf{y}\|_{[C(I)]^n}\end{aligned}$$

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 2.13 Linearly independent solutions, fundamental system. The solutions $y_i(x) \;:\; I \to \mathbb{R}$, $i = 1, \ldots, k$, of (2.6) 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 (2.6). □

Definition 2.14 Wronski[1] matrix, Wronski determinant. Let $y_i(x)$, $i = 1, \ldots, k$, be solutions of (2.6). 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 2.15 Properties of the Wronski matrix and Wronski determinant. *Let $I = [a, b]$ and let $y_1(x), \ldots, y_n(x)$ be solutions of (2.6).*

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 (2.6), e.g. $y_1(x), \ldots, y_k(x)$, linearly independent.

Proof:

[1] Joseph Marie Wronski (1758 – 1853)

i) Let S_n be the set of all permutations of $\{1,\dots,n\}$ and let $\sigma \in S_n$. Denote the entries of the Wronski matrix by $\mathcal{W}(x) = (y_{jk}(x))_{j,k=1}^n$. If $\sigma = (\sigma_1,\dots,\sigma_n)$, then let

$$\prod_{j=1}^{n} y_{j,\sigma_j}(x) = \left(y_{1,\sigma_1} y_{2,\sigma_2} \dots y_{n,\sigma_n}\right)(x).$$

Applying the Laplace[2] formula for determinants and the product rule yields

$$\begin{aligned}
\frac{d}{dx}\det(\mathcal{W}(x)) &= \frac{d}{dx}\left(\sum_{\sigma\in S_n}\left(\operatorname{sgn}(\sigma)\prod_{j=1}^{n} y_{j,\sigma_j}(x)\right)\right)\\
&= \sum_{\sigma\in S_n}\left(\operatorname{sgn}(\sigma)\sum_{i=1}^{n}\left(\prod_{j=1,j\neq i}^{n} y_{j,\sigma_j}(x)\right) y'_{i,\sigma_i}(x)\right)\\
&= \sum_{i=1}^{n}\left(\sum_{\sigma\in S_n}\left(\operatorname{sgn}(\sigma)\prod_{j=1,j\neq i}^{n} y_{j,\sigma_j}(x) y'_{i,\sigma_i}(x)\right)\right)\\
&= \sum_{i=1}^{n}\det\begin{pmatrix} \dots & \dots & \dots \\ \left(y_1^{(i-1)}(x)\right)' & \dots & \left(y_n^{(i-1)}(x)\right)' \\ \dots & \dots & \dots \end{pmatrix}.
\end{aligned}$$

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 $(y_1(x),\dots,y_n(x))$. The rows with dots in this matrix coincide with the respective rows of $\mathcal{W}(x)$. For $i = 1,\dots,n-1$, the determinants vanish, since in these cases there are two identical rows, namely row i and $i+1$. Thus, it is

$$\frac{d}{dx}\det(\mathcal{W}(x)) = \det\begin{pmatrix} y_1(x) & \dots & y_n(x) \\ y'_1(x) & \dots & y'_n(x) \\ & \vdots & \\ y_1^{(n-2)}(x) & \dots & y_n^{(n-2)}(x) \\ y_1^{(n)}(x) & \dots & y_n^{(n)}(x) \end{pmatrix}.$$

Now, one uses that $y_1(x),\dots,y_n(x)$ are solutions of (2.6) and one replaces the n-th derivative in the last row by (2.6). Using rules for the evaluation of determinants, one obtains

$$\frac{d}{dx}\det(\mathcal{W}(x)) = \sum_{i=1}^{n} -\frac{a_{i-1}(x)}{a_n(x)}\det\begin{pmatrix} y_1(x) & \dots & y_n(x) \\ y'_1(x) & \dots & y'_n(x) \\ & \vdots & \\ y_1^{(i-1)}(x) & \dots & y_n^{(i-1)}(x) \end{pmatrix}.$$

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 2.16 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* (2.6) *has a fundamental system in I. Each solution of* (2.6) *can be written as a linear combination of the solutions of an arbitrary fundamental system.*

[2] Pierre–Simon (Marquis de) Laplace (1749 – 1827)

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 2.12. It is $W(x_0) = 1$ for these solutions. From Lemma 2.15, iii), it follows that $\{y_1(x),\dots,y_n(x)\}$ is a fundamental system.

Let $y(x)$ be an arbitrary solution of (2.6) 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 (2.6). Since the solution of the initial value problem to (2.6) is unique, Theorem 2.12, it follows that $y(x) = \sum_{i=1}^n c_{i-1}y_i(x)$. ■

Theorem 2.17 Special solution of the inhomogeneous equation. *Let $\{y_1(x),\dots,y_n(x)\}$ be a fundamental system of the homogeneous equation (2.6) 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 (2.5).

Proof: The proof uses the principle of the variation of the constants. This principle will be explained in a simpler setting in Remark 2.29. For details of the proof, see the literature. ■

2.2.3 Linear n-th Order Differential Equations with Constant Coefficients

Definition 2.18 Linear differential equation of n-th order with constant coefficients. A linear n-th order differential equation with constant coefficients has the form

$$a_n y^{(n)}(x) + a_{n-1}y^{(n-1)}(x) + \dots + a_1 y'(x) + a_0 y(x) = f(x), \tag{2.8}$$

with $a_i \in \mathbb{R}$, $i = 0,\dots,n$, $a_n \neq 0$. □

The Homogeneous Equation

Remark 2.19 *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

$$\sum_{i=0}^n a_i y_h^{(i)}(x) = 0. \tag{2.9}$$

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 (2.9)

$$y_h(x) = e^{\lambda x}, \quad \lambda \in \mathbb{C}. \tag{2.10}$$

It follows that

$$y_h'(x) = \lambda e^{\lambda x}, \ldots, y_h^{(n)}(x) = \lambda^n e^{\lambda x}.$$

Inserting into (2.9) gives

$$\left(a_n \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1 \lambda + a_0\right) e^{\lambda x} = 0. \tag{2.11}$$

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

$$e^{\lambda x} = e^{ax}\left(\cos(bx) + i \sin(bx)\right) = e^{ax}\cos(bx) + ie^{ax}\sin(bx).$$

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 (2.11) 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 (2.9). 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 (2.10) is not sufficient in the case of multiple roots. □

Theorem 2.20 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* (2.9)

$$y_{h,1}(x) = e^{\lambda_0 x}, \; y_{h,2}(x) = xe^{\lambda_0 x}, \; \ldots, \; y_{h,k}(x) = x^{k-1}e^{\lambda_0 x}. \tag{2.12}$$

Proof: For $k = 2$.

$y_{h,1}(x), y_{h,2}(x)$ *solve* (2.9). This statement is already clear for $y_{h,1}(x)$ since this function has the form of the ansatz (2.10). For $y_{h,2}(x)$ it holds

$$\begin{aligned}
y_{h,2}'(x) &= (1 + \lambda_0 x)\, e^{\lambda_0 x}, \\
y_{h,2}''(x) &= \left(2\lambda_0 + \lambda_0^2 x\right) e^{\lambda_0 x}, \\
&\vdots \\
y_{h,2}^{(n)}(x) &= \left(n\lambda_0^{n-1} + \lambda_0^n x\right) e^{\lambda_0 x}.
\end{aligned}$$

Inserting into the left-hand side of (2.9) yields

$$e^{\lambda_0 x}\sum_{i=0}^{n} a_i(i\lambda_0^{i-1}+\lambda_0^i x) = e^{\lambda_0 x}\left(x\underbrace{\sum_{i=0}^{n} a_i\lambda_0^i}_{p(\lambda_0)} + \underbrace{\sum_{i=0}^{n} a_i i\lambda_0^i}_{p'(\lambda_0)}\right). \tag{2.13}$$

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 $p'(\lambda_0)=0$, (2.13) vanishes, and $y_{h,2}(x)$ is a solution of (2.9).

$y_{h,1}(x), y_{h,2}(x)$ *are linearly independent.* One has to show, Lemma 2.15, 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 2.21 *Complex roots.* The statement of Theorem 2.20 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 (2.9), 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 (2.9). □

Theorem 2.22 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 (2.9).

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 2.23 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,*

$$\begin{aligned} y_{h,1}(x) &= e^{ax}\cos(bx), \ldots, y_{h,k}(x) = x^{k-1}e^{ax}\cos(bx), \\ y_{h,k+1}(x) &= e^{ax}\sin(bx), \ldots, y_{h,2k}(x) = x^{k-1}e^{ax}\sin(bx) \end{aligned} \tag{2.14}$$

are real-valued, linearly independent solutions of (2.9).

Proof: The proof is similarly to the previous theorems. ■

Theorem 2.24 Fundamental system for (2.9)**.** *Let $p(\lambda)$ be the characteristic polynomial of* (2.9) *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* (2.12) *and* (2.14) *form a fundamental system of* (2.9).

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 (2.12) and (2.14) 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 2.16 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 2.25 *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
$$\begin{aligned} y_{h,1}(x) &= \cos(2x), \quad y_{h,2}(x) = \sin(2x) \\ y_h(x) &= c_1\cos(2x) + c_2\sin(2x), \quad c_1, c_2 \in \mathbb{R}. \end{aligned}$$

□

The Inhomogeneous Equation

Remark 2.26 *Goal.* Because of the superposition principle, a special solution of (2.8) has to be found. This section sketches several possibilities to obtain such a solution. □

Remark 2.27 *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 (2.8) with an appropriate ansatz. From (2.8) 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

  $$\begin{aligned} f(x) &= \left(b_0 + b_1 x + \ldots + b_m x^m\right) \cos(bx), \\ f(x) &= \left(b_0 + b_1 x + \ldots + b_m x^m\right) \sin(bx), \end{aligned}$$

  one can use the ansatz

  $$\begin{aligned} y_i(x) &= x^k \left(c_0 + c_1 x + \ldots + c_m x^m\right) \cos(bx) \\ &\quad + x^k \left(d_0 + d_1 x + \ldots + d_m x^m\right) \sin(bx), \end{aligned}$$

  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 2.28 *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

$$\begin{aligned} -a\cos x - b\sin x + a\sin x - b\cos x + 2a\cos x + 2b\sin x &= \cos x \implies \\ (-a - b + 2a)\cos x + (-b + a + 2b)\sin x &= \cos x. \end{aligned}$$

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 2.29 *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

$$y''(x) + a_1 y'(x) + a_0 y(x) = f(x). \tag{2.15}$$

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 (2.15) gives

$$\begin{aligned} f(x) &= c_1'(x) y_{h,1}'(x) + c_1(x) y_{h,1}''(x) + c_2'(x) y_{h,2}'(x) + c_2(x) y_{h,2}''(x) \\ &\quad + a_1\left(c_1(x) y_{h,1}'(x) + c_2(x) y_{h,2}'(x)\right) + a_0\left(c_1(x) y_{h,1}(x) + c_2(x) y_{h,2}(x)\right) \\ &= c_1(x) \underbrace{\left(y_{h,1}''(x) + a_1 y_{h,1}'(x) + a_0 y_{h,1}(x)\right)}_{=0} \\ &\quad + c_2(x) \underbrace{\left(y_{h,2}''(x) + a_1 y_{h,2}'(x) + a_0 y_{h,2}(x)\right)}_{=0} \\ &\quad + c_1'(x) y_{h,1}'(x) + c_2'(x) y_{h,2}'(x). \end{aligned}$$

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 2.30 *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

$$y_h(x) = c_1 e^{-3x} + c_2 x e^{-3x},$$

see Example 2.25. 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

$$\begin{aligned} c_1'(x) &= -\frac{e^{-6x}\left(\frac{x}{1+x}\right)}{(1-3x+3x)e^{-6x}} = -\frac{x}{1+x}, \\ c_2'(x) &= \frac{e^{-6x}\left(\frac{1}{1+x}\right)}{(1-3x+3x)e^{-6x}} = \frac{1}{1+x}. \end{aligned}$$

One obtains

$$\begin{aligned} c_1(x) &= -\int \frac{x}{1+x}\,dx = -\int \frac{1+x}{1+x}\,dx + \int \frac{1}{1+x}\,dx = -x + \ln|1+x|, \\ c_2(x) &= \int \frac{1}{1+x}\,dx = \ln|1+x|. \end{aligned}$$

Thus, one gets

$$y_i(x) = (-x + \ln|1+x|)\, e^{-3x} + \ln|1+x|\, x e^{-3x}$$

and one obtains for the general solution

$$y(x) = (-x + \ln|1+x| + c_1)\, e^{-3x} + (\ln|1+x| + c_2)\, x e^{-3x}.$$

Inserting this function into the equation proves the correctness of the result. □

2.3 Linear Systems of Ordinary Differential Equations of First Order

2.3.1 Definition, Existence and Uniqueness of a Solution

Definition 2.31 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

$$\mathbf{y}'(x) = A(x)\mathbf{y}(x) + \mathbf{f}(x) \tag{2.16}$$

with

$$\begin{aligned} \mathbf{y}(x) &= \begin{pmatrix} y_1(x) \\ \vdots \\ y_n(x) \end{pmatrix}, \ \mathbf{y}'(x) = \begin{pmatrix} y_1'(x) \\ \vdots \\ y_n'(x) \end{pmatrix}, \\ A(x) &= \begin{pmatrix} a_{11}(x) & \cdots & a_{1n}(x) \\ \vdots & \ddots & \vdots \\ a_{n1}(x) & \cdots & a_{nn}(x) \end{pmatrix}, \ \mathbf{f}(x) = \begin{pmatrix} f_1(x) \\ \vdots \\ f_n(x) \end{pmatrix}, \end{aligned}$$

where $a_{ij}(x), f_i(x) \in C(I)$. If $\mathbf{f}(x) \equiv \mathbf{0}$, then the system is called homogeneous. □

Theorem 2.32 Superposition principle for linear systems. *Consider the linear system of ordinary differential equations* (2.16)*, then the superposition principle holds:*

i) If $\mathbf{y}_1(x)$ *and* $\mathbf{y}_2(x)$ *are two solutions of the homogeneous systems, then* $c_1\mathbf{y}_1(x) + c_2\mathbf{y}_2(x)$, $c_1, c_2 \in \mathbb{R}$, *is a solution of the homogeneous system, too.*

ii) If $\mathbf{y}_0(x)$ *is a solution of the inhomogeneous system and* $\mathbf{y}_1(x)$ *is a solution of the homogeneous system, then* $\mathbf{y}_0(x) + \mathbf{y}_1(x)$ *is a solution of the inhomogeneous system.*

iii) If $\mathbf{y}_1(x)$ *and* $\mathbf{y}_2(x)$ *are two solutions of the inhomogeneous system, then* $\mathbf{y}_1(x) - \mathbf{y}_2(x)$ *is a solution of the homogeneous system.*

Proof: Direct calculations, exercise. ■

Cororllary 2.33 General solution of the inhomogeneous system.

i) If $\mathbf{y}_1(x), \mathbf{y}_2(x), \ldots, \mathbf{y}_k(x)$ *are solutions of the homogeneous system, then any linear combination* $\sum_{i=1}^{k} c_i\mathbf{y}_i(x)$, $c_1, \ldots, c_k \in \mathbb{R}$, *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 2.34 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* $\mathbf{y}(x) \ : \ I \to \mathbb{R}^n$ *of the initial value problem to* (2.16) *with the initial value* $\mathbf{y}(x_0) = \mathbf{y}_0 \in \mathbb{R}^n$.

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 (2.16) by $\mathbf{f}(x,\mathbf{y})$, it follows that

$$\begin{aligned}
\|\mathbf{f}(x,\mathbf{y}_1) - \mathbf{f}(x,\mathbf{y}_2)\|_\infty &= \max_{i=1,\ldots,n} |f_i(x,\mathbf{y}_1) - f_i(x,\mathbf{y}_2)| \\
&= \max_{i=1,\ldots,n} \left| \sum_{j=1}^{n} a_{ij}(x) y_{1,j}(x) + f_i(x) - \sum_{j=1}^{n} a_{ij}(x) y_{2,j}(x) - f_i(x) \right| \\
&= \max_{i=1,\ldots,n} \left| \sum_{j=1}^{n} a_{ij}(x) \left(y_{1,j}(x) - y_{2,j}(x) \right) \right| \\
&\leq n \max_{i,j=1,\ldots,n} |a_{ij}(x)| \max_{i=1,\ldots,n} |y_{1,i}(x) - y_{2,i}(x)| \\
&\leq nM \|\mathbf{y}_1 - \mathbf{y}_2\|_\infty ,
\end{aligned}$$

i.e. the right hand side satisfies a uniform Lipschitz condition with respect to $\mathbf{y}$ 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. ■

2.3.2 Solution of the Homogeneous System

Remark 2.35 *Scalar case.* Because of the superposition principle, one needs the general solution of the homogeneous system

$$\mathbf{y}'(x) = A(x)\mathbf{y}(x) \tag{2.17}$$

for finding the general solution of (2.16). The homogeneous system has always the trivial solution $\mathbf{y}(x) = \mathbf{0}$.

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 (2.17), it is possible to specify the general solution with the help of the exponential. □

Definition 2.36 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 2.37 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 componentwise, 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 real case.

Proof:

i) with induction and comparison test with a majorizing series, see literature,
ii) follows from i), exercise,
iii) follows from ii), exercise,
iv) follows from iii),
v) direct calculation with difference quotient, exercise.

■

Example 2.38 *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 2.37. 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

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

It holds that

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

□

Theorem 2.39 General solution of the homogeneous linear system of first order. *The general solution of* (2.17) *is*

$$\mathbf{y}_h(x) = e^{\int_{x_0}^{x} A(t)\ dt}\mathbf{c}, \quad \mathbf{c} \in \mathbb{R}^n, x_0 \in (a,b). \tag{2.18}$$

The integral is defined component-wise.

Proof: *i)* (2.18) *is a solution of* (2.17). 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

$$\mathbf{y}_h'(x) = \frac{d}{dx}\left(e^{\int_{x_0}^{x} A(t)\ dt}\mathbf{c}\right) = \frac{d}{dx}\left(\int_{x_0}^{x} A(t)\ dt\right) e^{\int_{x_0}^{x} A(t)\ dt}\mathbf{c} = A(x)e^{\int_{x_0}^{x} A(t)\ dt}\mathbf{c}.$$

ii) every solution of (2.17) *is of form* (2.18). Consider an arbitrary solution $\tilde{\mathbf{y}}_h(x)$ of (2.17) with $\tilde{\mathbf{y}}_h(x_0) \in \mathbb{R}^n$. Take in (2.18) $\mathbf{c} = \tilde{\mathbf{y}}_h(x_0)$. Then, it follows that

$$\mathbf{y}_h(x_0) = e^{\int_{x_0}^{x_0} A(t)\ dt}\tilde{\mathbf{y}}_h(x_0) = \underbrace{e^0}_{=I}\tilde{\mathbf{y}}_h(x_0) = \tilde{\mathbf{y}}_h(x_0).$$

That means, $e^{\int_{x_0}^{x} A(t)\ dt}\tilde{\mathbf{y}}_h(x_0)$ is a solution of (2.17) which has in x_0 the same initial value as $\tilde{\mathbf{y}}_h(x)$. Since the solution of the initial value problem is unique, Theorem 2.34, it follows that $\tilde{\mathbf{y}}_h(x) = e^{\int_{x_0}^{x} A(t)\ dt}\tilde{\mathbf{y}}_h(x_0)$. ■

2.3.3 Linear Systems of First Order with Constant Coefficients

Remark 2.40 *Linear system of first order differential equations with constant coefficients.* A linear system of first order differential equations with constant coefficients has the form

$$\mathbf{y}'(x) = A\mathbf{y}(x) + \mathbf{f}(x), \quad A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \in \mathbb{R}^{n\times n}. \tag{2.19}$$

Thus, the homogeneous system has the form

$$\mathbf{y}'(x) = A\mathbf{y}(x). \tag{2.20}$$

Its general solution is given by

$$\mathbf{y}_h(x) = e^{Ax}\mathbf{c}, \quad \mathbf{c} \in \mathbb{R}^n, \tag{2.21}$$

see Theorem 2.39. □

Remark 2.41 *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 (2.20). 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 2.42. 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 2.2.3. The other components of the solution vector of (2.20) can be obtained by back substitution. □

Example 2.42 *Elimination method, substitution method.* Find the solution of

$$\mathbf{y}'(x) = \begin{pmatrix} -3 & -1 \\ 1 & -1 \end{pmatrix} \mathbf{y}(x) \iff y_1'(x) = -3y_1(x) - y_2(x),\ y_2'(x) = y_1(x) - y_2(x).$$

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

$$y_2''(x) + y_2'(x) = -3\left(y_2'(x) + y_2(x)\right) - y_2(x) \iff y_2''(x) + 4y_2'(x) + 4y_2(x) = 0.$$

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

$$\mathbf{y} = \begin{pmatrix} -c_1 + c_2 \\ c_1 \end{pmatrix} e^{-2x} + \begin{pmatrix} -c_2 \\ c_2 \end{pmatrix} x e^{-2x}.$$

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

$$\mathbf{y} = \begin{pmatrix} C_1 \\ C_2 - C_1 \end{pmatrix} e^{-2x} + \begin{pmatrix} C_2 \\ -C_2 \end{pmatrix} x e^{-2x}.$$

If an initial condition is given, then corresponding constants can be determined. □

Remark 2.43 *Other methods for computing the general solution of the homogeneous system.* There are also other methods for computing the general solution of (2.20).

- 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

$$\mathbf{y}(x) = C\mathbf{z}(x) \quad \Longrightarrow \quad \mathbf{y}'(x) = C\mathbf{z}'(x).$$

Inserting into (2.20) yields

$$C\mathbf{z}'(x) = AC\mathbf{z}(x) \quad \Longleftrightarrow \quad \mathbf{z}'(x) = C^{-1}AC\mathbf{z}(x).$$

This is a triangular system for $\mathbf{z}(x)$, which is solved successively for the components of $\mathbf{z}(x)$. The solution of (2.20) is obtained by computing $C\mathbf{z}(x)$.

- 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 2.44 *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 $\mathbf{c}$ in (2.21) by $\mathbf{c}(x)$, inserts this expression into (2.19), obtains conditions for $\mathbf{c}'(x)$, and tries to compute $\mathbf{c}(x)$ from these conditions.
- *Appropriate ansatz (Störgliedansätze).* If each component of the right hand side $\mathbf{f}(x)$ 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 $\mathbf{f}(x)$ of (2.19) is $(n-1)$ times continuously differentiable, then one can proceed exactly as in the elimination method. One obtains for one component of $\mathbf{y}(x)$ an inhomogeneous ordinary differential equation of order n with constant coefficients, for which one has to find a special solution. A special solution for (2.19) is obtained by back substitution.

□

2.4 Implicit Runge–Kutta Schemes

Remark 2.45 *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 stages that depend not only on previously computed stages but also on not yet computed stages. Thus, one has to solve a nonlinear problem for computing these stages. 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. □

Remark 2.46 *Derivation of implicit Runge–Kutta methods.* Implicit Runge–Kutta schemes can be derived from the integral representation (1.3) 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 2.47 *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 $P_s(t)$ in 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 $2s$. □

Remark 2.48 *Simplifying order conditions.* The order conditions for an implicit Runge–Kutta scheme with s stages are the same as given in Theorem 1.26 and Remark 1.28. These conditions lead to a nonlinear system of equations for computing the parameters of the scheme. This computations are generally quite complicated.

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

$$\begin{aligned}
B(p) \quad &: \quad \sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \quad k = 1, \ldots, p, \\
C(l) \quad &: \quad \sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{1}{k} c_i^k, \quad i = 1, \ldots, s, \ k = 1, \ldots, l, \\
D(m) \quad &: \quad \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} = \frac{1}{k} b_j \left(1 - c_j^k\right), \quad j = 1, \ldots, s, \ k = 1, \ldots, m,
\end{aligned} \tag{2.22}$$

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 2.49 *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

$$y(x_0 + h) = h \int_0^1 f(x_0 + h\theta) \ d\theta.$$

A Runge–Kutta method with s stages gives

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

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

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

The Runge–Kutta scheme yields

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

Comparing the last two formulas, 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 (1.6) with the upper limit s

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

□

Example 2.50 *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.

$$\begin{aligned} \sum_{i=1}^{s} b_i c_i^{k-1} &= \frac{1}{k}, \quad k = 1, \ldots, 2s, \\ \sum_{j=1}^{s} a_{ij} c_j^{k-1} &= \frac{1}{k} c_i^k, \quad i = 1, \ldots, s,\ k = 1, \ldots, s, \\ \sum_{i=1}^{s} b_i c_i^{k-1} a_{ij} &= \frac{1}{k} b_j \left(1 - c_j^k\right), \quad j = 1, \ldots, s,\ k = 1, \ldots, s. \end{aligned} \tag{2.23}$$

  An example is the implicit mid point rule, whose coefficients can be derived by setting $s = 1$ in (2.23). One obtains the following conditions

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

  Consequently, the implicit mid point rule is given by

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

- *Gauss–Radau*[3] *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 ($s = 1$):
  - $\begin{array}{c|c} 0 & 1 \\ \hline & 1 \end{array}$ $\quad s = 1,\ p = 1,$
  - $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$ $\quad s = 1,\ p = 1,$ implicit Euler scheme.

  The first scheme does not satisfy condition (1.6).
- *Gauss–Lobatto*[4] *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[5]–Nicolson[6] scheme

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

[3] Rodolphe Radau (1835 – 1911)
[4] Rehuel Lobatto (1797 – 1866)
[5] John Crank (1916 – 2006)
[6] P. Nicolson

- other scheme

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

The second scheme does not satisfy condition (1.6).

□

Remark 2.51 *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. □

2.5 Stability Theory

Remark 2.52 *On the stability theory.* The stability theory studies numerical methods for solving ordinary differential equations for the linear initial value problem

$$y'(x) = \lambda y(x), \quad y(0) = 1,\ \lambda \in \mathbb{C}. \tag{2.24}$$

It will turn out the even at the simple initial value problem (2.24) the most important stability properties of numerical methods can be studied. The solution of (2.24) 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 $\mathrm{Re}(\lambda) > 0$, then the difference

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

becomes for each $\delta_0 \neq 0$ arbitrarily large if x is sufficiently large. That means, the initial value problem (2.24) is not stable in this case.

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 in the data result only in small changes of the solution. This is the case that 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 (2.24) 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 2.53 *Behavior of different one-step methods for one step of the model problem* (2.24).

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 (2.24)

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

   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 (2.24), 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 2.50. For the linear differential equation (2.24), one gets

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

   Hence, for the trapezoidal rule one has that $\lim_{|z|\to\infty} |R(z)| = 1$, 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 2.54 Stability function. Let $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^s$, $\hat{\mathbb{C}} = \mathbb{C} \cup \infty$, and consider a Runge–Kutta method with s stages and with the parameters $(A, \mathbf{b}, \mathbf{c})$. Then, the function

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

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

Remark 2.55 *Stability functions from Example 2.53.* All stability functions from Example 2.53 can be written in the form given in Definition 2.54. One obtains, e.g., for the trapezoidal rule

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

from what follows that

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

□

Theorem 2.56 Form of the stability function of Runge–Kutta schemes. *Given a Runge–Kutta scheme with s stages and with the parameters $(A, \mathbf{b}, \mathbf{c})$, then the stability function $R(z)$ is a rational function 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

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

which can be checked easily by multiplication with $(I - zA)$ and using that $A^s = 0$ since A is strictly lower triangular. It follows that $R(z)$ is a polynomial in z of degree 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 $(I - zA)\zeta = \mathbf{1}$. 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 which 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 z. Multiplying with $z\mathbf{b}^T$ from the left hand side gives just a rational function with polynomials of at most degree s in the numerator and in the denominator.

There is only one case that 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 2.57 Solution of the initial value problem (2.24) **obtained with a Runge–Kutta scheme.** *Consider a Runge–Kutta method with s stages and with the parameters $(A, \mathbf{b}, \mathbf{c})$. 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* (2.24)*. 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 down the Runge–Kutta scheme for (2.24) and by induction. *exercise* ■

Definition 2.58 Stability domain. The stability domain of a one-step method is the set

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

□

Remark 2.59 *Requirement for the stability domain.* The stability domain of the initial value problem (2.24) is, see Remark 2.52,

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

since $R(z) = e^z$. In this domain, the solution decreases (for $\text{Re}(z) < 0$) or its absolute value is constant (for $\text{Re}(z) = 0$). One should now require from a universally usable numerical method for solving initial value problems that $\mathbb{C}_0^- \subseteq S$. □

Definition 2.60 A-stable method. If for the stability domain S of a one-step method holds that $\mathbb{C}_0^- \subseteq S$, then this one-step method is called A-stable. □

Lemma 2.61 Property of an A-stable method. *Consider an A-stable one-step method, then it is* $|R(\infty)| \leq 1$.

Proof: The statement follows directly from $\mathbb{C}_0^- \subseteq S$ and $|R(z)| \leq 1$ for all $z \in S$. ■

Remark 2.62 *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 2.63 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 holds that $|R(\infty)| = 0$. □

Example 2.64 *Stability of some one-step methods.* The types of stability defined in Definitions 2.60 and 2.63 are of utmost importance for the quality of a numerical method. (numerical demonstrations)

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. One has to use very small steps lengths in order to get stable simulations.

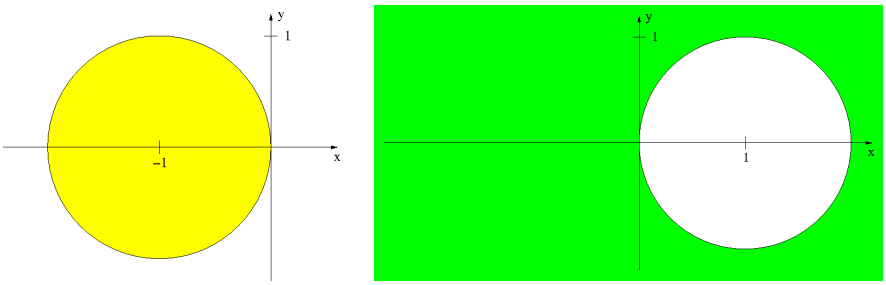

Figure 2.2: Stability domain of the explicit Euler method (left) and the implicit Euler method (right).

   The smallness of the step lengths for stable simulations 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 2.53 it is already known that $|R(\infty)| = 0$ such 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 (2.24).
In general, one can apply in the implicit Euler method much larger steps than, e.g., in 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 $z = a + ib$

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

It follows that

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

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 2.65 *Extension of the theory to linear systems.* Consider the linear system of ordinary differential equations with constant coefficients

$$\mathbf{y}'(x) = A\mathbf{y}(x), \quad \mathbf{y}(0) = \mathbf{y}_0, \ A \in \mathbb{R}^{n\times n}, \ \mathbf{y}_0 \in \mathbb{R}^n. \tag{2.25}$$

It will be assumed that the matrix A possesses n mutually different eigenvalues $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$. The solution of (2.25) has the form, see (2.21),

$$\mathbf{y}(x) = e^{Ax}\mathbf{y}_0.$$

Since A has n mutually different eigenvalues, this matrix can be diagonalized, i.e., there exists a matrix $Q \in \mathbb{R}^{n\times n}$ such that

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

The columns $\mathbf{q}_i$ of Q are the eigenvectors of A. Using the substitution

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

one obtains the differential equation

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

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

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

It follows that the general solution of (2.25) has the form

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

Inserting this expression into the initial value problem gives

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

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

$$\mathbf{y}(x) = \sum_{i=1}^{n} \left(Q^{-1}\mathbf{y}_0\right)_i e^{\lambda_i x}\mathbf{q}_i, \tag{2.26}$$

where $\left(\left(Q^{-1}\mathbf{y}_0\right)_i\right.$ is the i-th component of $Q^{-1}\mathbf{y}_0$. 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 consideration 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 (2.26) with respect to stability is the term with the eigenvalue of A with the largest absolute value of its real part. □

Definition 2.66 Stiff system of ordinary differential equations. The linear system of ordinary differential equations

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

is called stiff, if all eigenvalues λ_i of 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 2.67 *On Definition 2.66.* Definition 2.66 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, only the eigenvalue with the largest absolute value of the real part. □

Remark 2.68 *Local stiffness for general ordinary differential equations.* The notation of stiffness can be extended in some sense from linear differential equations to general differential equations. The differential equation

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

can be transformed, by introducing the functions

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

to the autonomous form

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

By linearizing at the initial value $\tilde{\mathbf{y}}_0$, one obtains a differential equation of the form $\tilde{\mathbf{y}}'(x) = A\tilde{\mathbf{y}}(x)$. 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 and they do not describe the behavior of a numerical method in the whole domain of definition of the nonlinear problem. □

2.6 Rosenbrock Methods

Remark 2.69 *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 defining 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 2.70 *Linearly implicit Runge–Kutta method.* Consider, without loss of generality, the autonomous initial value problem

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

see Remark 1.30. DIRK methods, see Remark 2.51, has 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

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

The quasi Newton method for solving the j-th equation leads to an iterative scheme of the form

$$\mathbf{K}_j^{(n+1)} = \mathbf{K}_j^{(n)} - (I - a_{jj}hJ)^{-1}\left[\mathbf{K}_j^{(n)} - \mathbf{f}\left(\mathbf{y}_k + h\sum_{l=1}^{j-1} a_{jl}\mathbf{K}_l + ha_{jj}\mathbf{K}_j^{(n)}\right)\right],$$

$n = 0, 1, \ldots$. In this scheme, one uses usually the approximation of the derivative $J = \mathbf{f}_\mathbf{y}(\mathbf{y}_k)$ 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 $(I - a_{jj}hJ)$ is non-singular 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 $\mathbf{K}_j^{(0)}$ is available. With the ansatz (linear combination of the already computed stages)

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

where the coefficients d_{jl}, $l = 1, \ldots, j-1$, still need to be determined, one obtains an implicit method with linear systems of equations

$$\begin{aligned} (I - a_{jj}hJ)\,\mathbf{K}_j &= \mathbf{f}\left(\mathbf{y}_k + h\sum_{l=1}^{j-1}(a_{jl} + d_{jl})\,\mathbf{K}_l\right) - hJ\sum_{l=1}^{j-1} d_{jl}\mathbf{K}_l, \; j = 1, \ldots, s, \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + h\sum_{j=1}^{s} b_j\mathbf{K}_j. \end{aligned} \tag{2.27}$$

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 computational complex than the original implicit methods and the first goal stated in Remark 2.69 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 2.71 Stability of linearly implicit Runge–Kutta methods. *Consider a Runge–Kutta method with the parameters $(A, \mathbf{b}, \mathbf{c})$, where $A \in \mathbb{R}^{s\times s}$ is a non-singular lower triangular matrix. Then, the corresponding linearly implicit Runge–Kutta method (2.27) with $J = \mathbf{f_y}(\mathbf{y}_k)$ 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 (2.27) has the form

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

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

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

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

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

Let h be chosen in such a way that z is not an eigenvalue of A. Then, one obtains by inserting this equation into the second equation of (2.27)

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

Now one can see that in the parentheses there is the stability function $R(z)$ of the original Runge–Kutta method, see Definition 2.54. ■

Remark 2.72 *On the stability and consistency.* Since the most important stability properties of a numerical method for solving 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 2.73 *Linearly implicit Euler method.* The implicit Euler method has the Butcher tableau

1	1
	1

With (2.27), it follows that the linearly implicit Euler method has the form

$$(I - h\mathbf{f}_{\mathbf{y}}(\mathbf{y}_k))\,\mathbf{K}_1 = \mathbf{f}\,(\mathbf{y}_k)\,, \quad \mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{K}_1$$

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 2.74 *Rosenbrock*[7] *methods.* Another possibility for simplifying the usage of linearly implicit methods and decreasing the numerical costs consists in using for all stages the same coefficient $a_{jj} = a$. In this case, all linear systems of equations in (2.27) possess the same system matrix $(I - ahJ)$. Then, one nees only one LU decomposition of this matrix and can solve all systems in (2.27) with this decomposition. This approach is called Rosenbrock methods or Rosenbrock–Wanner[8] methods (ROW methods)

$$\begin{aligned}(I - ahJ)\,\mathbf{K}_j &= \mathbf{f}\left(\mathbf{y}_k + h\sum_{l=1}^{j-1}(a_{jl} + d_{jl})\,\mathbf{K}_l\right) - hJ\sum_{l=1}^{j-1} d_{jl}\mathbf{K}_l,\ j = 1,\ldots,s, \\ \mathbf{y}_{k+1} &= \mathbf{y}_k + h\sum_{j=1}^{s} b_j\mathbf{K}_j.\end{aligned} \tag{2.28}$$

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 2.75 *The method* `ode23s`. In MATLAB, one can find for solving stiff ordinary differential equations the Rosenbrock method `ode23s`, see Shampine and Reichelt (1997). This method has the form

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

From the equation for the second stage, it follows that $d_{21} = a$. Then, one obtains with (2.28) $a_{21} = 1/2 - d_{21} = 1/2 - a$. Using the condition that the nodes are the sums of the rows of the matrix, it follows that the corresponding Butcher tableau looks like

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

□

Theorem 2.76 Order of `ode23s`. *The Rosenbrock method* `ode23s` *is of second order.*

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.

[7] Howard H. Rosenbrock (1920 – 2010)
[8] Gerhard Wanner, born 1942

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

$$(I - ahJ)\mathbf{K} = \mathbf{f}.$$

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

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

It follows, using the linear system of equations, that

$$\frac{1}{2}\,\|\mathbf{K}\|_2 \leq \|(I - ahJ)\mathbf{K}\|_2 = \|\mathbf{f}\|_2 \quad \Longleftrightarrow \quad \|\mathbf{K}\|_2 \leq 2\,\|\mathbf{f}\|_2 .$$

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

One obtains for the first stage of `ode23s` by recursive insertion

$$\begin{aligned} \mathbf{K}_1 &= \mathbf{f}(\mathbf{y}_k) + ahJ\mathbf{K}_1 = \mathbf{f}(\mathbf{y}_k) + ahJ\left(\mathbf{f}(\mathbf{y}_k) + ahJ\mathbf{K}_1\right) \\ &= \mathbf{f}(\mathbf{y}_k) + ahJ\mathbf{f}(\mathbf{y}_k) + h^2a^2J^2\mathbf{K}_1 \\ &= \mathbf{f}(\mathbf{y}_k) + ahJ\mathbf{f}(\mathbf{y}_k) + \mathcal{O}\left(h^2\right). \end{aligned} \tag{2.29}$$

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

$$\begin{aligned} \mathbf{K}_2 &= \mathbf{f}\left(\mathbf{y}_k + \frac{1}{2}h\mathbf{K}_1\right) - ahJ\mathbf{K}_1 + ahJ\mathbf{K}_2 \\ &= \mathbf{f}\left(\mathbf{y}_k\right) + \frac{1}{2}h\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)\mathbf{K}_1 - ahJ\mathbf{K}_1 + ahJ\mathbf{K}_2 + \mathcal{O}\left(h^2\right) \\ &\overset{(2.29)}{=} \mathbf{f}\left(\mathbf{y}_k\right) + \frac{1}{2}h\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)\mathbf{f}\left(\mathbf{y}_k\right) - ahJ\mathbf{f}(\mathbf{y}_k) + ahJ\mathbf{K}_2 + \mathcal{O}\left(h^2\right) \\ &= \mathbf{f}\left(\mathbf{y}_k\right) + \frac{1}{2}h\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)\mathbf{f}\left(\mathbf{y}_k\right) - ahJ\mathbf{f}(\mathbf{y}_k) + ahJ\mathbf{f}(\mathbf{y}_k) + \mathcal{O}\left(h^2\right) \\ &= \mathbf{f}\left(\mathbf{y}_k\right) + \frac{1}{2}h\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)\mathbf{f}\left(\mathbf{y}_k\right) + \mathcal{O}\left(h^2\right). \end{aligned}$$

Inserting these results gives for on step of `ode23s`

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}\left(\mathbf{y}_k\right) + \frac{1}{2}h^2\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)\mathbf{f}\left(\mathbf{y}_k\right) + \mathcal{O}\left(h^3\right). \tag{2.30}$$

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

$$\begin{aligned} \mathbf{y}(x_{k+1}) &= \mathbf{y}(x_k) + h\mathbf{y}'(x_k) + \frac{h^2}{2}\mathbf{y}''(x_k) + \mathcal{O}\left(h^3\right) \\ &= \mathbf{y}(x_k) + h\mathbf{f}\left(\mathbf{y}_k\right) + \frac{h^2}{2}\frac{\partial \mathbf{f}\left(\mathbf{y}\right)}{\partial x}(x_k) + \mathcal{O}\left(h^3\right) \\ &= \mathbf{y}(x_k) + h\mathbf{f}\left(\mathbf{y}_k\right) + \frac{h^2}{2}\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)\mathbf{y}'(x_k) + \mathcal{O}\left(h^3\right) \\ &= \mathbf{y}(x_k) + h\mathbf{f}\left(\mathbf{y}_k\right) + \frac{h^2}{2}\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)\mathbf{f}\left(\mathbf{y}_k\right) + \mathcal{O}\left(h^3\right). \end{aligned}$$

Starting with the exact value, then the first three terms of (2.30) correspond to the Taylor series expansion of the solution $\mathbf{y}(x)$ 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 2.77 *To the proof of Theorem 2.76.* Note that it is not needed in the proof of Theorem 2.76 that J is the exact derivative $\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)$. The method `ode23s` remains a second order method if J is only an approximation of $\mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)$ and even if J is an arbitrary matrix. However, the transfer of the stability properties from the original method to `ode23s` is only guarantied for the choice $J = \mathbf{f}_{\mathbf{y}}\left(\mathbf{y}_k\right)$, see Theorem 2.71. □

Theorem 2.78 Stability function of `ode23s`. *Assume that $J = \mathbf{f}_{\mathbf{y}}(\mathbf{y}_k)$, then the stability function of the Rosenbrock method* `ode23s` *has the form*

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

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

Remark 2.79 *On the order of* `ode23s`. It remains the question if 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 third 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 the J every now and then. □

Cororllary 2.80 Stability of `ode23s`. *If $J = \mathbf{f}_{\mathbf{y}}(\mathbf{y}_k)$, then the Rosenbrock method* `ode23s` *is L-stable.*

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