Numerical Differential Equations

An Undergraduate Course

Ronald H.W. Hoppe
Contents

1. Ordinary Differential Equations 1

1.1 Initial-value Problems: Theoretical foundations and preliminaries 1

  1.1.1 Basic definitions 1

  1.1.2 Existence and uniqueness 3

  1.1.3 Continuous Dependence on the initial data 4

  1.1.4 Elementary examples of numerical integrators 6

  1.1.5 Stiff and non-stiff systems 9

1.2 One-step methods for initial-value problems 13

  1.2.1 Definitions 13

  1.2.2 Convergence, consistency, and stability 13

  1.2.3 Explicit Runge-Kutta methods 21

  1.2.4 Asymptotic expansion of the global discretization error 26

  1.2.5 Explicit extrapolation methods 28

  1.2.6 Adaptive step size control for explicit one-step methods 31

1.3 Multi-step methods 35

  1.3.1 Basic definitions 35

  1.3.2 Convergence, consistency, and stability 37

  1.3.3 Extrapolatory and interpolatory multi-step methods 45

  1.3.4 Predictor-corrector methods 49

  1.3.5 BDF (Backward Difference Formulas) 51
1.4 Numerical integration of stiff systems 52
1.4.1 Linear stability theory 52
1.4.2 A-stability, L-stability, and a(α)-stability 54
1.4.3 Implicit and semi-implicit Runge-Kutta methods 57
1.4.4 Linear multi-step methods and Dahlquist’s second barrier 61
1.4.5 A(α)-stability of implicit BDF 63
1.4.6 Numerical solution of differential-algebraic systems 64
1.5 Numerical solution of boundary-value problems 68
1.5.1 Theoretical foundations and preliminaries 68
1.5.2 Shooting methods 71
1.5.3 Finite difference approximations 75
1.5.4 Galerkin approximations 77
Literature 80
Partial Differential Equations

2.1 Theoretical foundations and preliminaries

2.2 Linear second order elliptic boundary value problems
  2.2.1 Preliminaries
  2.2.2 Finite difference approximations
  2.2.3 Sobolev spaces
  2.2.4 Finite element approximations

2.3 The heat equation
  2.3.1 Preliminaries
  2.3.2 Method of lines and Rothe's method
  2.3.3 Finite difference approximations
  2.3.4 Finite element approximations

2.4 The wave equation
  2.4.1 Preliminaries
  2.4.2 Finite difference approximations
  2.4.3 Finite element approximations

Literature

Exercises
1. Ordinary Differential Equations

1.1 Initial-value Problems: Theoretical foundations and preliminaries

1.1.1 Basic definitions

An ordinary differential equation is an equation which contains an unknown function in one independent variable along with certain derivatives of that function. The order of the differential equation corresponds to the highest derivative that occurs in the equation.

**Definition 1.1. (First order system)**

(i) Assume that $g : I \times D_1 \times D_2 \to \mathbb{R}^d, g = (g_1, ..., g_d)^T$, where $I := [a,b] \subset \mathbb{R}, D_i \subset \mathbb{R}^d, 1 \leq i \leq 2$, and $y : I \to \mathbb{R}^d, y = (y_1, ..., y_d)^T$ with $y' = (y'_1, ..., y'_d)^T, y'_i := \frac{dy_i}{dx}, 1 \leq i \leq d$. Assume further that $(y(x), y'(x)) \in D_1 \times D_2, x \in I$. Then

$$g(x, y(x), y'(x)) = 0 \quad , \quad x \in I, \quad (1.1a)$$

is called a nonlinear system of ordinary differential equations of first order in implicit form.

A function $y \in C^1(I)^d$ satisfying (1.1a) is called a solution of the system of ordinary differential equations of first order.

(ii) Let $\alpha \in \mathbb{R}^d, \alpha = (\alpha_1, ..., \alpha_d)^T$, and assume that $y \in C^1(D)$ satisfies (1.1a) and

$$y(a) = \alpha. \quad (1.1b)$$

Then, (1.1a),(1.1b) is called an initial value problem for the system of ordinary differential equations of first order. (iii) Assume that $B : I \times D \to \mathbb{R}^{d \times d}, B = (B_{ij})_{i,j=1}^d, D \subset \mathbb{R}^d$, and $f : I \times D \to \mathbb{R}^d, f = (f_1, ..., f_d)^T$. Assume further that $y : I \to \mathbb{R}^d$ as in (i) with $y(x) \in D, x \in I$. Then

$$B(x, y(x)) y'(x) = f(x, y(x)), \quad x \in I, \quad (1.2)$$

is called a quasilinear system of ordinary differential equations of first order in implicit form. In case $B = B(y), f = f(y)$, the system (1.2) is said to be autonomous.

(iv) In the special case $B = I$ we obtain

$$y'(x) = f(x, y(x)), \quad x \in I. \quad (1.3)$$
The system (1.3) is referred to as a system of ordinary differential equations of first order in explicit form.

(v) Assume that $A, B : I \rightarrow \mathbb{R}^{d \times d}$, $A = (A_{ij})_{i,j=1}^d$, $B = (B_{ij})_{i,j=1}^d$ and $f : I \rightarrow \mathbb{R}^d$, $f = (f_1, ..., f_d)^T$. Let further $y$ be as in (iii). Then

\begin{align}
(1.4a) & \quad B(x) y'(x) = A(x) y(x) + f(x) \quad x \in I, \\
(1.4b) & \quad y'(x) = A(x) y(x) + f(x) \quad x \in I,
\end{align}

are called linear systems of ordinary differential equations of first order in implicit resp. explicit form.

**Definition 1.2. (n-th order system)**

(i) Assume that $g : I \times D \rightarrow \mathbb{R}^d$, $I := [a, b] \subset \mathbb{R}$, $D \subset \mathbb{R}^{(n+1)d}$, $n \in \mathbb{N}$, and $y : I \rightarrow \mathbb{R}^d$ with $y^{(i)}(x) := d^i y(x)/dx^i$, $1 \leq i \leq n$. Assume further that $(y(x), y^{(1)}(x), ..., y^{(n)}(x))^T \in D$, $x \in I$. Then

\begin{align}
(1.5a) & \quad g(x, y(x), y^{(1)}(x), ..., y^{(n)}(x)) = 0, \quad x \in I,
\end{align}

is called a system of ordinary differential equations of $n$-th order in implicit form.

(ii) Assume that $\alpha \in \mathbb{R}^{nd}$, $\alpha = (\alpha_0, \cdots, \alpha_{n-1})^T$, $\alpha_i \in \mathbb{R}^d$, $0 \leq i \leq n-1$. Moreover, suppose that

\begin{align}
(1.5b) & \quad y^{(i)}(a) = \alpha_i, \quad 0 \leq i \leq n-1.
\end{align}

Then, (1.5a),(1.5b) is called an initial value problem for the system of ordinary differential equations of $n$-th order. (iii) The definitions of quasilinear resp. linear systems of $n$-th order and of systems of $n$-th order in explicit form can be given analogously.

In the sequel we will restrict ourselves to systems of first order, since higher order equations can be equivalently written as a first order system as the following result shows.

**Lemma 1.3. (Equivalence of n-th order equation and first order system)**

Any ordinary differential equation of $n$-th order is equivalent to a system of ordinary differential equations of first order.

**Proof.** Without restriction of generality we assume that $d = 1$ and that the ordinary differential equation of $n$-th order is explicit, i.e.,

\[ y^{(n)}(x) = f(x, y(x), y'(x), ..., y^{(n-1)}(x)), \quad x \in I. \]
Setting \( y_i(x) := y^{(i-1)}(x), x \in I, 1 \leq i \leq n \), there holds

\[
\begin{align*}
  y'_1(x) &= y_2(x) \\
  y'_2(x) &= y_3(x) \\
  & \quad \text{...} \\
  y'_{n-1}(x) &= y_n(x) \\
  y'_n(x) &= f(x, y_1(x), \ldots, y_{n-1}(x))
\end{align*}
\]

i.e., \((y_1, \ldots, y_n)\) satisfies a system of first order. \(\square\)

### 1.1.2 Existence and uniqueness

We consider an initial-value problem for a system of first order ordinary differential equations in explicit form. The continuity of the right-hand side of the ordinary differential equation is sufficient for the existence of a solution.

**Theorem 1.4. (Peano’s existence theorem)**

Assume that \( I := [a, b] \subset \mathbb{R}, D \subset \mathbb{R}^d, f \in C(I \times D), \) and \( \alpha \in D. \) Then, the initial value problem

\[
\begin{align*}
  (1.6a) & \quad y'(x) = f(x, y(x)), \quad x \in I, \\
  (1.6b) & \quad y(a) = \alpha
\end{align*}
\]

has a solution \( y \in C^1(D). \)

However, the continuity of the right-hand side does not guarantee uniqueness as the following example shows:

**Example 1.5. (Non-uniqueness)**

We consider the initial-value problem

\[
\begin{align*}
  y'(x) &= \sqrt{|y(x)|}, \quad x \in \mathbb{R}, \\
  y(2) &= 1.
\end{align*}
\]

If \( y = y(x), x \in \mathbb{R} \) is a solution, so is \( z(x) := -y(-x), \) i.e., it suffices to consider positive solutions. Separation of variables results in

\[
\int \frac{dy}{\sqrt{y}} = \int dx \quad \Rightarrow \quad 2\sqrt{y} = x + C \quad \Rightarrow \quad y(x; C) = \frac{(x + C)^2}{4}.
\]
The condition \( y(2) = 1 \) yields \((C + 2)^2 = 4 \implies C = 0\). Multiple solutions:
\[
y(x; a) = \begin{cases} 
  x^2/4 & , \quad x > 0 \\
  0 & , \quad a \leq x \leq 0 \ (a \leq 0) \\
  - (x - a)^2/4 & , \quad x < a
\end{cases}
\]

\[
y(x) = \begin{cases} 
  x^2/4 & , \quad x > 0 \\
  0 & , \quad x \leq 0
\end{cases}
\]

If the right-hand side additionally satisfies a Lipschitz-condition in its second argument, we do have uniqueness.

**Theorem 1.6. (Existence and uniqueness theorem of Picard-Lindelöf)**
Assume that \( f : I \times D \to \mathbb{R}^d, I := [a, b] \subset \mathbb{R}, D \subset \mathbb{R}^d \), and \( \alpha \in D \). Assume further that
\[
(1.7a) \quad f \in C(I \times D),
\]
and that there exists a constant \( L > 0 \) such that for all \((x, y_i) \in I \times D, 1 \leq i \leq 2\), it holds
\[
(1.7b) \quad \| f(x, y_1) - f(x, y_2) \| \leq L \| y_1 - y_2 \|.
\]
Then, the initial value problem (1.6a),(1.6b) has one and only one solution.

**1.1.3 Continuous Dependence on the initial data**
We study the continuous dependence of the solution on the initial data.

**Lemma 1.7. (Gronwall’s lemma)**
Let \( \Phi : I \to \mathbb{R}, I := [a, b] \subset \mathbb{R}, \Phi \in C(I) \), and assume that
\[
\Phi(x) \leq \alpha + \beta \int_a^x \Phi(s) \, ds, \quad x \in I, \quad \alpha \geq 0, \beta > 0.
\]
Then, it holds
\[
\Phi(x) \leq \alpha \exp(\beta (x - a)), \quad x \in I.
\]

**Proof.** Let \( \varepsilon > 0 \). We consider
\[
\Psi(x) := (\alpha + \varepsilon) \exp(\beta (x - a)), \quad x \in I \implies \Psi'(x) = \beta \psi(x) \]
\[
\implies \psi(x) = \psi(a) + \beta \int_a^x \Psi(s) \, ds.
\]
We show

\[ \Phi(x) < \Psi(x), \quad x \in I. \]

Obviously, we have

\[ \Phi(a) \leq \alpha \leq \alpha + \epsilon = \Psi(a). \]

In contradiction to the assertion we assume the existence of \( x \in (a, b] \) such that \( \Phi(x) \geq \Psi(x) \). Set

\[ x_0 := \min \{ x \in (a, b] \mid \Phi(x) = \Psi(x) \}. \]

We have \( \Phi(x) \leq \Psi(x), x \in [a, x_0] \), whence

\[ \Phi(x_0) \leq \alpha + \beta \int_a^{x_0} \Phi(s) \, ds < \alpha + \epsilon + \beta \int_a^{x_0} \Psi(s) \, ds = \Psi(x_0) \]

contradicting the assumption \( \Phi(x_0) = \Psi(x_0) \). \( \square \)

**Theorem 1.8. (Continuous dependence on the initial data)**

*Under the assumptions of the Theorem of Picard-Lindelöf let \( y_i, 1 \leq i \leq 2 \), two solutions of the system \( y'(x) = f(x, y(x)), x \in I \), with respect to the initial conditions \( y_i(a) = \alpha_i \). Then, it holds

\[ \|y_1(x) - y_2(x)\| \leq \|\alpha_1 - \alpha_2\| \exp(L(x-a)), \quad x \in I. \]

*Proof.* In view of

\[ y_i(x) = \alpha_i + \int_a^x f(s, y_i(s)) \, ds, \quad 1 \leq i \leq 2, \]

we have

\[ \|y_1(x) - y_2(x)\| \leq \|\alpha_1 - \alpha_2\| + \int_a^x \|f(s, y_1(s)) - f(s, y_2(s))\| \, ds \leq \]

\[ \|\alpha_1 - \alpha_2\| + \frac{L}{\beta} \int_a^x \|y_1(s) - y_2(s)\| \, ds. \]

Gronwall’s lemma allows to conclude. \( \square \)
1.1.4 Elementary examples of numerical integrators

We consider an initial-value problem for a scalar ordinary differential equation of first order in explicit form
\[
y'(x) = f(x, y(x)), \quad x \in [a, b],
y(a) = \alpha.
\]

We introduce a partition of the interval \( I \) according to
\[
I_h := \{a =: x_0 < x_1 < \cdots < x_N := b\}, \quad N \in \mathbb{N},
\]
with step sizes \( h_k := x_{k+1} - x_k, 0 \leq k \leq N - 1 \). The integration of the differential equation over the subinterval \([x_k, x_{k+1}]\) yields
\[
y(x_{k+1}) - y(x_k) = \int_{x_k}^{x_{k+1}} f(x, y(x)) \, dx.
\]

The idea is to approximate the integral on the right-hand side by a quadrature formula.

**Example 1.9. (Explicit and implicit Euler method)**

The integral on the right-hand side of (1.8) describes the area enclosed by the graph of the function \( f(\cdot, y(\cdot)) \) and the interval \([x_k, x_{k+1}]\). In the explicit Euler method it is approximated by the area of the rectangle \([x_k, x_{k+1}] \times [0, f(x_k, y(x_k))]\) (cf. Figure 1 left). In the implicit Euler method we approximate the integral by the area of the rectangle \([x_k, x_{k+1}] \times [0, f(x_{k+1}, y(x_{k+1}))]\) (cf. Figure 1 right).

\[\text{Figure 1. The explicit (left) and the implicit (right) Euler method.}\]

We replace the unknown function values \( y(x_k) \) and \( y(x_{k+1}) \) by some approximations \( y_k \) and \( y_{k+1} \) and thus obtain:
(i) **Explicit Euler method:** For \( k = 0, 1, \cdots, N - 1 \) compute

\[
\begin{align*}
(1.9a) & \quad y_{k+1} = y_k + h_k f(x_k, y_k), \\
(1.9b) & \quad y_0 = \alpha.
\end{align*}
\]

We note that \( y_{k+1} \) can be explicitly computed by the evaluation of the right-hand side in (1.9a).

(ii) **Implicit Euler method:** For \( k = 0, 1, \cdots, N - 1 \) compute

\[
\begin{align*}
(1.10a) & \quad y_{k+1} = y_k + h_k f(x_k+1, y_{k+1}), \\
(1.10b) & \quad y_0 = \alpha.
\end{align*}
\]

We note that (1.10a) requires the solution of a nonlinear equation, i.e., \( y_{k+1} \) is implicitly given.

**Example 1.10. (Implicit trapezoidal rule)**

In the implicit trapezoidal rule the integral on the right-hand side of (1.8) is approximated by the area of the trapeze as depicted in Figure 2:

\[
\int_{x_k}^{x_{k+1}} f(x, y(x)) \, dx \approx h_k \left( f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1})) \right).
\]

![Figure 2. The implicit trapezoidal rule.](image)

**Implicit trapezoidal rule:** For \( k = 0, 1, \cdots, N - 1 \) compute

\[
\begin{align*}
(1.11a) & \quad y_{k+1} = y_k + h_k \left( f(x_k, y_k) + f(x_{k+1}, y_{k+1}) \right), \\
(1.11b) & \quad y_0 = \alpha.
\end{align*}
\]
Example 1.11. (Explicit and implicit midpoint rule)
In the explicit midpoint rule, the differential equation \( y'(x) = f(x, y(x)) \) is integrated over the interval \([x_k, x_{k+2}]\) resulting in
\[
y(x_{k+2}) - y(x_k) = \int_{x_k}^{x_{k+2}} f(x, y(x)) \, dx.
\]
In the explicit midpoint rule we use a quadrature formula which approximates the integral by the area of the rectangle \([x_k, x_{k+2}] \times [0, f(x_{k+1}, y(x_{k+1}))]\) (cf. Figure 3 left):
\[
\int_{x_k}^{x_{k+2}} f(x, y(x)) \, dx \approx (h_k + h_{k+1}) f(x_{k+1}, y(x_{k+1})).
\]
In the implicit midpoint rule, the differential equation \( y'(x) = f(x, y(x)) \) is integrated over the interval \([x_k, x_{k+1}]\) and a quadrature formula is used which approximates the integral by the area of the rectangle \([x_k, x_{k+1}] \times [0, f(x_k + h_k/2, y(x_k + h_k/2))]\) (cf. Figure 3 right):
\[
\int_{x_k}^{x_{k+1}} f(x, y(x)) \, dx \approx h_k f(x_k + h_k/2, y(x_k + h_k/2)).
\]
Since we do not want to evaluate the function \( f \) at intermediate points, the function value \( f(x_k + h_k/2, y(x_k + h_k/2)) \) is approximated by the arithmetic mean of the function values with respect to \( x_k \) and \( x_{k+1} \):
\[
f(x_k + h_k/2, y(x_k + h_k/2)) \approx \frac{1}{2} \left( f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1})) \right).
\]

Figure 3. The explicit (left) and the implicit (right) midpoint rule.
Again, we replace the unknown function values \(y(x_k), y(x_{k+1}),\) and \(y(x_{k+2})\) by some approximations \(y_k, y_{k+1},\) and \(y_{k+2}\. We note that the explicit midpoint rule requires two start values \(y_0\) and \(y_1\. Given \(y_0 = \alpha, y_1\) can be computed, e.g., by the explicit Euler method. We thus obtain:

(i) **Explicit midpoint rule:** For \(k = 0, 1, \ldots, N-2\) compute

\[
\begin{align*}
  y_{k+2} &= y_k + (h_k + h_{k+1}) f(x_k, y_k), \\
  y_0 &= \alpha, \quad y_1 = \alpha + h_0 f(a, \alpha).
\end{align*}
\]

We note that \(y_{k+2}\) can be explicitly computed by the evaluation of the right-hand side in (1.12a).

(ii) **Implicit midpoint rule:** For \(k = 0, 1, \ldots, N-1\) compute

\[
\begin{align*}
  y_{k+1} &= y_k + \frac{h_k}{2} \left( f(x_k, y_k) + f(x_{k+1}, y_{k+1}) \right), \\
  y_0 &= \alpha.
\end{align*}
\]

We note that \(y_{k+1}\) is implicitly given, i.e., the computation requires the solution of a nonlinear equation.

**Remark 1.12. (Comparison of explicit and implicit methods)**

Since explicit methods only require function evaluations, whereas implicit methods require the solution of nonlinear equations and are thus computationally more expensive, the question is why one should use implicit methods instead of explicit methods. The answer is that some systems of ordinary differential equations require the solution by implicit methods as will be illustrated in the next subsection.

**1.1.5 Stiff and non-stiff systems**

A system of ordinary differential equations is called stiff, if the solution has components with extremely different growth behavior. Otherwise, it is said to be non-stiff.

**Example 1.13. (Stiff system)**

We consider the following initial-value problem for a linear system of first order

\[
\begin{pmatrix}
  y_1'(x) \\
  y_2'(x)
\end{pmatrix} = \frac{1}{2} \begin{pmatrix}
  101 & 99 \\
  99 & 101
\end{pmatrix} \begin{pmatrix}
  y_1(x) \\
  y_2(x)
\end{pmatrix}, \quad x \geq 0,
\]

\[
\begin{pmatrix}
  y_1(0) \\
  y_2(0)
\end{pmatrix} = \begin{pmatrix}
  2 \\
  0
\end{pmatrix}.
\]
We diagonalize the matrix $A$ by a similarity transformation $TAT^*$, where $T := (e_1|e_2)$ with $e_i, 1 \leq i \leq 2$, being the orthonormal eigenvectors associated with the eigenvalues $\lambda_1, \lambda_2$ of $A$:

Computation of the eigenvalues of $A$ as the zeroes of the characteristic polynomial

$$\det(\lambda I - A) = \lambda^2 - 101\lambda + 100 \implies \lambda_1 = 100, \ \lambda_2 = 1.$$ 

The orthonormal eigenvectors are given by

$$e_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} - 1 \\ \frac{\sqrt{2}}{2} - 1 \end{pmatrix},$$

$$e_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} - 1 \\ -\frac{\sqrt{2}}{2} - 1 \end{pmatrix}.$$ 

Using the transformation matrix

$$T := \begin{pmatrix} \frac{\sqrt{2}}{2} - 1 & \frac{\sqrt{2}}{2} - 1 \\ \frac{\sqrt{2}}{2} - 1 & -\frac{\sqrt{2}}{2} - 1 \end{pmatrix}, \quad \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix} = T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix},$$

we obtain the transformed linear system

$$\begin{pmatrix} \tilde{y}_1'(x) \\ \tilde{y}_2'(x) \end{pmatrix} = -TAT^*T \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \quad x \geq 0,$$

$$\begin{pmatrix} \tilde{y}_1(0) \\ \tilde{y}_2(0) \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ \sqrt{2} \end{pmatrix}.$$

It has the solution

$$\tilde{y}_1(x) = \sqrt{2} \exp(-100x), \ \tilde{y}_2(x) = \sqrt{2} \exp(-x).$$

The back transformation

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = T \begin{pmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{pmatrix}$$

results in

$$y_1(x) = \exp(-100x) + \exp(-x), \ y_2(x) = \exp(-100x) - \exp(-x).$$

The linear system is stiff, since the solution component $\exp(-100x)$ decays much faster than the solution component $\exp(-x)$.

We now consider the numerical solution of the initial-value problem from the example above with respect to an equidistant grid with grid
points \( x_k = kh, k \in \mathbb{N}, h > 0 \), using the explicit Euler method. The application of this method to the transformed linear system results in
\[
\begin{align*}
\dot{y}_1, k &= \dot{y}_1, k-1 - 100h\dot{y}_1, k-1 = \\
(1 - 100h)\dot{y}_1, k-1 &= (1 - 100h)^k\ddot{y}_1, 0, \\
\dot{y}_2, k &= \dot{y}_2, k-1 - h\dot{y}_2, k-1 = \\
(1 - h)\dot{y}_2, k-1 &= (1 - h)^k\ddot{y}_2, 0.
\end{align*}
\]
The back transformation yields
\[
\begin{align*}
y_1, k &= (1 - 100h)^k + (1 - h)^k, \\
y_2, k &= (1 - 100h)^k - (1 - h)^k.
\end{align*}
\]
We want that the approximate solution decays as the solution of the continuous problem, i.e., \( y_1, k, y_2, k \to 0 \) for \( k \to \infty \). Hence, we must have
\[
|1 - 100h| < 1 \iff h < \frac{1}{50}.
\]
On the other hand, we consider the numerical solution by the implicit Euler method. The application of this method to the transformed linear system gives
\[
\begin{align*}
\ddot{y}_1, k &= \ddot{y}_1, k-1 + 100h\ddot{y}_1, k \implies \\
\ddot{y}_1, k &= (1 + 100h)^{-1}\ddot{y}_1, k-1 = (1 + 100h)^{-k}\ddot{y}_1, 0, \\
\ddot{y}_2, k &= \ddot{y}_2, k-1 + h\ddot{y}_2, k \implies \\
\ddot{y}_2, k &= (1 + h)^{-1}\ddot{y}_2, k-1 = (1 + h)^{-k}\ddot{y}_2, 0.
\end{align*}
\]
Transforming back, we obtain
\[
\begin{align*}
y_1, k &= (1 + 100h)^{-k} + (1 + h)^{-k}, \\
y_2, k &= (1 + 100h)^{-k} - (1 - h)^{-k}.
\end{align*}
\]
Obviously, we have \( y_1, k, y_2, k \to 0 \) for \( k \to \infty \) without any restriction on the step size \( h \).

The application of the explicit and implicit Euler method to the stiff system from Example 1.13 shows that the explicit Euler method is much less suited than the implicit Euler method, since it requires the choice of a very small step size to guarantee that the approximate solution behaves in the same way as the exact solution. This does not only yield a significant higher amount of computational work but also may lead to rounding errors. The next sections will show indeed that
explicit methods are well suited for non-stiff problems, whereas stiff problems require the application of implicit schemes.
1.2 One-step methods for initial-value problems

1.2.1 Definitions

We consider the initial-value problem

\begin{align}
  y'(x) &= f(x, y(x)), \quad x \in I := [a, b], \\
  y(a) &= \alpha.
\end{align}

We assume \( f \in C(I \times D), D \subset \mathbb{R}^d, \alpha \in D \), and suppose that there exists \( L \geq 0 \) such that

\[ \|f(x, y_1) - f(x, y_2)\| \leq L \|y_1 - y_2\| \]

for all \((x, y_i) \in I \times D, 1 \leq i \leq 2\).

**Definition 1.15.** (Explicit and implicit one-step methods)

Let \( I_h := \{x_k = a + kh \mid 0 \leq k \leq N\}, N \in \mathbb{N} \), be an equidistant partition of \( I = [a, b] \) of step size \( h = (b - a)/N \) and \( \Phi : I_h \times I_h \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}_+ \rightarrow \mathbb{R}^d \). Then, the scheme

\begin{align}
  y_{k+1} &= y_k + h \Phi(x_k, x_{k+1}, y_k, y_{k+1}, h), \quad 0 \leq k \leq N - 1, \\
  y_0 &= \alpha
\end{align}

is called a one-step method with increment function \( \Phi \). The one-step method is said to be explicit, if \( \Phi \) does not depend on \( y_{k+1} \), and is called implicit otherwise.

**Remark 1.16.** (Interpretation as a difference equation)

Setting \( I'_h := I_h \setminus \{x_N\} \) and introducing the grid function \( y_h : I_h \rightarrow \mathbb{R}^d \), the one-step method (1.16a),(1.16b) can be equivalent written as the difference equation of first order

\begin{align}
  y_h(x + h) &= y_h(x) + h \Phi(x, x + h, y_h(x), y_h(x + h), h), \quad x \in I'_h, \\
  y_h(a) &= \alpha.
\end{align}

1.2.2 Convergence, consistency, and stability

**Definition 1.17.** (Convergence and order of convergence)

The grid function

\[ e_h(x) := y_h(x) - y(x), \quad x \in I_h, \]

is called the global discretization error. The one-step method (1.16a), (1.16b) is said to be convergent, if it holds

\[ \max_{x \in I_h} \|e_h(x)\| \rightarrow 0 \quad (h \rightarrow 0). \]
It is said to be convergent of order $p$, if
\[
(1.20) \quad \max_{x \in I_h} \| e_h(x) \| = O(h^p) \quad (h \to 0).
\]

The convergence can be characterized by consistency and stability. Consistency guarantees that the one-step method (1.16a), (1.16b) is a meaningful approximation of the initial-value problem (1.14a), (1.14b). We can not expect that the exact solution $y$ of the initial-value problem satisfies the one-step method, but if we insert the exact solution into the equivalent difference equation, we should have
\[
\begin{align*}
\frac{y(x + h) - y(x)}{h} &= \Phi(x, x + h, y(x), y(x + h), h), \\
\downarrow & \quad h \to 0 \\
y'(x) &= f(x, y(x)).
\end{align*}
\]

**Definition 1.18. (Consistency, order of consistency)**
The grid function
\[
\tau_h(x) := h^{-1} (y(x + h) - y(x)) - \Phi(x, x + h, y(x), y(x + h), h),
\]
is called the local discretization error. The one-step method (1.16a), (1.16b) is said to consistent with the initial-value problem (1.14a), (1.14b), if
\[
(1.21) \quad h \sum_{x \in I'_h} \| \tau_h(x) \| \to 0 \quad (h \to 0).
\]

It is said to be consistent of order $p$, if
\[
(1.22) \quad h \sum_{x \in I'_h} \| \tau_h(x) \| = O(h^p) \quad (h \to 0).
\]

**Lemma 1.19. (Sufficient condition for consistency)**
Assume that
\[
(1.23) \quad \max_{x \in I'_h} \| \tau_h(x) \| \to 0 \quad (h \to 0).
\]
Then, the one-step method (1.16a), (1.16b) is consistent with the initial-value problem (1.14a), (1.14b).

The condition (1.23) holds true if and only if
\[
(1.24) \quad \max_{x \in I'_h} \| \Phi(x, x + h, y(x), y(x + h), h) - f(x, y(x)) \| \to 0
\]
as $h \to 0$.

**Proof.** The proof is left as an exercise. \qed
Example 1.20. (Consistency order: explicit Euler method)
The explicit Euler method reads
\[ y_h(x + h) = y_h(x) + h \ f(x, y_h(x)), \quad y_h(a) = \alpha. \]
Hence, the local discretization error is given by
\[ \tau_h(x) = h^{-1} (y(x + h) - y(x)) - f(x, y(x)). \]
Taylor expansion leads
\[
\begin{align*}
y(x + h) &= y(x) + h \ y'(x) + \frac{1}{2} h^2 \ y''(x) + O(h^3) \quad \Rightarrow \\
h^{-1} (y(x + h) - y(x)) &= y'(x) + \frac{1}{2} h \ y''(x) + O(h^2) = \\
f(x, y(x)) + \frac{1}{2} h \ y''(x) + O(h^2).
\end{align*}
\]
It follows that
\[ \tau_h(x) = \frac{1}{2} h \ y''(x) + O(h^2). \]
Hence, if \( y \in C^2(I) \), the explicit Euler method is consistent of order \( p = 1 \).

Example 1.21. (Consistency order: implicit trapezoidal rule)
The implicit trapezoidal rule reads
\[ y_h(x + h) = y_h(x) + \frac{h}{2} \ [f(x, y_h(x)) + f(x + h, y_h(x + h))], \quad y_h(a) = \alpha. \]
By Taylor expansion we obtain
\[
\begin{align*}
y(x + h) &= y(x) + h \ y'(x) + \frac{1}{2} h^2 \ y''(x) + \frac{1}{6} h^3 \ y'''(x) + O(h^4) \quad \Rightarrow \\
h^{-1} (y(x + h) - y(x)) &= y'(x) + \frac{1}{2} h \ y''(x) + \frac{1}{6} h^2 \ y'''(x) + O(h^3) \quad \Rightarrow \\
h^{-1} (y(x + h) - y(x)) &= f(x, y(x)) + \\
\frac{1}{2} h \ (f_x(x, y(x)) + f_y(x, y(x))) \ f(x, y(x))) + \frac{1}{6} h^2 \ y'''(x) + O(h^3),
\end{align*}
\]
and
\[
\begin{align*}
f(x + h, y(x + h)) &= f(x, y(x)) + h \ f_x(x, y(x)) + \\
h \ f_y(x, y(x)) \ y'(x) + O(h^2) \quad \Rightarrow \\
\frac{1}{2} \ (f(x, y(x)) + f(x + h, y(x + h))) &= f(x, y(x)) + \\
\frac{1}{2} h \ (f_x(x, y(x)) + f_y(x, y(x))) \ f(x, y(x))) + O(h^2).
\end{align*}
\]
It follows that
\[ \tau_h(x) = O(h^2), \]
i.e., for \( y \in C^3(I) \) the implicit trapezoidal rule if of consistency order \( p = 2 \).

In contrast to the consistency which related the one-step method and the initial-value problem, stability is alone a property of the one-step method. It states the continuous dependence of the solution of the one-step method on its data.

We restrict ourselves to the explicit one-step method
\begin{align}
(1.25a) \quad y_h(x + h) &= y_h(x) + h \Phi(x, y_h(x), h), \\
(1.25b) \quad y_h(a) &= \alpha.
\end{align}

**Definition 1.22. (Asymptotic stability)**

Consider the perturbed one-step method
\begin{align}
(1.26a) \quad z_h(x + h) &= z_h(x) + h \left( \Phi(x, z_h(x), h) + \sigma_h(x) \right), \\
(1.26b) \quad z_h(a) &= \alpha + \beta_h.
\end{align}

The explicit one-step method (1.25a),(1.25b) is called asymptotically stable, if there exist \( h_{\text{max}} > 0 \) and for each \( \varepsilon > 0 \) a number \( \delta > 0 \) such that for all perturbations \( \sigma_h, \beta_h \) with
\[ \|\beta_h\| + h \sum_{x \in I'_h} \|\sigma_h(x)\| < \delta \]
for \( h < h_{\text{max}} \) the solutions of the perturbed one-step method (1.26a), (1.26b) satisfy
\[ \max_{x \in I_h} \|(y_h - z_h)(x)\| + h \sum_{x \in I'_h} \|(D_hy_h - D_hz_h)(x)\| < \varepsilon, \]
where \((D_hy_h)(x) := h^{-1} (y_h(x + h) - y_h(x)) \), \( x \in I'_h \).

A significant tool in the proof of the continuous dependence of the exact solution on the initial data was Gronwall’s lemma (cf. Lemma 1.7). A discrete analog of Gronwall’s lemma will play a prominent role in the proof of the asymptotic stability of the one-step method (1.25a),(1.25b).
Lemma 1.23. (Discrete Gronwall’s lemma)
Consider a grid function $v_h : I_h \to \mathbb{R}^d$ with $v_j := v_h(x_j), 0 \leq j \leq N,$ and assume that there exist $\delta \geq 0$ and $\eta \geq 0$ such that

\begin{align}
(1.27a) \quad & \|v_{j+1}\| \leq \delta h \sum_{k=0}^{j} \|v_k\| + \eta, \quad 0 \leq j \leq N - 1, \\
(1.27b) \quad & \|v_0\| \leq \eta,
\end{align}

hold true. Then we have

\begin{align}
(1.28) \quad & \|v_j\| \leq \eta \exp(\delta (x_j - a)), \quad 0 \leq j \leq N.
\end{align}

Proof. Suppose that $\|v_j\| \leq M, 0 \leq j \leq N$. By induction we prove that

\begin{align}
(1.29) \quad & \|v_j\| \leq M \frac{\delta^m (x_j - a)^m}{m!} + \eta \sum_{\ell=0}^{m-1} \frac{\delta^\ell (x_j - a)^\ell}{\ell!}.
\end{align}

(i) Induction basis: Obviously, the assertion (1.28) holds true for $j = 0, m \in \mathbb{N},$ and $m = 0, 0 \leq j \leq N$.

(ii) Induction assumption: The assertion (1.28) holds true for some $m \in \mathbb{N}$ and $0 \leq j \leq N$.

(iii) Induction conclusion: We have

$$\|v_{j+1}\| \leq \delta h \sum_{k=0}^{j} \left( M \frac{\delta^m (x_k - a)^m}{m!} + \eta \sum_{\ell=0}^{m-1} \frac{\delta^\ell (x_k - a)^\ell}{\ell!} \right) + \eta.$$

Using the elementary inequality

$$h \sum_{k=0}^{j} (x_k - a)^\ell \leq \int_a^{x_{j+1}} (x - a)^\ell \; dx = \frac{(x_{j+1} - a)^{\ell+1}}{\ell + 1},$$

for $m + 1$ and $j + 1$ we obtain

$$\|v_{j+1}\| \leq M \frac{\delta^{m+1} (x_{j+1} - a)^{m+1}}{(m + 1)!} + \eta \sum_{\ell=0}^{m} \frac{\delta^\ell (x_{j+1} - a)^\ell}{\ell!}.$$ 

\[\Box\]

Theorem 1.24. (Sufficient condition for asymptotic stability)
Assume that the increment function $\Phi$ satisfies the following Lipschitz condition: There exist a neighborhood $U \subset I \times D$ of the graph $\{(x, y(x)) \mid x \in I\}$ of the exact solution $y$ of the initial-value problem
and numbers \( h_{\text{max}} > 0 \) and \( L_\Phi \geq 0 \) such that for all \( 0 < h < h_{\text{max}} \) and all \( (x, y_i) \in U, 1 \leq i \leq 2 \), it holds
\[
\| \Phi(x, y_1, h) - \Phi(x, y_2, h) \| \leq L_\Phi \| y_1 - y_2 \|.  
\]
Then, for all perturbations \( \sigma_h, \beta_h \) with
\[
\| \beta_h \| + h \sum_{x \in I_h} \| \sigma_h(x) \| \leq \delta
\]
and \( x \in I_h \) resp. \( x \in I'_h \) the solution \( y_h \) of the unperturbed one-step method (1.25a),(1.25b) and the solution \( z_h \) of the perturbed one-step method (1.26a),(1.26b) satisfy
\[
\| (y_h - z_h)(x) \| \leq \left( \| \beta_h \| + h \sum_{x \in I_h} \| \sigma_h(x) \| \right) \exp(L_\Phi(x - a)),
\]
\[
\| (D_h y_h - D_h z_h)(x) \| \leq L_\Phi \| (y_h - z_h)(x) \| + \| \sigma_h(x) \|. 
\]
In particular, this implies asymptotic stability of the one-step method (1.25a),(1.25b).

**Proof.** Without restriction of generality let \( U = I \times D \). We set
\[
w_h(x) := z_h(x) - y_h(x), \quad x \in I_h.
\]
With \( w_j := w_h(x_j) \) it follows that
\[
w_{j+1} = w_j + h \left( \Phi(x_j, z_j, h) + \sigma_j - \Phi(x_j, y_j, h) \right), \quad w_0 = \beta_h \implies
\]
\[
w_{j+1} = w_0 + h \sum_{k=0}^j \left( \sigma_k + \Phi(x_k, z_k, h) - \Phi(x_k, y_k, h) \right).
\]
The Lipschitz condition (1.30) yields
\[
\| w_{j+1} \| \leq \| \beta_h \| + h \sum_{k=0}^N \| \sigma_k \| + L_\Phi h \sum_{k=0}^j \| w_k \|.
\]
Then, the discrete Gronwall’s lemma implies (1.31a). The assertion (1.31b) is a direct consequence of (1.31a) and
\[
(D_h w_h)(x) = \Phi(x, z_h(x), h) + \sigma_h(x) - \Phi(x, y_h(x), h).
\]
\[\square\]

An important result is that consistency and asymptotic stability of the one-step method (1.25a),(1.25b) implies convergence with the order of convergence being the order of consistency.
Theorem 1.25. (Consistency and stability imply convergence)

Assume that the one-step method (1.25a),(1.25b) is consistent with the initial-value problem (1.14a),(1.14b) and asymptotically stable. Then it holds

\begin{align}
(1.32a) \quad & \max_{x \in I_h} \|(y_h - y)(x)\| \to 0 \quad (h \to 0), \\
(1.32b) \quad & \sum_{x \in I_h} \|(D_h y_h - D_h y)(x)\| \to 0 \quad (h \to 0).
\end{align}

Proof. We set \( z_h(x) := y(x), \) \( x \in I_h. \) It follows that \( \sigma_h(x) = \tau_h(x) \) and \( \beta_h = 0. \) In view of the consistency, for each \( \varepsilon > 0 \) there exist \( h_{\max}(\varepsilon) > 0 \) and \( \delta(\varepsilon) > 0 \) such that for \( 0 < h < h_{\max}(\varepsilon) \) we have

\[ h \sum_{x \in I'_h} \|\tau_h(x)\| < \delta(\varepsilon). \]

The asymptotic stability implies (1.32a),(1.32b). \( \square \)

Theorem 1.26. (Characterization of convergence)

Assume that the one-step method (1.25a),(1.25b) satisfies the Lipschitz condition (1.30). Then, the following two statements are equivalent:

(i) The one-step method (1.25a),(1.25b) is convergent and (1.32a), (1.32b) hold true.

(ii) The one-step method (1.25a),(1.25b) is consistent with the initial-value problem (1.14a),(1.14b).

For consistent one-step methods there exists \( h_{\max} > 0 \) such that for all \( 0 < h < h_{\max} \) the following a priori error estimate holds true:

\[ \|(y_h - y)(x)\| \leq (h \sum_{x \in I'_h} \|\tau_h(x)\|) \exp(L_\Phi (x - a)). \]

If additionally \( \max_{x \in I'_h} \|\tau_h(x)\| \to 0 \quad (h \to 0) \) is satisfied, then we have

\[ \max_{x \in I'_h} \|(D_h y_h - D_h y)(x)\| \to 0 \quad (h \to 0). \]

Proof. We first show that (ii) implies (i). Theorem 1.24 gives the asymptotic stability of the one-step method and Theorem 1.25 implies convergence in the sense of (1.32a),(1.32b). The a priori estimate (2.90) and (1.34) follow from Theorem 1.24 with \( z_h(x) := y(x) \) so that \( \sigma_h(x) = \tau_h(x) \) and \( \beta_h = 0. \)
Next, we show that (i) implies (ii). In view of the convergence, for sufficiently small $h$ the Lipschitz condition (1.30) can be applied to
\[ \Phi(x, y_h(x), h) - \Phi(x, y(x), h), \]
whence
\[
\|\tau_h(x)\| = \|D_h y(x) - \Phi(x, y(x), h)\| = \\
\|(D_h y(x) - \Phi(x, y(x), h) - D_h y_h(x) + \Phi(x, y_h(x), h)\| \leq \\
\|(D_h y_h - D_h y)(x)\| + L_\Phi \|y_h - y\|(x) |.
\]
Multiplication by $h$ and summation over $x \in I_h'$ allows to conclude. \(\square\)

**Remark 1.27. (Implicit one-step methods)**
The definition of asymptotic stability can be generalized to implicit one-step methods. The corresponding stability and convergence results hold true as well.

**Remark 1.28. (Lipschitz condition for the increment function)**
If the increment function $\Phi$ only involves the right-hand side $f$ of the initial-value problem, the Lipschitz condition (1.30) follows from the corresponding Lipschitz condition for $f$. 
1.2.3 Explicit Runge-Kutta methods

The explicit Euler method (1.9a),(1.9b) uses approximate slopes \( f(x_k, y_k) \approx f(x_k, y(x_k)) \) of the graph of the exact solution in \( x_k \) (cf. Figure 4).

\[
\begin{align*}
\text{(1.35a)} & \quad y_{k+1} = y_k + h \left[ b_1 f(x_k, y_k) + b_2 f(x_k + a_2 h, y_k + a_2 h f(x_k, y_k)) \right], \quad k \geq 0, \\
\text{(1.35b)} & \quad y_0 = \alpha,
\end{align*}
\]

where \( b_i \in \mathbb{R}, 1 \leq i \leq 2 \). The goal is to determine \( a_{21}, b_1, b_2 \) such that the one-step method (1.35a),(1.35b) has the order of consistency \( p = 2 \).

Assuming \( f \) to be sufficiently smooth, Taylor expansion yields

\[
\begin{align*}
\text{(1.36a)} & \quad \frac{y(x + h) - y(x)}{h} = y'(x) + \frac{h}{2} y''(x) + O(h^2) = \\
& \quad f(x, y(x)) + \frac{h}{2} \left( f_x(x, y(x)) + f_y(x, y(x)) f(x, y(x)) \right) + O(h^2), \\
\text{(1.36b)} & \quad b_1 f(x, y(x)) + b_2 f(x + a_2 h, y(x) + a_2 h f(x, y(x))) = \\
& \quad (b_1 + b_2) f(x, y(x)) + \\
& \quad h \left( a_{21} b_2 f_x(x, y(x)) + a_{21} b_2 f_y(x, y(x)) f(x, y(x)) \right) + O(h^2).
\end{align*}
\]

A comparison of the coefficients in (1.36a) and (1.36b) results in

\[ b_1 + b_2 = 1, \quad a_{21} b_2 = \frac{1}{2}. \]
Choosing $\beta := b_2$ gives
\[ b_1 = 1 - \beta, \quad a_{21} = \frac{1}{2\beta}. \]

**Definition 1.29. (Method of Runge and Heun)**
The explicit one-step method

\[ y_{k+1} = y_k + (1 - \beta) \ h \  f(x_k, y_k) + \beta \ h \ f(x_k + \frac{1}{2\beta} h, y_k + \frac{h}{2\beta} f(x_k, y_k)), \quad k \geq 0, \]

is called the method of Runge. In particular, for the special choice $\beta = \frac{1}{2}$ it is called the method of Heun.

We now study the question whether by a suitable choice of $\beta$ we can achieve an order of consistency $p = 3$. To this end we determine a further term in the Taylor expansion:

\[ \frac{y(x+h) - y(x)}{h} = y'(x) + \frac{h}{2} y''(x) + \frac{h^2}{6} y'''(x) + O(h^3) = \]

\[ f + \frac{h}{2} \left( f_x + f_y f \right) + \frac{h^2}{6} \left( f_{xx} + 2f f_{xy} + f^2 f_{yy} + f_x f_y + f f_y^2 \right) + O(h^3), \]

\( (1 - \beta) f(x, y(x)) + \beta f(x + \frac{h}{2\beta}, y(x) + \frac{h}{2\beta} f(x, y(x))) = \)

\[ f + \frac{h}{2} \left( f_x + f_y f \right) + \frac{h^2}{8\beta} \left( f_{xx} + 2f f_{xy} + f^2 f_{yy} \right) + O(h^3). \]

Again, a comparison of coefficients yields

\[ \tau_h(x) = \frac{h^2}{6} \left( 1 - \frac{3}{4\beta} \right) \left( f_{xx} + 2f f_{xy} + f^2 f_{yy} \right) + O(h^3). \]

Hence, the consistency order $p = 3$ can not be achieved by any choice of $\beta$. However, if we choose $\beta = \frac{3}{4}$, the sum of the coefficients in front of the leading error term of $\tau_h$ is minimized.

The construction of the method of Runge can be generalized and gives rise to explicit Runge-Kutta methods of higher order.
Definition 1.30. (Explicit $s$-stage Runge-Kutta method)

The explicit one-step method

\begin{align}
    y_{k+1} &= y_k + \sum_{j=1}^{s} b_j k_j, \quad k \geq 0, \tag{1.38a} \\
y_0 &= \alpha, \tag{1.38b}
\end{align}

with

\begin{align*}
k_1 &= h \, f(x_k, y_k), \\
k_2 &= h \, f(x_k + a_{21}h, y_k + a_{21}k_1), \\
k_3 &= h \, f(x_k + a_{31}h + a_{32}h, y_k + a_{31}k_1 + a_{32}k_2), \\
\vdots & \quad \vdots \\
k_i &= h \, f(x_k + c_i h, y_k + \sum_{j=1}^{i-1} a_{ij} k_j), \quad c_i := \sum_{j=1}^{i-1} a_{ij}, 1 \leq i \leq s,
\end{align*}

is called an explicit $s$-stage Runge-Kutta method. It is uniquely determined by the $s(s + 1)/2$ parameters $a_{ij}, 2 \leq i \leq s, 1 \leq j \leq i - 1$, and $b_i, 1 \leq i \leq s$.

An explicit $s$-stage Runge-Kutta method can be described by the so-called Butcher scheme

\begin{center}
\begin{array}{c|ccc}
   c_1 &= 0  & a_{21} \\
c_2 & a_{31} & a_{32} \\
   \cdot & \cdot & \cdot \\
   \cdot & \cdot & \cdot & \cdot \\
   \cdot & \cdot & \cdot & \cdot \\
   \cdot & \cdot & \cdot & \cdot & \cdot \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} \\
\hline
   b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array}
\end{center}

Lemma 1.31. (Consistency of $s$-stage Runge-Kutta methods)

Under the assumption

\begin{equation}
    \sum_{i=1}^{s} b_i = 1 \tag{1.39}
\end{equation}

the explicit $s$-stage Runge-Kutta method (1.38a),(1.38b) is consistent with the initial-value problem (1.14a),(1.14b).
Proof. For the local discretization error we obtain
\[
\tau_h(x) = \frac{y(x + h) - y(x)}{h} - \sum_{i=1}^{s} b_i f(x + c_i h, y(x) + \sum_{j=1}^{i-1} a_{ij} k_j),
\]
where \(k_j\) is given as in (1.38a) with \(y_h\) replaced by \(y(x)\). For \(h \to 0\) it follows that
\[
\tau_h(x) \to y'(x) - \sum_{i=1}^{s} b_i f(x, y(x)) = y'(x) - f(x, y(x)) = 0.
\]

The goal is to determine the parameters \(a_{ij}\) and \(b_i\) in such a way that the highest possible consistency order \(p\) can be achieved. This can be realized by graph-theoretical concepts (Butcher trees). For details we refer to the textbooks by Butcher and Hairer/Norsett/Wanner.

The following table contains the number \(N(s)\) of parameters to be determined and the number \(M(s)\) of equations to be satisfied as well as the maximal achievable consistency order \(p(s)\):

<table>
<thead>
<tr>
<th>(s)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N_s)</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>15</td>
<td>21</td>
<td>28</td>
<td>36</td>
<td>45</td>
<td>55</td>
</tr>
<tr>
<td>(M_s)</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>17</td>
<td>37</td>
<td>85</td>
<td>200</td>
<td>486</td>
<td>1205</td>
</tr>
<tr>
<td>(p_s)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
</tbody>
</table>

If \(M(s) < N(s)\), the 'free' parameters are chosen such that the leading term in the local discretization error gets minimized.

**Example 1.32. (Kutta’s third order method)**

The 4 equations to be satisfied by the 6 unknown parameters are given by
\[
\begin{align*}
b_1 + b_2 + b_3 &= 1, \\
b_2 a_{21} + b_3 (a_{31} + a_{32}) &= \frac{1}{2}, \\
b_2 a_{21}^2 + b_3 (a_{31}^2 + 2 a_{31} a_{32} + a_{32}^2) &= \frac{1}{3}, \\
b_3 a_{32} a_{21} &= \frac{1}{6}.
\end{align*}
\]

The associated Butcher scheme reads
Example 1.33. (Classical fourth order Runge-Kutta method)

The Butcher scheme for the classical fourth order Runge-Kutta method is given by

\[
\begin{array}{cccccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 \\
1 & -1 & 2 & 0 \\
\hline
\frac{5}{6} & \frac{5}{3} & \frac{1}{6}
\end{array}
\]
1.2.4 Asymptotic expansion of the global discretization error

An asymptotic expansion of the global discretization error is a prerequisite for extrapolation methods that will be treated in the subsequent subsection.

We assume that the explicit one-step method (1.25a),(1.25b) has consistency order $p$. Then, the local discretization error admits the asymptotic expansion

$y(x+h) - y(x) - h \Phi(x, y(x), h) =
\begin{align*}
d_{p+1}(x) \ h^{p+1} + d_{p+2}(x) \ h^{p+2} + \cdots
\end{align*}$

(1.40)

A natural question is whether the global discretization error has a similar asymptotic expansion.

**Theorem 1.34. (Theorem of Gragg)**

Let $f \in C^{p+k}(I \times D)$ and let $\Phi \in C^{p+k}(I \times D \times \mathbb{R}_+), k \geq 1$, be the increment function of an explicit one-step method with $\Phi(x, y, 0) = f(x, y), (x, y) \in I \times D$. Moreover, assume that there exist functions $d_{p+i} \in C^{k-i}(I), 1 \leq i \leq k$, such that the local discretization error satisfies

$y(x+h) - y(x) - h \Phi(x, y(x), h) = \sum_{i=1}^{k} d_{p+i}(x) \ h^{p+i} + O(h^{p+k+1}),$

where $x \in I\_h$. Then, the global discretization error admits the asymptotic expansion

$e_h(x) = \sum_{i=0}^{k-1} e_{p+i}(x) \ h^{p+i} + E_{p+k}(x; h) \ h^{p+k}, \ x \in I\_h.$

Here, the coefficient functions $e_{p+i}, 0 \leq i \leq k-1$, are solutions of the linear initial-value problems

$e'_{p+i}(x) = f_y(x, y(x)) \ e_{p+i}(x) - d_{p+i+1}(x), \ x \in I,$

$e_{p+i}(a) = 0.$

Further, there exist $h_{\text{max}} > 0$ and a constant $C > 0$, independent of $h$, such that

$\|E_{p+k}(x; h)\| \leq C, \quad x \in I\_h, \ 0 < h < h_{\text{max}}.$

**Remark 1.35. (Implicit one-step methods)**

Gragg's theorem can be generalized to implicit one-step methods. However, the iterative solution of the associated nonlinear system of equations may perturb the asymptotic expansion.
Example 1.36. (Asymptotic expansions of the explicit/implicit Euler method)

The explicit and the implicit Euler methods (1.9a), (1.9b), and (1.10a),(1.10b) admit asymptotic expansions of the global discretization error in $h$.

Particular interest is on one-step methods that admit an asymptotic expansion of the global discretization error in $h^2$.

Theorem 1.37. (Theorem of Stetter)

Assume that the one-step method

$y_h(x) = y_h(x + h) = y(x) + h \Phi(x, x + h, y_h(x), y(x), x \in I'_h,$

$y_h(a) = \alpha,$

is symmetric, i.e.,

$\Phi(x, x + h, y_h(x), y_h(x + h), h) = \Phi(x + h, x, y_h(x + h), y_h(x), -h).$

Under the assumptions of Theorem 1.34 it holds

$e_{2m+1}(x) = 0, \ x \in I,$

i.e., there exists an asymptotic expansion of the global discretization error in $h^2$.

Example 1.38. (Explicit midpoint rule)

Although the explicit midpoint rule (cf. Example 1.11) formally is a two-step method, it can be equivalently written as a system of two one-step methods to which the assumptions of Theorem 1.37 apply. Hence, the explicit midpoint rule admits an asymptotic expansion of the global discretization error in $h^2$. 
1.2.5 Explicit extrapolation methods

We consider an explicit one-step method (1.25a),(1.25b) which admits an asymptotic expansion of the global discretization error in $h^\gamma$.

Example 1.39. (Extrapolation in case of an asymptotic expansion in $h$)

In Theorem 1.34 we assume $p = 1, \gamma = 1$. For the step sizes $h$ and $h/2$ we obtain

\begin{align*}
(1.41a) \quad y_h(x) - y(x) &= e_1(x) h + E_2(x; h) h^2, \\
(1.41b) \quad y_{h/2}(x) - y(x) &= \frac{1}{2} e_1(x) h + \frac{1}{4} E_2(x; h) h^2.
\end{align*}

Multiplication of (1.41b) with 2 and subtraction of (1.41a) yield

\[(2 y_{h/2}(x) - y_h(x)) - y(x) = O(h^2).\]

Now, consider the polynomial $p_1 \in P_1(\mathbb{R})$ with $p_1(h_i) = y_{h_i}(x), 1 \leq i \leq 2$, where $h_1 := h$ and $h_2 := h/2$. We obtain

\[p_1(t) = \frac{2}{h} (y_h(x) - y_{h/2}(x)) (t - \frac{h}{2}) + y_{h/2}(x) \implies p_1(0) = 2 y_{h/2}(x) - y_h(x) = \hat{y}_h(x).\]

Hence, the construction of the approximation $\hat{y}_h(x)$ of order $h^2$ corresponds to an extrapolation to the step size $h = 0$.

Example 1.40. (Extrapolation in case of an asymptotic expansion in $h$) In Theorem 1.34 we assume $p = 2, \gamma = 2$. For the step sizes $h$ and $h/2$ we obtain

\begin{align*}
(1.42a) \quad y_h(x) - y(x) &= e_2(x) h^2 + E_4(x; h) h^4, \\
(1.42b) \quad y_{h/2}(x) - y(x) &= \frac{1}{4} e_2(x) h^2 + \frac{1}{16} E_4(x; h) h^4.
\end{align*}

Multiplication of (1.42b) with 4 and subtraction of (1.42a) implies

\[4 y_{h/2}(x) - 3 y_h(x) = O(h^4) \implies 4 y_{h/2}(x) - 3 y_h(x) \overbrace{- y(x) = O(h^4)}^{\hat{y}_h(x)}.\]
We consider the polynomial $p_1 \in P_1(\mathbb{R})$ in $t^2$ with $p_1(h_i) = y_{h_i}(x), 1 \leq i \leq 2$, where $h_1 := h$ and $h_2 := h/2$. It follows that

$$p_1(t^2) = \frac{4}{3h^2} \left( y_h(x) - y_{h/2}(x) \right) \left( t^2 - \frac{h^2}{4} \right) + y_{h/2}(x) \quad \Rightarrow$$

$$p_1(0) = \frac{4}{3} y_{h/2}(x) - \frac{1}{3} y_h(x) = \hat{y}_h(x).$$

Again, the construction of the approximation $\hat{y}_h(x)$ of order $h^4$ corresponds to an extrapolation to the step size $h = 0$.

In practice, more than two step sizes are used for extrapolation: Given a basic step size $H > 0$, we choose a step size sequence $h_i := H/n_i, \quad n_i \in \mathbb{N}, \quad i = 1, 2, \ldots, (1.43)$ characterized by the sequence $F := \{ n_1, n_2, \ldots \}$, and compute the approximations $y(H; h_i) := y_{h_i}, \quad i = 1, 2, \ldots.$

The associated sequence $T_{i,1} = y(H; h_i), i = 1, 2, \ldots,$ forms the first column of an extrapolation tableau. In case of polynomial extrapolation of Aitken-Neville type we determine a polynomial in $h^\gamma$

$$\tilde{T}_{ik}(h) := a_0 + a_1 h^\gamma + \cdots + a_{k-1} h^{(k-1)\gamma},$$

such that

$$\tilde{T}_{ik}(h_j) := y(H; h_j), \quad j = i, i-1, \ldots, i-k+1,$n_i \in \mathbb{N}, i = 1, 2, \ldots,$

and extrapolate to $h = 0$: $T_{ik} := \tilde{T}_{ik}(0)$. This gives rise to the recursion

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}} {\left( \frac{n_i}{n_{i-k+1}} \right)^{\gamma} - 1}, \quad i \geq k, \quad k \geq 2.$$

The extrapolation tableau looks as follows:

\[
\begin{array}{cccccccccc}
  y(H; h_1) & : & T_{11} \\
  y(H; h_2) & : & T_{21} & T_{22} \\
  y(H; h_3) & : & T_{31} & T_{32} & T_{33} \\
  \vdots & : & \vdots & \vdots & \vdots & \vdots & \vdots \\
  \vdots & : & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
  y(H; h_i) & : & T_{i1} & T_{i2} & T_{i3} & \cdots & T_{ii} \\
\end{array}
\]
Common choices of the step size sequence are given by:

(i) **Harmonic sequence**

\[ n_i = i, \quad i \in \mathbb{N}, \]

\[ F_H = \{1, 2, 3, 4, 5, \ldots\}. \]

(ii) **Romberg sequence**

\[ n_i = 2^i, \quad i \in \mathbb{N}, \]

\[ F_R = \{2, 4, 8, 16, 32, \ldots\}. \]

(iii) **Bulirsch sequence**

\[ n_i = \begin{cases} 
3 \cdot 2^{(i-2)/2}, & i \text{ even} \\
2^{(i+1)/2}, & i \text{ odd}
\end{cases}, \]

\[ F_B = \{2, 3, 4, 6, 8, 12, 16, \ldots\}. \]
1.2.6 Adaptive step size control for explicit one-step methods

Let \( y_{k+1} \) be an approximation computed by an explicit one-step method of order \( p \):

\[
y_{k+1} = y_k + h \, \Phi(x_k, y_k, h).
\]

We are looking for an easily computable a posteriori estimator for the global discretization error

\[
e_{k+1} := y_{k+1} - y(x_{k+1}).
\]

**Lemma 1.41. (Error estimator)**

Assume that \( \hat{y}_{k+1} \) is a more accurate approximation of \( y(x_{k+1}) \) according to

\[
\| \hat{y}_{k+1} - y(x_{k+1}) \| \leq q \| y_{k+1} - y(x_{k+1}) \|, \quad 0 \leq q < 1.
\]

Then

\[
\hat{e}_{k+1} := \| \hat{y}_{k+1} - y_{k+1} \|
\]

is an estimator of \( e_{k+1} \) in the sense that

\[
\frac{1}{1+q} \hat{e}_{k+1} \leq e_{k+1} \leq \frac{1}{1-q} \hat{e}_{k+1}.
\]

**Proof.** The assertions can be easily deduced by an application of the right- and left-hand side of the triangle inequality. \( \square \)

We now assume that \( \hat{y}_{k+1} \) is an approximation of \( y \) at \( x_{k+1} \) of order \( p + 1 \), i.e.,

\[
\| \hat{y}_{k+1} - y(x_{k+1}) \| \leq C \, h^{p+1}.
\]

Given a prespecified accuracy \( \varepsilon > 0 \), for a ’reasonable’ new step size \( h := x_{k+2} - x_{k+1} \) we should have

(1.44)

\[
C \, \hat{h}^{p+1} = \rho \varepsilon,
\]

where \( 0 < \rho < 1 \) represents a ’safety factor’. If \( y_{k+1} \) satisfies

\[
\hat{e}_{k+1} := \| \hat{y}_{k+1} - y_{k+1} \| \leq \varepsilon = C \, h^{p+1},
\]

then \( C = \hat{e}_{k+1}/\hat{h}^{p+1} \). Inserting into (1.44) and solving for \( \hat{h} \) yields

\[
\hat{h} = h \left( \frac{\rho \varepsilon}{\hat{e}_{k+1}} \right)^{1/(p+1)}.
\]

There are two strategies to determine \( \hat{y}_{k+1} \):

(i) extrapolation,

(ii) embedded Runge-Kutta methods of higher order.
We first consider extrapolation: Let $y_{k+1}$ and $\bar{y}_{k+1}$ be the approximations associated with the step sizes $h$ and $h/2$, i.e.,

$$(1.45a) \quad y_{k+1} - y(x_{k+1}) \doteq C \ h^p,$$

$$(1.45b) \quad \bar{y}_{k+1} - y(x_{k+1}) \doteq C \left(\frac{h}{2}\right)^p = 2^{-p} \ C \ h^p.$$

Subtraction of (1.45b) from (1.45a) yields

$$y_{k+1} - \bar{y}_{k+1} \doteq C \ h^p \ (1 - 2^{-p}) = C \ h^p \ \frac{2^p - 1}{2p},$$

and hence,

$$C \doteq \frac{2^p - 1}{2p} \ h^{-p} \ (y_{k+1} - \bar{y}_{k+1}).$$

If we insert $C$ into (1.45b) and solve for $y(x_{k+1})$, we obtain

$$\hat{y}_{k+1} = \bar{y}_{k+1} + \frac{\bar{y}_{k+1} - y_{k+1}}{2^p - 1}.$$

Embedded Runge-Kutta methods of higher order are based on Runge-Kutta-Fehlberg methods: Let $y_{k+1}$ be an approximation of $y(x_{k+1})$ obtained by an explicit $s$-stage Runge-Kutta method of order $p$:

$$y_{k+1} = y_k + \sum_{i=1}^{s} b_i \ k_i,$$

$$k_i = h \ f(x_k + c_i h, y_k + \sum_{j=1}^{i-1} a_{ij} \ k_j),$$

$$c_i = \sum_{j=1}^{i-1} a_{ij}.$$

We compute $\hat{y}_{k+1}$ as the solution of an explicit $(s+t)$-stage 'embedded' Runge-Kutta method of order $p + 1$:

$$\hat{y}_{k+1} = y_k + \sum_{i=1}^{s+t} \hat{b}_i \ k_i,$$

$$k_i = h \ f(x_k + c_i h, y_k + \sum_{j=1}^{i-1} a_{ij} \ k_j),$$

$$c_i = \sum_{j=1}^{i-1} a_{ij}.$$

The Butcher scheme of the embedded $(s+t)$-stage Runge-Kutta method reads
A drawback is that the computation is continued with $y_k$ instead of $\hat{y}_k$. Therefore, in practice one considers Dormand-Prince type embedded Runge-Kutta methods: We proceed as above but continue with $\hat{y}_k$, i.e.,

$$y_{k+1} = \hat{y}_k + \sum_{i=1}^{s} b_i k_i,$$

$$k_i = h \, f(x_k + c_i \, h, \, \hat{y}_k + \sum_{j=1}^{i-1} a_{ij} k_j),$$

and use the so-called Fehlberg trick for the reduction of the number of free parameters. For $t = 1$ the Fehlberg trick works as follows: In the following step $x_{k+1} \mapsto x_{k+2}$ use

$$k_{s+1} = h \, f(x_k + c_{s+1} \, h, \, \hat{y}_k + \sum_{j=1}^{s} a_{s+1,j} k_j)$$

as

$$k_1 = h \, f(x_{k+1}, \, \hat{y}_{k+1}).$$

In view of $\hat{y}_{k+1} = \hat{y}_k + \sum_{j=1}^{s+1} \hat{b}_j k_j$ this gives rise to

$$c_{s+1} = 1 \quad \hat{b}_{s+1} = 0 \quad \hat{b}_i = a_{s+1,i}, \quad 1 \leq i \leq s.$$

**Example 1.42. (Dormand/Prince embedded 7-stage Runge-Kutta-Fehlberg method)**

The Butcher scheme for the Dormand/Prince embedded 7-stage Runge-Kutta-Fehlberg method is as follows:
<table>
<thead>
<tr>
<th></th>
<th>1/5</th>
<th>1/5</th>
</tr>
</thead>
<tbody>
<tr>
<td>3/10</td>
<td>3/40</td>
<td>9/40</td>
</tr>
<tr>
<td>4/5</td>
<td>44/45</td>
<td>56/15</td>
</tr>
<tr>
<td>8/9</td>
<td>19372/6561</td>
<td>25360/6561</td>
</tr>
<tr>
<td>1</td>
<td>9017/3168</td>
<td>355/33</td>
</tr>
<tr>
<td>1</td>
<td>35/384</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>35/384</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5179/57600</td>
<td>0</td>
</tr>
</tbody>
</table>
1.3 Multi-step methods

1.3.1 Basic definitions

In the sequel we will use the grid-point sets

\[ I_h := \{ x_j = a + j h \mid 0 \leq j \leq N \}, \quad h := (b - a)/N, \]
\[ I'_h := \{ x_j \in I_h \mid m \leq j \leq N \}, \quad m \geq 2. \]

Definition 1.43. (Multi-step method)

Given real numbers \( \alpha_0, \alpha_1, \ldots, \alpha_m \neq 0 \), a function \( \Phi : I_h \times \mathbb{R}^{d(m+1)} \times \mathbb{R}_+ \to \mathbb{R}^d \) and vectors \( \alpha^{(0)}, \ldots, \alpha^{(m-1)} \in \mathbb{R}^d \), the scheme

\[
\frac{1}{h} \sum_{k=0}^{m} \alpha_k y_{j+k} = \Phi(x_{j+m}, y_j, \ldots, y_{j+m}; h), \quad 0 \leq j \leq N - m, \tag{1.46a}
\]

\[
y_j = \alpha^{(j)}, \quad 0 \leq j \leq m - 1, \tag{1.46b}
\]

is called a multi-step method. The multi-step method is said to be explicit, if \( \Phi(x, z_0, \ldots, z_m; h) \) does not depend on \( z_m \), and implicit otherwise.

Using the grid function \( y_h : I_h \to \mathbb{R}^d \), the multi-step method (1.46a), (1.46b) can be written as a difference equation of \( m \)-th order:

\[
\frac{1}{h} \sum_{k=0}^{m} \alpha_k y_h(x + kh) = \Phi(x + mh, y_h(x), \ldots, y_h(x + mh); h), \tag{1.47a}
\]

\[
y_h(a + jh) = \alpha^{(j)}, \quad 0 \leq j \leq m - 1. \tag{1.47b}
\]

Example 1.44. (Milne-Simpson method)

For the scalar first order ordinary differential equation

\[ y'(x) = f(x, y(x)), \quad x \in [a, b], \]

integration over \([x_k, x_{k+2}] \subset [a, b]\) yields

\[ y(x_{k+2}) - y(x_k) = \int_{x_k}^{x_{k+2}} f(x, y(x)) \, dx. \]

Approximating the right-hand side by the Simpson rule, we obtain the Milne-Simpson rule

\[ y_{k+2} = y_k + \frac{h}{3} \left( f(x_k, y_k) + 4 f(x_{k+1}, y_{k+1}) + f(x_{k+2}, y_{k+2}) \right). \]
Definition 1.45. (Linear multi-step methods)
If there exist real numbers $\beta_0, \beta_1, \cdots, \beta_m$, such that the function $\Phi$ is given by
\[
\Phi(x, z_0, \cdots, z_m; h) = \sum_{k=0}^{m} \beta_k f(x_k, z_k), \quad x_m \in I'_h,
\]
the multi-step method (1.46a),(1.46b) is said to be a linear multi-step method.

Remark 1.46. (Start ramp)
For $j > 1$ the determination of the initial vectors $\alpha^{(j)}, 0 \leq j \leq m - 1$, can be done by a multi-step method of a lower step number. This is called the start ramp.

Definition 1.47. (Characteristic polynomials)
For a multi-step method (1.46a),(1.46b) the polynomial
\[
\rho(z) := \sum_{k=0}^{m} \alpha_k z^k
\]
is called the characteristic polynomial of the multi-step method. In case of a linear multi-step method, we can assign another characteristic polynomial associated with the increment function $\Phi$:
\[
\sigma(z) := \sum_{k=0}^{m} \beta_k z^k.
\]
1.3.2 Convergence, consistency, and stability

Definition 1.48. (Convergence and convergence order)
If \( y \in C^1([a,b]) \) is the exact solution of the initial-value problem (1.6a),(1.6b) and \( y_h : I_h \to \mathbb{R}^d \) is the approximation obtained by the multi-step method (1.46a),(1.46b), again we denote by \( e_h(x) := y_h(x) - y(x) \), \( x \in I_h \), the global discretization error. The solution \( y_h \) of the multi-step method (1.46a),(1.46b) is said to convergence to the solution \( y \) of the initial-value problem, if
\[
\max_{x \in I_h} \| e_h(x) \| \to 0 \quad (h \to 0).
\]
The multi-step method (1.46a),(1.46b) is said to convergence of order \( p > 0 \), if there exists a constant \( C > 0 \), independent of \( h \), such that
\[
\max_{x \in I_h} \| e_h(x) \| \leq C h^p \quad (h \to 0).
\]

As in the case of one-step methods, the convergence of multi-step methods can be characterized by consistency and stability.

Definition 1.49. (Consistency and order of consistency)
Let \( y \in C^1([a,b]) \) be the exact solution of the initial-value problem (1.6a),(1.6b). Then, the grid function
\[
\tau_h(x) := \frac{1}{h} \sum_{k=0}^{m} \alpha_k y(x - (m - k)h) - \Phi(x, y(x - mh), \ldots, y(x); h),
\]
\[
\tau_h(x_j) := y(x_j) - \alpha^{(j)}, \quad 0 \leq j \leq m - 1,
\]
is called the local discretization error of the multi-step method (1.46a), (1.46b).
The multi-step method is said to be consistent with the initial-value problem, if
\[
\max_{0 \leq j \leq m-1} |\tau_h(x_j)| + h \sum_{x \in I'_h} |\tau_h(x)| \to 0 \quad (h \to 0).
\]
The multi-step method has the order of consistency \( p > 0 \), if there exists a constant \( C > 0 \), independent of \( h \), such that
\[
\max_{0 \leq j \leq m-1} |\tau_h(x_j)| + h \sum_{x \in I'_h} |\tau_h(x)| \leq C h^p \quad (h \to 0).
\]
Theorem 1.50. (Consistency of multi-step methods)
The multi-step method \((1.46a),(1.46b)\) is consistent with the given initial value problem, if
\[
\alpha^{(j)} \to \alpha \quad (h \to 0), \quad 0 \leq j \leq m - 1,
\]
\[
(1.50b) \quad \rho(1)y = 0,
\]
\[
(1.50c) \quad h \sum_{x \in I'_h} |\Phi(x, y(x - mh), \cdots, y(x); h) - \\
\quad \rho'(1)f(x, y(x))| \to 0 \quad (h \to 0).
\]

Proof. The equivalence
\[
\alpha^{(j)} \to \alpha \quad (h \to 0) \iff |\tau_h(x_j)| \to 0 \quad (h \to 0), \quad 0 \leq j \leq m - 1,
\]
is obvious. Moreover, by Taylor expansion we obtain
\[
\tau_h(x) = \frac{1}{h} \sum_{k=0}^{m} \alpha_k y(x - mh) + \\
\sum_{k=1}^{m} k\alpha_k y'(x - mh) - \Phi(x, y(x - mh), \cdots, y(x); h) + O(h).
\]
Observing
\[
y(x - mh) = y(x) + O(h),
\]
\[
y'(x - mh) = y'(x) + O(h) = f(x, y(x)) + O(h),
\]
we see that \((1.50b)\) and \((1.50c)\) are sufficient for \(h \sum_{x \in I'_h} |\tau_h(x)| \to 0 \quad (h \to 0). \quad \square
\]

For linear multi-step methods the following result holds true:

Theorem 1.51. (Order of consistency of linear multi-step methods)
Let \(f \in C^p(I \times D)\) and assume
\[
\tau_h(x_j) = O(h^p), \quad 0 \leq j \leq m - 1.
\]
Then, for a linear multi-step method each of the following conditions is equivalent to the method having an order of consistency \(p\).

\(C_1:\) For each \(\ell = 0, 1, \cdots, p\) it holds
\[
\sum_{k=0}^{m} (k^\ell \alpha_k - \ell \int k^{\ell-1} \beta_k) = 0,
\]
C_2: z = 0 is a zero of order ≥ p + 1 of the entire function
\[ \varphi(z) := \rho(\exp(z)) - z \sigma(\exp(z)). \]

C_3: The initial-value problem
\[ y'(x) = y(x), \ x \in I, \ y(a) = 1, \]
has the order of consistency p.

C_4: The order of consistency is p for a class of initial-value problems whose solutions span the linear space of polynomials of degree p.

Proof. We first prove the equivalence of C_1 and C_4. We note that \( f \in C^p(I \times D) \) implies \( y \in C^{p+1}(I) \). Taylor expansion yields

\[ h \tau_h(x + mh) = \sum_{k=0}^{m} \left( \alpha_k y(x + kh) - h\beta_k \frac{f(x + kh, y(x + kh))}{y'(x + kh)} \right), \]

\[ y(x + kh) = y(x) + \sum_{\ell=1}^{p} \frac{1}{\ell!} k^\ell y^{(\ell)}(x) + O(h^{p+1}), \]

\[ y'(x + kh) = \sum_{\ell=1}^{p} \frac{1}{\ell!} k^\ell y^{(\ell+1)}(x) + O(h^p) = \]

\[ \sum_{\ell=1}^{p} \frac{1}{(\ell - 1)!} k^{\ell-1} h^{\ell-1} y^{(\ell)}(x) + O(h^p) \implies \]

\[ (1.51) \]

\[ h \tau_h(x + mh) = \]

\[ \sum_{k=0}^{m} \left( \alpha_k y(x) + \sum_{\ell=1}^{p} \frac{\alpha_k}{\ell!} k^\ell - \frac{\beta_k}{(\ell - 1)!} k^{\ell-1} h^{\ell-1} y^{(\ell)}(x) \right) + O(h^p). \]

It follows that

\[ C_1 \implies |\tau_h(x + mh)| = O(h^p). \]

Conversely, the consistency order p implies C_1, if in (1.51) we use the functions \( y = 1, \ y = x, \cdots, \ y = x^p. \)

The equivalence of C_4 can be shown similarly.

We next show the equivalence of C_3. If the linear multi-step method is consistent of order p, then C_3 obviously holds true. Conversely, assume that C_3 is satisfied. Inserting \( y(x) = \exp(x - a) \) into (1.51) and
summing over $I_h$ yields
\[
\frac{h}{m} \tau_h(x + mh) = O(h^p)
\]

\[
\sum_{\ell = 0}^{p} \frac{1}{\ell!} \left( \sum_{k=0}^{m} (k^\ell \alpha_k - \ell \ k^{\ell-1} \beta_k) \right) h^{\ell-1} \sum_{x + mh \in I_h} h \exp(x - a) + O(h^p) \rightarrow \int_{a}^{b} \exp(x-a) dx \quad (h \to 0)
\]

\[
= O(h^p)
\]

\[
\implies \sum_{k=0}^{m} (k^\ell \alpha_k - \ell \ k^{\ell-1} \beta_k) = 0, \quad 0 \leq \ell \leq p.
\]

Finally, we prove the equivalence of $C_2$. Taylor expansion of $\varphi(z)$ around $z = 0$ results in
\[
\varphi(z) = \rho(\exp(z)) - z \sigma(\exp(z)) = \sum_{\ell=0}^{p} \varphi^{(\ell)}(0) z^\ell + O(h^p) \implies
\]

\[
\varphi^{(\ell)}(0) = 0 \iff \sum_{k=0}^{m} (k^\ell \alpha_k - \ell \ k^{\ell-1} \beta_k) = 0, \quad 0 \leq \ell \leq p.
\]

\[
\square
\]

For the stability of the multi-step method (1.47a),(1.47b) we consider the linear space $C(I_h)$ of grid functions on $I_h$ equipped with the norm
\[
|||z_h||| := \sum_{j=0}^{m-1} |z_h(x_j)| + h \sum_{x \in I_h'} |z_h(x)|.
\]

The multi-step method define a mapping
\[
A_h : C(I_h) \to C(I_h)
\]

\[
A_h z_h(x_j) :=
\begin{cases}
  z_j - \alpha^{(j)}, & 0 \leq j \leq m - 1, \\
  \frac{1}{h} \sum_{k=0}^{m} \alpha_k z_{j-m+k} - \Phi(x_j, z_{j-m}, \ldots, z_j, h), & m \leq j \leq N
\end{cases}
\]

**Definition 1.52. (Local Lipschitz stability of multi-step methods)**

The multi-step method (1.47a),(1.47b) is said to be locally Lipschitz
stable in \( z_h \in C(I_h) \), if there exist positive numbers \( h_{\text{max}}, \delta, \) and \( \eta \) such that for all \( 0 < h < h_{\text{max}} \) and all \( w_h \in C(I_h) \) with \( \|A_hz_h - A_hw_h\| \leq \delta \) it holds
\[
\|z_h(x) - w_h(x)\| \leq \eta \|A_hz_h - A_hw_h\|, \quad x \in I_h.
\] (1.52)

The number \( \eta \) is called the stability barrier, whereas the number \( \delta \) is referred to as the stability threshold.

We suppose that the increment function \( \Phi \) satisfies the following Lipschitz condition:

There exist a neighborhood \( U \subset I \times \mathbb{R}^d \) of the graph \( (x, y(x)) \), \( x \in I \), of the solution \( y \in C^1(I) \) of the initial-value problem (1.6a),(1.6b) and numbers \( h_{\text{max}} > 0 \) and \( L > 0 \) such that for all \( (x, y_k), (x, y'_k) \in U \cap (I_h \times \mathbb{R}^d), 0 \leq k \leq m \), and for all \( 0 \leq h < h_{\text{max}} \) it holds
\[
\|\Phi(x, y_0, \cdots, y_m, h) - \Phi(x, y'_0, \cdots, y'_m, h)\| \leq L \sum_{k=0}^{m} \|y_k - y'_k\|. \tag{1.53}
\]

**Theorem 1.53. (Characterization of local Lipschitz stability)**

Assume that (1.53) holds true. Then, the multi-step method (1.47a), (1.47b) is locally Lipschitz stable in \( r_h y := y|_{I_h} \) if and only if the multi-step method with \( \Phi \equiv 0 \) is locally Lipschitz stable.

**Proof.** Without restriction of generality let \( U = I \times \mathbb{R}^d \). The local Lipschitz stability in \( r_h y \) is equivalent to the fact that for \( v_h := r_h y - w_h \) and \( h < h_{\text{max}} \) it holds
\[
\|v_h(x)\| \leq \eta \left( \sum_{t \in I_h, t \leq x} \|A_hr_h y - A_hw_h(t)\| + \sum_{j=0, t_j \leq x}^{m-1} \|v_h(t_j)\| \right). \tag{1.54}
\]

Now, let \( \Psi : I_h \times \mathbb{R}^{d(m+1)} \times \mathbb{R}_+ \to \mathbb{R}^d \) another increment function that satisfies (1.53) and let \( \tilde{A}_h \) the mapping associated with the increment function \( \Phi + \Psi \), i.e.,
\[
A_hz_h(x_j) = \tilde{A}_h z_h(x_j) + \Psi(x_j, z_{j-m}, \cdots, z_j, h). \tag{1.55}
\]

If we insert (1.55) into
\[
\|v_h(x)\| \leq \eta \left( \sum_{t \in I_h, t \leq x} \|A_hr_h y - A_hw_h(t)\| + \sum_{j=0, t_j \leq x}^{m-1} \|v_h(t_j)\| \right), \tag{1.56}
\]
for $x = x_\ell$, $0 \leq \ell \leq N$, we obtain

(1.57)

$$\|v_h(x_\ell)\| \leq \eta \left( h \sum_{t \in I_h} \| (A_h r_h y - A_h w_h)(t) \| + \sum_{j=0}^{m-1} \|v_h(t_j)\| \right) +$$

(1.58)

$$h \sum_{j=m}^\ell \left\| \Psi(x_j, (r_h y)_{j-m}, \ldots, (r_h y)_j, h) - \Psi(x_j, w_{j-m}, \ldots, w_j, h) \right\| \leq$$

(1.59)

$$\eta \left( \sum_{t \in I_h} \| (A_h r_h y - A_h w_h)(t) \| + \sum_{j=0}^{m-1} \|v_h(t_j)\| + h L \sum_{j=m}^\ell \sum_{k=0}^m \|v_h(t_{j-k})\| \right).$$

The second term on the right-hand side in (1.57) can be estimated by means of

(1.60)

$$\eta h L \sum_{j=m}^\ell b \sum_{k=0}^m \|v_h(t_{j-k})\| \leq \eta h L \|v_\ell\| + \eta (m+1) h L \sum_{k=0}^{\ell-1} \|v_k\|. $$

Inserting (1.60) into (1.57), for $\eta h_{\text{max}} L < 1$ it follows that

$$\|v_\ell\| \leq \frac{\eta}{1 - \eta h_{\text{max}} L} \left( \sigma + (m+1) L h \sum_{k=0}^{\ell-1} \|v_k\| \right).$$

An application of the discrete Gronwall’s lemma (see 1.23) yields

(1.61) $$\|v_h(x)\| \leq \eta' \sigma \exp(\eta'(m+1)L(x-a)), \quad x \in I_h.$$  

The equivalence follows from (1.61) by choosing $\Psi = -\Phi$ (necessity) and $\Psi = \Phi$ (sufficiency).

□

By using tools from the theory of difference equations and complex analysis the local Lipschitz stability can be characterized algebraically in terms of Dahlquist’s root condition:

**Definition 1.54. (Dahlquist’s root condition)**

Dahlquist’s root condition holds true, if the roots $\xi_j$, $1 \leq j \leq m$, of the characteristic polynomial $\rho = \rho(z)$ satisfy the condition: $|\xi_j| \leq 1$ and roots with $|\xi_j| = 1$ are simple.
Theorem 1.55. (Algebraic characterization of Lipschitz stability)

Under the assumption (1.53), Dahlquist’s root condition is necessary and sufficient for the local Lipschitz stability of the multi-step method (1.47a),(1.47b) in \( r_h y \).

As in the case of one-step methods, consistency and stability imply convergence.

Theorem 1.56. (Convergence of multi-step methods)

Under the Lipschitz condition (1.53) assume that the multi-step method (1.47a), (1.47b) is consistent with the initial-value problem (1.6a),(1.6b) and Lipschitz stable in \( r_h y \) with the stability barrier \( \eta > 0 \). Then, there exist numbers \( h_{\text{max}} > 0 \) and \( \delta > 0 \) such that for all \( 0 \leq h < h_{\text{max}} \) the equation \( A_h y_h = g_h \) is uniquely solvable in \( U \) for all \( g_h \in C(I_h) \) with \( |||g_h||| < \delta \) and the a priori estimate

\[
(1.62) \quad \max_{x \in I_h} ||(y_h - y)(x)|| \leq \eta \left( |||g_h||| + |||\tau_h||| \right)
\]

holds true. In particular, for \( g_h \equiv 0 \) we have

\[
\max_{x \in I_h} ||y_h(x) - y(x)|| \to 0 \quad (h \to 0).
\]

The order of convergence corresponds to the order of consistency.

Proof. Without restriction of generality we assume \( U = I \times \mathbb{R}^d \). In view of the Lipschitz continuity of the increment function \( \Phi \), the Banach fixed point theorem implies the existence of a solution \( y_h \in C(I_h) \). Choosing \( z_h := y_h \) and \( w_h := r_h y \), the Lipschitz stability with stability barrier \( \eta \) implies that for \( 0 \leq h < h_{\text{max}} \)

\[
(1.63) \quad ||y_h(x) - y(x)|| \leq \eta |||\tau_h|||, \quad x \in I_h.
\]

The consistency yields \( |||\tau_h||| \to 0(h \to 0) \) so that \( y_h \) is uniquely determined for sufficiently small \( h \) and \( \delta \). The estimate (1.62) follows readily from (1.63). \( \square \)

There is an upper bound on the achievable order of convergence of a locally Lipschitz stable linear multi-step method, known as Dahlquist’s first barrier.
Theorem 1.57. (Dahlquist’s first barrier)

For the order of convergence $p$ of a locally Lipschitz stable linear multi-step method it holds

$$p \leq \begin{cases} 
  m + 2 & , \text{ if } m \text{ is even,} \\
  m + 1 & , \text{ if } m \text{ is odd,} \\
  m & , \text{ if it is explicit}
\end{cases}$$

Stable multi-step methods of order $m + 2$ are symmetric, i.e.,

$$\alpha_k = -\alpha_{m-k}, \quad 0 \leq k \leq m,$$
$$\beta_k = \beta_{m-k}, \quad 0 \leq k \leq m.$$
1.3.3 Extrapolatory and interpolatory multi-step methods

Extrapolatory and interpolatory linear multi-step methods can be constructed by numerical quadrature: The idea is to integrate \( y'(x) = f(x, y(x)) \) over \([x^*, x + mh]\)

\[
y(x + mh) = y(x^*) + \int_{x^*}^{x + mh} f(s, y(s)) \, ds
\]

and to replace the integrand by an interpolating polynomial of degree \( r \) with respect to the interpolation points

\[(x + jh, f(x + jh, y(x + jh))), \quad 0 \leq j \leq r,
\]

where \( r = m \) (interpolation) or \( r = m - 1 \) (extrapolation):

<table>
<thead>
<tr>
<th>Type</th>
<th>Name</th>
<th>( r )</th>
<th>( x^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extrapolatory</td>
<td>Adams-Bashforth</td>
<td>( m - 1 )</td>
<td>( x + (m - 1)h )</td>
</tr>
<tr>
<td>method</td>
<td>Nyström</td>
<td>( m - 1 )</td>
<td>( x + (m - 2)h )</td>
</tr>
<tr>
<td>Interpolatory</td>
<td>Adams-Moulton</td>
<td>( m )</td>
<td>( x + (m - 1)h )</td>
</tr>
<tr>
<td>method</td>
<td>Milne-Simpson</td>
<td>( m )</td>
<td>( x + (m - 2)h )</td>
</tr>
</tbody>
</table>

(i): Adams-Bashforth method

The extrapolatory Adams-Bashforth method reads

\[
y_{j+m} = y_{j+m-1} + \int_{x_{j+m-1}}^{x_{j+m}} p_{m-1}(s) \, ds,
\]

where \( p_{m-1} \in P_{m-1}(\mathbb{R}) \) is the interpolating polynomial with \( p_{m-1}(x_{j+k}) = f_{j+k}, 0 \leq k \leq m - 1 \) (cf. Figure 5).

![Figure 5. Adams-Bashforth method](image-url)
Newton’s representation of the interpolating polynomial is as follows:

\[ p_{m-1}(s) = \sum_{k=0}^{m-1} (-1)^k \binom{-1}{k} \nabla^k f_{j+m-1}, \]

where \( t = \frac{1}{h} (s - x_{j+m-1}) \) and \( \nabla^k f = \nabla^{k-1} f - \nabla^{k-1} f_{\ell-1} \). Inserting (1.65) into (1.64) yields

\[ y_{j+m} = y_{j+m-1} + \sum_{k=0}^{m-1} (-1)^k \int_{x_{j+m-1}}^{x_j+m} \binom{-1}{k} ds \nabla^k f_{j+m-1}. \]  

Setting

\[ \gamma_k := (-1)^k \frac{1}{h} \int_{x_{j+m-1}}^{x_j+m} \binom{-1}{k} ds \]

from (1.66) we obtain the Adams-Bashforth method in the form

\[ y_{j+m} = y_{j+m-1} + h \sum_{k=0}^{m-1} \gamma_k \nabla^k f_{j+m-1}, \quad 0 \leq j \leq N - m, \]

\[ y_k = \alpha^{(k)}, \quad 0 \leq k \leq m - 1. \]

**Lemma 1.58. (Recursion formula for the Adams-Bashforth method)**

The coefficients \( \gamma_k, 0 \leq k \leq m - 1 \), of the Adams-Bashforth method (1.67a),(1.67b) can be computed by means of the recursion formula

\[ \gamma_k + \frac{1}{2} \gamma_{k-1} + \frac{1}{3} \gamma_{k-2} + \cdots + \frac{1}{k+1} \gamma_0 = 1. \]

**Proof.** The proof follows from the relationship

\[ \sum_{k=0}^{\infty} (-1)^k \int_{0}^{1} \binom{-1}{k} dt z^k = -\frac{(1 - z)^{-t}}{\log(1 - z)} \bigg|_{t=0}. \]

**Theorem 1.59. (Convergence of the Adams-Bashforth method)**

Assume \( y \in C^m(I) \) and suppose that \( f : I \times D \to \mathbb{R}^d \) satisfies a Lipschitz condition. Then, the Adams-Bashforth method (1.67a),(1.67b) is locally Lipschitz stable and consistent of order \( m \). The order of convergence corresponds to the order of consistency.
An equivalent formulation of the Adams-Bashforth method can be derived by an evaluation of the differences $\nabla^k f_{j+m-1}$. It reads

\begin{equation}
y_{j+m} = y_{j+m-1} + h \sum_{k=0}^{m-1} \alpha_{mk} f_{j+k}, \quad 0 \leq j \leq N - m, \tag{1.68a}
\end{equation}

\begin{equation}
y_k = \alpha^{(k)}, \quad 0 \leq k \leq m - 1, \tag{1.68b}
\end{equation}

where the coefficients $\alpha_{mk}, 0 \leq k \leq m - 1$, are given by

$$\alpha_{mk} = \sum_{\ell=m-k-1}^{m-1} \gamma^\ell (-1)^{m-k-1} \binom{\ell}{m-k-1}.$$ 

(ii): Adams-Moulton method

The interpolatory Adams-Moulton method reads

\begin{equation}
y_{j+m} = y_{j+m-1} + \int_{x_{j+m-1}} x_{j+m} p_m(s) \, ds \tag{1.69}
\end{equation}

where $p_m \in P_m(\mathbb{R})$ is the interpolating polynomial with $p_m(x_{j+k}) = f_{j+k}, 0 \leq k \leq m$ (cf. Figure 6).

Here, Newton’s representation of the interpolating polynomial reads

\begin{equation}
p_m(s) = \sum_{k=0}^{m} (-1)^k \binom{-t}{k} \nabla^k f_{j+m}. \tag{1.70}
\end{equation}
Inserting (1.70) into (1.69) yields
\[(1.71a)\quad y_{j+m} = y_{j+m-1} + h \sum_{k=0}^{m} \nu_k \nabla^k f_{j+m}, 0 \leq j \leq N - m,\]
\[(1.71b)\quad y_k = \alpha(k), \ 0 \leq k \leq m - 1,\]
\[(1.71c)\quad \nu_k := (-1)^k \int_{-1}^{0} \frac{(-k)}{\ell_{k}} \, dt, \ 0 \leq k \leq m.\]

**Lemma 1.60. (Recursion formula for the Adams-Moulton method)**

The coefficients \(\nu_k, 0 \leq k \leq m,\) of the Adams-Moulton method (1.71a)-(1.71c) satisfy the recursion formula
\[\nu_k + \frac{1}{2} \nu_{k-1} + \frac{1}{3} \nu_{k-2} + \cdots + \frac{1}{k+1} \nu_0 = 0.\]

**Theorem 1.61. (Convergence of the Adams-Moulton method)**

Assume \(y \in C^{m+1}(I)\) and suppose that \(f : I \times D \to \mathbb{R}^d\) satisfies a Lipschitz condition. Then, for \(h\) sufficiently small the Adams-Moulton method (1.71a)-(1.71c) is well defined. It is locally Lipschitz stable and consistent of order \(m + 1.\) The order of convergence corresponds to the order of consistency.

By an evaluation of the differences \(\nabla^k f_{j+m}\) we obtain the following equivalent formulation of the Adams-Moulton method
\[(1.72a)\quad y_{j+m} = y_{j+m-1} + h \sum_{k=0}^{m} \beta_{mk} f_{j+k}, 0 \leq j \leq N - m,\]
\[(1.72b)\quad y_k = \alpha(k), \ 0 \leq k \leq m - 1,\]
where the coefficients \(\beta_{mk}, 0 \leq k \leq m,\) are given by
\[\beta_{mk} := (-1)^{m-k} \sum_{\ell=m-k}^{m} \nu_\ell \left(\frac{\ell}{m-k}\right), 0 \leq k \leq m.\]

**Remark 1.62. (Nyström and Milne-Simpson method)**

Nyström’s method and the method of Milne-Simpson can be derived similarly.
1.3.4 Predictor-corrector methods

Extrapolatory (explicit) multi-step methods have the advantage that they are computationally cheap and the disadvantage that the order of convergence is lower than corresponding interpolatory (implicit) methods. On the other hand, for interpolatory (implicit) methods the advantage of a higher order of convergence is marred by the fact that they are computationally more expensive. Predictor-corrector methods combine the advantages and avoid the disadvantages.

The idea is to compute an approximate solution based on an interpolatory multi-step method by fixed point iteration and to use a start iterate computed by an extrapolatory multi-step method:

(i) P: Predictor (e.g. Adams-Bashforth method)

\[ y_{j+m}^{(0)} = y_{j+m-1} + h \sum_{k=0}^{m-1} \alpha_{mk} f_{j+k}. \]

(ii) E: Evaluate

\[ f_{j+m}^{(0)} := f(x_{j+m}, y_{j+m}^{(0)}). \]

(iii) C: Corrector (e.g. Adams-Moulton method)

\[ y_{j+m}^{(1)} = y_{j+m-1} + h \beta_{mm} f_{j+m}^{(0)} + h \sum_{k=0}^{m-1} \beta_{mk} f_{j+k}. \]

Variants of the predictor-corrector methods

I. \( P(EC)^\ell E \) method

(a) \( y_{j+m}^{(0)} = y_{j+m-1}^{(\ell)} + h \sum_{k=0}^{m-1} \alpha_{mk} f_{j+k}^{(\ell)} \)

\[ f_{j+m}^{(r-1)} = f(x_{j+m}, y_{j+m}^{(r-1)}), \]

(b) \( y_{j+m}^{(r)} = y_{j+m-1} + h \beta_{mm} f_{j+m}^{(r-1)} + h \sum_{k=0}^{m-1} \beta_{mk} f_{j+k}^{(\ell)} \) \( 1 \leq r \leq \ell \),

(c) \( f_{j+m}^{(\ell)} = f(x_{j+m}, y_{j+m}^{(\ell)}). \)
II. $P(\text{EC})^\ell$ method

(a) $y_{j+m}^{(0)} = y_{j+m-1}^{(\ell)} + h \sum_{k=0}^{m-1} \alpha_{mk} f_{j+k}^{(\ell-1)}$

$$f_{j+m}^{(r-1)} = f(x_{j+m}, y_{j+m}^{(r-1)})$$

(b) $y_{j+m}^{(r)} = y_{j+m-1}^{(\ell)} + h \beta_{mn} f_{j+m}^{(r-1)} + h \sum_{k=0}^{m-1} \beta_{mk} f_{j+k}^{(\ell-1)}$

$$1 \leq r \leq \ell.$$
1.3.5 BDF (Backward Difference Formulas)

The idea behind the Backward Difference Formulas (BDF) is to use the interpolating polynomial

\[ p_m(x_{j+k}) = y_{j+k}, \quad 0 \leq k \leq m, \]

for the approximation of \( y'(x) = f(x, y(x)) \) in \( x_{j+m-r}, r = 0 \) or \( r = 1 \):

\[ p'_m(x_{j+m-r}) = f(x_{j+m-r}, y_{j+m-r}). \]

Newton’s representation of the interpolating polynomial

\[ p_m(x) = \sum_{k=0}^m (-1)^k \binom{x}{k} \nabla^k y_{j+m}, \quad t := (x - x_{j+m})/h \]

yields

\[ p'_m(x_{j+m-r}) = \frac{1}{h} \sum_{k=0}^m \rho_{rk} \nabla^k y_{j+m} = f(x_{j+m-r}, y_{j+m-r}), \]

(1.73a)

\[ \rho_{rk} = (-1)^k \frac{d}{dt} \binom{t}{k} \big|_{t=-r}, \quad 0 \leq k \leq m. \]

(1.73b)

Remark 1.63. (Implicit/explicit BDF)

For \( r = 0 \) we obtain an implicit method, whereas \( r = 1 \) yields an explicit method. For \( r = 0 \) it follows that

\[ \rho_{00} = 0, \quad \rho_{0k} = \frac{1}{k}, \quad 1 \leq k \leq m. \]
1.4 Numerical integration of stiff systems

1.4.1 Linear stability theory

For linear systems \( y' = Ay \) the growth behavior of the solutions is determined by the eigenvalues \( \lambda \in \sigma(A) \), where \( \sigma(A) \) stands for the spectrum of the matrix \( A \). For nonlinear autonomous systems \( y' = f(y) \) local information about the growth behavior, i.e., with respect to a function \( y^* \in C(I) \), can be obtained by means of the eigenvalues of the Jacobian \( J_f(y^*) := f_y(y^*) \).

We consider the autonomous initial-value problem

\[
\begin{align*}
\tag{1.74a} y'(x) & = f(y(x)), \quad x \in I, \\
\tag{1.74b} y(a) & = \alpha.
\end{align*}
\]

Now, if \( B \) is a regular \( d \times d \) matrix, the affine transformation

\[ y \mapsto By =: \bar{y} \]

leads to the transformed initial-value problem

\[
\begin{align*}
\tag{1.75a} \bar{y}'(x) & = \bar{f}(\bar{y}(x)), \quad x \in I, \\
\tag{1.75b} \bar{y}(a) & = B\alpha,
\end{align*}
\]

where \( \bar{f}(\bar{y}) := Bf(B^{-1}\bar{y}) \). The theorem of Picard-Lindelöf implies the relationship

\[ \bar{y} = By, \]

which is referred to as the affine covariance of the autonomous system.

**Definition 1.64. (Affine covariance of numerical integrators)**

A numerical integrator for the approximate solution of an autonomous first order system is said to be affine covariant, if the following holds true:

If \( B \) is a regular \( d \times d \) matrix and \( y_h \in C(I_h) \) is an approximation computed with the integrator, then \( \bar{y}_h = By_h \) is an approximation of \( \bar{y} = By \).

Now, for the initial-value problem (1.74b), the eigenvalues of the Jacobian \( f_y \) determine the local stability. The spectrum of the Jacobian is invariant with respect to similarity transformations

\[ f_y \mapsto B f_y B^{-1}. \]

If the matrix \( A := f_y \) is diagonalizable, there exist a regular matrix \( B \) such that

\[ B A B^{-1} = \text{diag}(\lambda_1, \ldots, \lambda_n). \]
where $\lambda_i \in \sigma(A), 1 \leq i \leq d$. We conclude that due to the affine covariance it suffices to consider the scalar test problem

\begin{align}
(1.76a) & \quad y'(x) = \lambda y(x), \quad x > 0, \\
(1.76b) & \quad y(0) = 1 \quad (\lambda \in \mathbb{C}),
\end{align}

whose solution is given by $y(x) = \exp(\lambda x), x \geq 0$. 

1.4.2 A-stability, L-stability, and a(α)-stability

One-step methods for the numerical integration of the test problem (1.76a),(1.76b) give rise to rational functions

\[ y_1 = R_{\ell m}(z) \quad y_0 = \frac{P_\ell(z)}{Q_m(z)} \quad y_0, \]

where \( R_{\ell m}(z) \) is an approximation of \( \exp(z) \).

**Definition 1.65. (Padé approximation of the exponential function)**

Let \( \ell, m \in \mathbb{N}_0 \) and \( P_\ell, Q_m \) polynomials of degree \( \ell \) and \( m \). Assume that

\[ \exp(z) Q_m(z) - P_\ell(z) = O(|z|^\ell+m+1) \quad (z \to 0). \]

Then, the rational function

\[ R_{\ell m}(z) = \frac{P_\ell(z)}{Q_m(z)} \]

is called a Padé approximation of \( \exp(z) \) of index \((m, \ell)\).

**Definition 1.66. (Stability region)**

The stability region of a one-step method with \( R = R_{\ell m} \) is given by

\[ G := \{ z \in \mathbb{C} \mid |R(z)| \leq 1 \}. \]

**Example 1.67. (Stability region of the explicit/implicit Euler method and of the implicit trapezoidal rule)**

For the explicit Euler method we have

\[ R_{EE}(z) = 1 + z. \]

Hence, the stability region is given by

(1.77) \[ G_{EE} := \{ z = (z_1, z_2) \in \mathbb{C} \mid (z_1 + 1)^2 + z_2^2 \leq 1 \}, \]

i.e., \( G_{EE} \) is the circle in the complex plane with center \((-1, 0)\) and radius 1.

For the implicit Euler method it holds

\[ R_{IE}(z) = \frac{1}{1 - z}. \]

Consequently, the stability region is given by

(1.78) \[ G_{IE} := \{ z = (z_1, z_2) \in \mathbb{C} \mid (z_1 - 1)^2 + z_2^2 \geq 1 \}, \]
i.e., $G_{IE}$ is the complement of the interior of the circle in the complex plane with center $(1, 0)$ and radius 1. The implicit trapezoidal rule gives rise to

$$R_{IT}(z) = \frac{1 + z/2}{1 - z/2}.$$  

It follows that the stability region is given by

$$(1.79) \quad G_{IT} := \{ z = (z_1, z_2) \in \mathbb{C} \mid z_1 \leq 0 \},$$

i.e., $G_{IT}$ coincides with the left part $\mathbb{C}_-$ of the complex plane.

Dahlquist has required that decreasing analytical solutions should be approximated by numerical integrators whose solutions are also decreasing. This immediately leads to the notion of A-stability.

**Definition 1.68. (A-stability)**

A one-step method with Padé approximation $R = R(z)$ and stability region $G$ is called A-stable, if

$$\mathbb{C}_- \subseteq G.$$  

It follows from Example 1.67 that the implicit Euler method and the implicit trapezoidal rule are A-stable, but the explicit Euler method is not.

**Lemma 1.69. (Ehle’s conjecture)**

One-step methods with Padé approximations of index $(m, \ell)$, $\ell \leq m \leq \ell + 2$, (diagonal and subdiagonal Padé approximations) are A-stable.

**Proof.** For the proof of Ehle’s conjecture we refer to Hairer and Wanner. □

The implicit trapezoidal rule is A-stable, but $R_{IT}(z) \to -1$ for $\text{Re}(z) \to -\infty$. This motivates the notion of L-stability.

**Definition 1.70. (L-stability)**

An A-stable one-step method with Padé approximation $R = R(z)$ is called L-stable, if

$$R(z) \to 0 \quad \text{for} \quad \text{Re}(z) \to -\infty.$$  

The implicit Euler method is L-stable, but the trapezoidal rule is not.
Lemma 1.71. (L-stability of subdiagonal Padé approximations)

A-stable one-step methods with Padé approximations of index \((\ell+1, \ell)\) are L-stable.

**Proof.** We refer to Hairer and Wanner. \qed

**Definition 1.72. (Super-stability)**

An L-stable one-step method with stability region \(G\) is called super-stable, if

\[
G \setminus \mathbb{C}_- \neq \emptyset.
\]

The implicit Euler method is super-stable.

**Definition 1.73. (A(\alpha)-stability)**

A numerical integrator is called A(\alpha)-stable, if

\[
G = G(\alpha) = \{ z \in \mathbb{C} \mid |\arg z - \pi| \leq \alpha \}.
\]
1.4.3 Implicit and semi-implicit Runge-Kutta methods

Definition 1.74. (Implicit s-stage Runge-Kutta method)
An implicit s-stage Runge-Kutta method is given as follows: Given $a_{ij} \in \mathbb{R}, 1 \leq i, j \leq s,$ and $b_i \in \mathbb{R}, 1 \leq i \leq s,$ we set $c_i := \sum_{j=1}^{s} a_{ij}, 1 \leq i \leq s,$ and define

\[
k_1 = h \, f(x_k + c_1 \, h, y_k + \sum_{j=1}^{s} a_{1j} \, k_j),
\]

\[
\vdots \quad \vdots
\]

\[
k_s = h \, f(x_k + c_s \, h, y_k + \sum_{j=1}^{s} a_{sj} \, k_j),
\]

\[
y_{k+1} = y_k + \sum_{j=1}^{s} b_j \, k_j.
\]

The Butcher scheme is given by

\[
c_1 \begin{array}{cccc}
  a_{11} & a_{12} & \cdots & a_{1,s-1} & a_{1,s} \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  a_{s1} & a_{s2} & \cdots & a_{s,s-1} & a_{ss} \\
\end{array}
\begin{array}{c}
b_1 \\
b_2 \\
b_{s-1} \\
b_s
\end{array}
\]

Implicit Runge-Kutta methods require the solution of a nonlinear system: Setting $k_i = h \, f(x_k + c_i h, g_i)$ it follows that

\[
g_i = y_k + h \sum_{j=1}^{s} a_{ij} \, f(x_k + c_j \, h, g_j), \quad 1 \leq i \leq s.
\]

The application to the linear test problem (1.76a),(1.76b) yields

\[
\begin{pmatrix}
g_1 \\
\vdots \\
g_s
\end{pmatrix} = \begin{pmatrix}
1 \\
\vdots \\
1
\end{pmatrix} \begin{pmatrix}
y_k + h \lambda A \begin{pmatrix}
g_1 \\
\vdots \\
g_s
\end{pmatrix}
\end{pmatrix} \Rightarrow \begin{pmatrix}
(I_s - z \, A) g = e \\
k = z \, g
\end{pmatrix},
\]

where $A := (a_{ij})_{i,j=1}^{s}$ and $I_s$ stands for the $s \times s$ unit matrix. If $I_s - zA$ is regular, we have

\[
y_{k+1} = y_k + b^T k = \left(1 + z \, b^T \, (I_s - z \, A)^{-1} \, e \right) \, y_k.
\]
From (1.80) we deduce the stability function

\[ R_s(z) = 1 + b^T \left( \frac{1}{z} I_s - A \right)^{-1} e. \]

If \( A \) is regular, we have

\[ R_s(\infty) = 1 - b^T A^{-1} e. \]

In order to ensure L-stability we must have \( R_s(\infty) = 0 \).

**Example 1.75. (Fehlberg trick for implicit Runge-Kutta methods)**

The Fehlberg trick for implicit Runge-Kutta methods reads

\[ a_{sj} = b_j, \quad 1 \leq j \leq s \implies \]

\[ c_s = \sum_{j=1}^{s} a_{sj} = \sum_{j=1}^{s} b_j = 1 \implies \]

\[ b^T = e_s^T A, \quad \text{where} \quad e_s := (0, \ldots, 0, 1)^T \implies \]

\[ R_s(\infty) = 1 - e_s^T A A^{-1} e = 0. \]

We now consider the solution of the nonlinear system

\[ G(g_1, \ldots, g_s) = \begin{cases} 
  g_1 - y_k - h \sum_{j=1}^{s} a_{1j} f(x_k + c_1 h, g_j) \\
  g_s - y_k - h \sum_{j=1}^{s} a_{sj} f(x_k + c_s h, g_j) 
\end{cases} = 0 \]

by a simplified Newton method. For the Jacobi matrix in \( g_i = y_k \) we obtain

\[ \frac{\partial G_{i \ell}}{\partial y_{k}} \bigg|_{y_k = y_k} = \delta_{i \ell} I_d - h a_{i \ell} f_y(x_k + c_i h, g_k) \bigg|_{y=y_k} = \]

\[ \delta_{i \ell} I_d - h a_{i \ell} f_y(x_k, y_k) + O(h^2). \]

It follows that

\[ J = \begin{pmatrix} 
  I_d - ha_{11} A & -ha_{12} A & \cdots & -ha_{1s} A \\
  -ha_{21} A & I_d - ha_{22} A & \cdots & -ha_{2s} A \\
  \vdots & \vdots & \ddots & \vdots \\
  -ha_{s1} A & -ha_{s2} A & \cdots & I_d - ha_{ss} A
\end{pmatrix}. \]
Since $J|_{h=0} = I$, for sufficiently small $h > 0$ the Jacobi matrix $J$ is regular.

The simplified Newton method is as follows: Choose $g^{(0)} = (g_1^{(0)}, \ldots, g_s^{(0)})^T$ with $g_i^{(0)} = y_k, 1 \leq i \leq s$. For $k = 0, 1, 2, \cdots$ compute

$$J \Delta g^{(k)} = -G(g^{(k)}),$$

$$g^{(k+1)} = g^{(k)} + \Delta g^{(k)}.$$

We consider the following simplifications of implicit Runge-Kutta methods:

**Definition 1.76.** (Diagonally Implicit Runge-Kutta (DIRK) method)

The Butcher scheme of a Diagonally Implicit Runge-Kutta (DIRK) method is given by

$$
\begin{array}{c|cccc}
  c_1 & a_{11} & 0 & 0 & \cdots & 0 \\
  c_2 & a_{21} & a_{22} & 0 & \cdots & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  c_s & a_{s1} & a_{s2} & a_{s3} & \cdots & a_{ss} \\
  b_1 & b_2 & b_3 & \cdots & b_s \\
\end{array}
$$

The implementation of DIRK methods requires the LR-decomposition of $s$ matrices of the form

$$I_d - h a_{ii} A = L_i R_i, \quad 1 \leq i \leq s.$$

**Definition 1.77.** (Singly Diagonally Implicit Runge-Kutta (SDIRK) method)

The Butcher scheme of a Singly Diagonally Implicit Runge-Kutta (SDIRK) method is given by

$$
\begin{array}{c|cccc}
  c_1 & \gamma & 0 & 0 & \cdots & 0 \\
  c_2 & a_{21} & \gamma & 0 & \cdots & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  c_s & a_{s1} & a_{s2} & a_{s3} & \cdots & \gamma \\
  b_1 & b_2 & b_3 & \cdots & b_s \\
\end{array}
$$

The implementation of SDIRK methods requires the LR-decomposition of only one matrix of the form

$$I_d - h \gamma A = L R.$$
**Definition 1.78. (Semi-implicit Runge-Kutta methods)**

Semi-implicit Runge-Kutta methods are characterized by the implementation of only one Newton step

\[
J \Delta g^{(0)} = - G(g^{(0)}), \\
g^{(1)} = g^{(0)} + \Delta g^{(0)}.
\]

For Rosenbrock methods and W-methods we refer to Hairer and Wanner.
1.4.4 Linear multi-step methods and Dahlquist’s second barrier

We consider the linear multi-step method

\[ \sum_{k=0}^{m} \alpha_k y_{j+k} = h \sum_{k=0}^{m} \beta_k f(x_{j+k}, y_{j+k}) \]

with the characteristic polynomials

\[ \rho(z) = \sum_{k=0}^{m} \alpha_k z^k, \quad \sigma(z) = \sum_{k=0}^{m} \beta_k z^k. \]

The application to the scalar test problem (1.76a), (1.76b) yields (observe \( z = \lambda h \)):

\[ \sum_{k=0}^{m} (\alpha_k - \beta_k z) y_{j+k} = 0. \]

The ansatz \( y_\ell = z^\ell \) gives rise to the characteristic equation

\[ \rho(z) - z \sigma(z) = 0. \]

Lemma 1.79. (Stability region of linear multi-step methods)

Let \( \xi_i = \xi_i(z), 1 \leq i \leq m, \) be the roots of the characteristic equation (1.82). Then, the stability region of the linear multi-step method (1.81) is given by

\[ G = \{ z \in \mathbb{C} \mid |\xi_i(z)| \leq 1, 1 \leq i \leq m, \text{ and } \xi_i(z) \text{ simple, if } |\xi_i(z)| = 1 \}. \]

Definition 1.80. (Root locus)

Parametrization of \( \partial G \) according to

\[ |\xi| = 1 \rightarrow \xi = \exp(i \varphi), \quad \varphi \in [0, 2\pi), \]

defines the curve

\[ \Gamma := \{ z \in \mathbb{C} \mid z = \frac{\rho(\exp(i\varphi))}{\sigma(\exp(i\varphi))}, \varphi \in [0, 2\pi) \}, \]

which is called the root locus.

Remark 1.81. (Stability region and root locus)

Since \( \partial G \) only contains simple roots, it holds

\[ \partial G \subset \Gamma. \]
Example 1.82. (Explicit midpoint rule)
The application of the explicit midpoint rule to the test problem (1.76a),(1.76b) yields
\[ y_{k+2} = y_k + 2h \lambda y_{k+1} = y_k + 2z y_{k+1}, \]
and hence, the characteristic polynomials are given by
\[ \rho(\xi) = \xi^2 - 1, \quad \sigma(\xi) = 2. \]
The characteristic equation reads
\[ \xi^2 - 2z \xi - 1 = 0, \]
and has the zeroes
\[ \xi_{1,2} = z \pm \sqrt{1+z^2}. \]
Consequently, the root locus is given by
\[
\begin{align*}
z &= \frac{\rho(\exp(i\varphi))}{\sigma(\exp(i\varphi))} = \frac{\exp(2i\varphi) - 1}{2 \exp(i\varphi)} = \\
&= \frac{1}{2} \left( \exp(i\varphi) - \exp(-i\varphi) \right) = i \sin \varphi \quad \Rightarrow \\
\Gamma &= \{ z \in \mathbb{C} \mid z = i\tau, \tau \in [-1, +1] \}.
\end{align*}
\]
The stability region turns out to be
\[ G = \partial G = \{ z \in \mathbb{C} \mid z = i\tau, \tau \in (-1, +1) \}. \]
Since \( \mathbb{C}_- \not\subset G \), the explicit midpoint rule is not A-stable.

The second Dahlquist barrier limits the order of A-stable linear multi-step methods:

Theorem 1.83. (Second Dahlquist barrier)
A consistent A-stable linear multi-step method is implicit and has an order of consistency \( p \leq 2 \) with the error constant \( C^* := C/\sigma(1) \leq -\frac{1}{12} \). The implicit trapezoidal rule is the only A-stable method with \( p = 2 \) and \( C^* = -\frac{1}{12} \).
1.4.5 A(α)-stability of implicit BDF

The implicit BDF

\[ \sum_{k=1}^{m} \frac{1}{k} \nabla^k y_{j+m} = h f(x_{j+m}, y_{j+m}), \quad 0 \leq j \leq N - m, \]

have the following stability properties:

(i) They are A-stable for \( m = 1, 2 \) (\( m = 1 \) corresponds to the implicit Euler method).

(ii) They are A(α)-stable for \( 3 \leq m \leq 6 \) with

\[
\begin{array}{c|cccc}
 m & 3 & 4 & 5 & 6 \\
\hline
\alpha & 86^\circ & 76^\circ & 50^\circ & 16^\circ \\
\end{array}
\]

(iii) They are unstable for \( m \geq 6 \).
1.4.6 Numerical solution of differential-algebraic systems

Given \( F : D_1 \times D_2 \times I \to \mathbb{R}^d, D_i \subset \mathbb{R}^d, I := [a, b] \subset \mathbb{R}, \) and \( \alpha \in \mathbb{R}^d, \)
we consider the following initial-value problem for an implicit system of first order ordinary differential equations

\[
\begin{align*}
(1.83a) & \quad F(y'(x), y(x), x) = 0, \quad x \in I, \\
(1.83b) & \quad y(a) = \alpha.
\end{align*}
\]

The numerical treatment of (1.83a),(1.83b) significantly depends on the notion of the (differential) index which is due to Gear.

**Definition 1.84. (Differential index of an implicit first order system)**

Consider the system obtained from (1.83a) by differentiation

\[
(1.84) \quad F(y', y, x) = 0,
\]

\[
\frac{d}{dx} F(y', y, x) = \frac{\partial F}{\partial y'} y'' + \frac{\partial F}{\partial y} y' + \frac{\partial F}{\partial x} = 0,
\]

\[
\frac{d^\mu}{dx^\mu} F(y', y, x) = \frac{\partial F}{\partial y'} y^{(\mu+1)} + \cdots = 0.
\]

The smallest number \( \mu \in \mathbb{N}_0, \) for which (1.84) can be explicitly solved for \( y' \), is called the differential index of (1.83a),(1.83b) and will be denoted by \( \nu. \)

If (1.83a),(1.83b) has the differential index \( \nu, \) then the initial values \( y_a \) are said to be consistent, if

\[
\begin{align*}
F(y', y, x)|_{y=y(a), x=a} &= 0, \\
\frac{d^\nu}{dx^\nu} F(y', y, x)|_{y=y(a), x=a} &= 0.
\end{align*}
\]

A particular case occurs when the equations can be split into a differential part and a purely algebraic part:

**Definition 1.85. (Differential-algebraic system)**

Let \( I := [a, b] \subset \mathbb{R}, D_i \subset \mathbb{R}^d, 1 \leq i \leq 3, \) and

\[
F_1 : D_1 \times D_2 \times D_3 \times I \to \mathbb{R}^d, \quad F_2 : D_2 \times D_3 \times I \to \mathbb{R}^d.
\]

Then the system

\[
\begin{align*}
(1.85a) & \quad F_1(y'(x), y(x), z(x), x) = 0, \quad x \in I, \\
(1.85b) & \quad F_2(y(x), z(x), x) = 0, \quad x \in I,
\end{align*}
\]
is called a semi-implicit system of first order differential equations in separated form. The characteristic feature is an a priori separation into the differential variable $y$ and the algebraic variable $z$. Therefore, the system is also said to be a differential-algebraic system.

The differential index of a differential-algebraic system can be determined by the differentiation of the algebraic part.

**Example 1.86. (Systems of index 1)**

We consider the differential-algebraic system

(1.86a) \[ y' = f(y, z), \]
(1.86b) \[ 0 = g(y, z). \]

We assume that $\partial g/\partial z$ admits a bounded inverse in a neighborhood of the solution. Differentiation of (1.86b) yields

\[
0 = \frac{\partial g}{\partial y} y' + \frac{\partial g}{\partial z} z' \implies \\
z' = - \left( \frac{\partial g}{\partial z} \right)^{-1} \frac{\partial g}{\partial y} y' = -\left( \frac{\partial g}{\partial z} \right)^{-1} \frac{\partial g}{\partial y} f(y, z).
\]

Hence, the system (1.86a),(1.86b) is of differential index $\nu = 1$. The initial values $(y(a), z(a))^T$ are consistent, if $g(y(a), z(a)) = 0$.

**Example 1.87. (Systems of index 2)**

For the differential-algebraic system

(1.87a) \[ y' = f(y, z), \]
(1.87b) \[ 0 = g(y), \]

we assume that $\frac{\partial g}{\partial y} \frac{\partial f}{\partial z}$ admits a bounded inverse in a neighborhood of the solution. By differentiation of (1.87b) it follows that

(1.87c) \[ 0 = \frac{\partial g}{\partial y} y' = \frac{\partial g}{\partial y} f(y, z). \]

The system (1.87a),(1.87c) is of index $\nu = 1$. Hence, the system (1.87a),(1.87b) is of index $\nu = 2$. The initial values $(y(a), z(a))^T$ are consistent, if $g(y(a)) = 0$, $\frac{\partial}{\partial y} (g(y) f(y, z))|_{y=y(a),z=z(a)} = 0$. 
**Example 1.88. (Systems of index 3)**

We consider the differential-algebraic system

\begin{align}
\dot{y} &= f(y, z), \\
\dot{z} &= k(y, z, u), \\
0 &= g(y).
\end{align}

We suppose that \( \frac{\partial g}{\partial y} \frac{\partial f}{\partial z} \frac{\partial k}{\partial u} \) admits a bounded inverse in a neighborhood of the solution. Differentiating (1.88c) twice results in

\begin{align}
0 &= \frac{\partial g}{\partial y} f(y, z), \\
0 &= \frac{\partial^2 g}{\partial y^2} f(y, z)^T f(y, z) + \frac{\partial g}{\partial y} \left( \frac{\partial f}{\partial y} f(y, z) + \frac{\partial f}{\partial z} k(y, z, u) \right).
\end{align}

The system (1.88a)-(1.88c) has the differential index \( \nu = 3 \). Hence, the system (1.88a)-(1.88c) has the differential index \( \nu = 3 \). The initial values \((y(a), z(a), u(a))^T\) are consistent, if

\begin{align*}
g(y(a)) &= 0, \\
\frac{\partial g}{\partial y} f(y, z)|_{y=y(a), z=z(a)} &= 0, \\
\frac{\partial^2 g}{\partial y^2} f^T f + \frac{\partial g}{\partial y} \left( \frac{\partial f}{\partial y} f + \frac{\partial f}{\partial z} k \right)|_{y=y(a), z=z(a), u=u(a)} &= 0.
\end{align*}

Numerical integrators for differential-algebraic systems can be obtained by a suitable adaption of integrators for stiff systems as will be explained in case of BDF. The idea is to replace \( y'(x_{j+m}) \) by the backward difference quotient

\[ D^m y_{j+m} := \frac{1}{h} \sum_{k=0}^{m} \alpha_k y_{j+k}, \]

and - given the start iterate \( y^{(0)}_{j+m} \) - to solve the nonlinear system

\[ F(D^m y_{j+m}, y_{j+m}, x_{j+m}) = 0, \quad 0 \leq j \leq N - m, \]

by a Newton-type method.

**Remark 1.89. (Extrapolation and adaptive step size control)**

As far as extrapolation techniques and adaptive step size control for stiff
systems and differential-algebraic systems based on implicit and semi-implicit integrators are concerned, we refer to Deuflhard and Bornemann.
1.5 Numerical solution of boundary-value problems

1.5.1 Theoretical foundations and preliminaries

**Definition 1.90.** (Two-point boundary-value problem)

Let $I := [a, b] \subset \mathbb{R}$, $D \subset \mathbb{R}^d$, and $f : I \times D \to \mathbb{R}^d$ as well as $r : D \times D \to \mathbb{R}^d$. Then

\begin{align}
(1.89a) & \quad y'(x) = f(x, y(x)), \quad x \in I, \\
(1.89b) & \quad r(y(a), y(b)) = 0,
\end{align}

is called a two-point boundary-value problem for the system of first order ordinary differential equations (1.89a).

As far as the existence and uniqueness of a solution is concerned, we formulate (1.89a),(1.89b) as an operator equation:

Compute $y \in C^1(I)$ such that

\begin{equation}
(1.90) \quad Ty = 0,
\end{equation}

where the operator $T : C^1(I) \to C^1(I) \times \mathbb{R}^d$ is given by

\begin{equation}
(1.91) \quad Ty := \begin{cases} 
  y(x) - y(a) - \int_a^x f(s, y(s)) \, ds, & x \in I, \\
  r(y(a), y(b)).
\end{cases}
\end{equation}

A solution $y \in C^1(I)$ is locally unique, if the following assumption holds true:

(A) There exists a neighborhood $U(y) \subset C^1(I)$ such that $T$ is injective on $U(y)$.

If $f$ is continuously differentiable in the second argument and $r$ is continuously differentiable in both arguments, the operator $T$ is Fréchet differentiable with the Fréchet derivative $T'(y) : C^1(I) \to C^1(I)$ in $y \in C^1(I)$ given by

\[
T' y^* \delta y = \begin{cases} 
  \delta y(x) - \delta y(a) - \int_a^x f_y(s, y^*(s)) \, \delta y(s) \, ds, \\
  \frac{\partial r}{\partial y_a}(y(a), y(b)) \, \delta y(a) + \frac{\partial r}{\partial y_b}(y(a), y(b)) \, \delta y(b).
\end{cases}
\]

Then, assumption (A) is satisfied, if

\begin{equation}
(1.92) \quad T' y \delta y = 0 \implies \delta y = 0.
\end{equation}
We note that the linear operator equation $T' y \delta y = 0$ is equivalent to the linear boundary-value problem

\begin{align}
(\delta y)'(x) &= f_y(x, y(x)) \ (\delta y)(x), \ x \in I, \\
A(\delta y)(a) + B(\delta y)(b) &= 0.
\end{align}

Let $W(x_0, x), x_0 \in I,$ be the Wronski matrix with respect to (1.93a), (1.93b), i.e.,

$$
\frac{dW}{dx}(x_0, x) = f_y(x, y^*(x)) \ W(x_0, x), \ x_0 \in I, \\
W(x_0, x_0) = I.
$$

We set

$$(\delta y)(x) = W(a, x) \ (\delta y)(a).$$

Since $W(a, x)$ is regular, we have

$$(\delta y)(a) = 0 \iff (\delta y)(x) = 0, \ x \in I.$$  

In view of $(\delta y)(b) = W(a, b)(\delta y)(a)$, from (1.93b) we deduce

$$
\left(A^* + B^* \ W(a, b)\right) \ (\delta y)(a) = 0.
$$

Hence, it holds $(\delta y)(a) = 0$ if and only if $\det\left(A + B \ W(a, b)\right) \neq 0$.

**Definition 1.91. (Sensitivity matrix)**

The matrix

$$E(x) := A \ W(a, x) + B \ W(b, x)$$

is called the sensitivity matrix in $x \in I$.

**Lemma 1.92. (Properties of sensitivity matrix)**

The sensitivity matrix satisfies

\begin{equation}
E(x) = E(a) \ W(a, x), \ x \in I.
\end{equation}

Hence, $E(x)$ is regular for all $x \in I$ if and only if $E(x)$ is regular for some $x \in I$.

**Proof.** It holds

$$
E(a) \ W(a, x) = (A + B \ W(b, a)) \ W(a, x) = A \ W(a, x) + B \underbrace{W(b, a) \ W(a, x)}_{= W(b, x)} = E(x).
$$

\[\square\]
We have thus shown:

**Theorem 1.93. (Local uniqueness of the two-point boundary-value problem)**

A solution \( y \in C^1(I) \) of the two-point boundary-value problem (1.89a), (1.89b) is locally unique, if the sensitivity matrix \( E(x) \) is regular for some \( x \in I \).

We consider the linear two-point boundary-value problem

\[
(1.95a) \quad (\delta y_f)'(x) = f_y(x, y(x)) (\delta y_f)(x) + \delta f(x), \quad x \in I,
\]

\[
(1.95b) \quad A (\delta y_f)(a) + B (\delta y_f)(b) = 0,
\]

with the Wronski matrix \( W(x_0, x) \) and the sensitivity matrix \( E(x) \).

**Definition 1.94. (Green’s function)**

The function

\[
G(x, s) := \begin{cases} 
E(x)^{-1} \ A \ W(a, s) & , \ a \leq s \leq x, \\
-E(x)^{-1} \ A \ W(b, s) & , \ x \leq s \leq b
\end{cases}
\]

is called Green’s function of the linear two-point boundary-value problem (1.95a),(1.95b).

**Theorem 1.95. (Solution of linear two-point boundary-value problems)**

Suppose that the sensitivity matrix \( E(x) \) is regular. Then, the unique solution of the linear two-point boundary-value problem (1.95a),(1.95b) is given by

\[
(\delta y_f)(x) = \int_a^b G(x, s) \ \delta f(s) \ ds.
\]
1.5.2 Shooting methods

Given $I := [a, b] \subset \mathbb{R}, D_i \subset \mathbb{R}, 1 \leq i \leq 2$, a function $f : I \times D_1 \times D_2 \to \mathbb{R}$, and real numbers $\alpha, \beta$, we consider the following two-point boundary-value problem for a second order ordinary differential equation

\begin{align}
(1.96a) & \quad y''(x) = f(x, y(x), y'(x)), \quad x \in I, \\
(1.96b) & \quad y(a) = \alpha, \ y(b) = \beta.
\end{align}

The idea is to consider the initial-value problem

\begin{align}
(1.97a) & \quad y''(x) = f(x, y(x), y'(x)), \quad x \in [a, b], \\
(1.97b) & \quad y(a) = \alpha, \ y'(a) = s,
\end{align}

and to determine the initial slope $s$ such that

\begin{equation}
(1.98) \quad y(b) = y(b; s) = \beta,
\end{equation}

where $y(\cdot; s)$ is the solution of the initial-value problem (1.97a),(1.97b). In other words: We are looking for a zero of the nonlinear map

\begin{equation}
(1.99) \quad F(s) := y(b; s) - \beta.
\end{equation}

Given a start iterate $s^{(0)}$, we solve (1.99) by Newton’s method

\begin{equation}
(\nu+1) = s^{(\nu)} - \frac{F(s^{(\nu)})}{F'(s^{(\nu)})}, \quad \nu \geq 0.
\end{equation}

The computation of $F(s^{(\nu)})$ requires the solution of the initial-value problem

\begin{align}
 y''(x) = f(x, y(x), y'(x)), \quad x \in I, \\
y(a) = \alpha, \ y'(a) = s^{(\nu)}.
\end{align}
As far as the computation of \( F'(s^{(\nu)}) \) is concerned, we observe \( F'(s) = \frac{\partial}{\partial s} y(b; s) \) and note that the function \( z(x; s) := \frac{\partial}{\partial x} y(x; s) \) satisfies the initial-value problem

\[
\begin{align*}
z''(x) &= f_y(x, y(x), y'(x)) z(x) + f_{y'}(x, y(x), y'(x)) z'(x), \quad x \in I, \\
z(a) &= 0, \quad z'(a) = 1.
\end{align*}
\]

The shooting method can be easily adapted to two-point boundary-value problems of the form (1.89a),(1.89b): We consider the initial-value problem

\[
\begin{align*}
y'(x; s) &= f(x, y(x; s)), \quad x \in I, \tag{1.100a} \\
y(a; s) &= s, \tag{1.100b}
\end{align*}
\]

and determine \( s \in \mathbb{R}^d \) such that

\[
F(s) := r(y(a; s), y(b; s)) = r(s, y(b, s)) = 0. \tag{1.101}
\]

Newton’s method for the solution of (1.101) reads

\[
F_s(s^{(\nu)}) \Delta s^{(\nu)} = -F(s^{(\nu)}),
\]

\[
s^{(\nu+1)} = s^{(\nu)} + \Delta s^{(\nu)},
\]

where

\[
F_s(s) = \begin{pmatrix} \frac{\partial r}{\partial y_a}(s, y(b; s)) + \frac{\partial r}{\partial y_b}(s, y(b; s)) \cdot W_s(b, a) \end{pmatrix} =: A + B
\]

and \( W_s(x, a) \) is the solution of

\[
\begin{align*}
\frac{d}{dx} W_s(x, a) &= f_y(x, y(x; s)) W_s(x, a), \\
W_s(a, a) &= I.
\end{align*}
\]

The computation of \( A, B, \) and \( f_y \) is done by numerical differentiation.

Significant drawbacks of the shooting method are as follows:

(i) The occurrence of singularities of the solution of the initial-value problem (1.100a),(1.100b) in \( I \) which prevents the computation of the trajectory \( y(x; s) \) over the entire interval \( I \).

(ii) Even if the two-point boundary-value problem is well conditioned, the associated initial-value problem can be ill conditioned.
Example 1.96. (Failure of the shooting method)
Consider the two-point boundary-value problem
\[
y''(x) = \lambda \sinh(\lambda y(x)), \quad x \in [0, 1],
y(0) = 0, \quad y(1) = 1.
\]
The associated initial-value problem
\[
y''(x) = \lambda \sinh(\lambda y(x)), \quad x \in [0, 1],
y(0) = 0, \quad y'(0) = s
\]
has a singularity (for \( s \to 0 \)) in
\[
x_s \doteq \frac{1}{\lambda} \ln\left(\frac{8}{|s|}\right), \quad (x_s > 1!),
\]
and hence, \( s \) has to be chosen such that
\[
|s| \leq 8 \exp(-\lambda), \quad \text{e.g.,} \quad |s| \leq 0.05 \quad \text{for} \ \lambda = 5.
\]
A remedy to circumvent the above drawbacks is the multiple shooting method.

The idea is to partition the interval \( I := [a, b] \) according to
\[
\Delta_m := \{x_1 := a < x_2 < \cdots < x_m := b\}, \quad m > 2,
\]
and to compute \( y(x; s_k), x \in [x_k, x_{k+1}], k = 1, \cdots, m - 1, \) as the solutions of the initial-value problems
\[
(1.102a) \quad y'(x) = f(x, y(x)), \quad x \in [x_k, x_{k+1}],
\]
\[
(1.102b) \quad y(x_k) = s_k.
\]
Here, $s_1, \ldots, s_{m-1}$, and $s_m$ have to be computed such that the composed function
\[ y(x) := y(x; s_k), \quad x \in [x_k, x_{k+1}], \ 1 \leq k \leq m-1, \]
\[ y(b) := s_m, \]
is continuous and satisfies the boundary conditions
\[ (1.103a) \quad F_k(s_k, s_{k+1}) := y(x_{k+1}; s_k) - s_{k+1} = 0, \quad 1 \leq k \leq m-1, \]
\[ (1.103b) \quad F_m(s_1, s_m) := r(s_1, s_m) = 0. \]

We note that (1.103a),(1.103b) represents a cyclic system in the $n \times m$ unknowns $s_k = (s_{k,1}, \ldots, s_{k,m})^T$, $1 \leq k \leq m$. The cyclic system can be solved by Newton’s method
\[ (1.104a) \quad J(s^{(\nu)}) \Delta s^{(\nu)} = -F(s^{(\nu)}), \quad \nu \geq 0, \]
\[ (1.104b) \quad s^{(\nu+1)} = s^{(\nu)} + \Delta s^{(\nu)}, \]
where
\[ J(s) = \left( \frac{\partial F_k}{\partial s_{t}} \right)_{k,t=1}^m. \]

The linear system (1.104a) for the Newton increments reads as follows:
\[
\begin{pmatrix}
G_1 & -I & 0 & \cdots & 0 \\
0 & G_2 & -I & \cdots & \\
& \ddots & \ddots & \ddots & \\
0 & & \ddots & 0 & \\
A & 0 & \cdots & 0 & B
\end{pmatrix}
\begin{pmatrix}
\Delta s_1^{(\nu)} \\
\Delta s_2^{(\nu)} \\
\vdots \\
\Delta s_{m-1}^{(\nu)} \\
\Delta s_m^{(\nu)}
\end{pmatrix}
= -
\begin{pmatrix}
F(s^{(1)}) \\
F(s^{(2)}) \\
\vdots \\
F(s^{(m-1)}) \\
F(s^{(m)})
\end{pmatrix}
\]

where $G_k := W(x_{k+1}, x_k)|_{y(x, s_k)}$, $1 \leq k \leq m-1$, are the Wronski matrices and $A := \frac{\partial r}{\partial y_a}$, $B := \frac{\partial r}{\partial y_b}$. 

1.5.3 Finite difference approximations

Given $I := [a, b] \subset \mathbb{R}$ and functions $f, s \in C(I)$, and $r \in C^1(I)$, satisfying

$$0 < c_0 \leq r(x) \leq c_1, \quad x \in I, \quad s(x) \geq 0, \quad x \in I,$$

we consider the two-point boundary value problem

$$(1.105a) \quad -(r(x) \, y'(x))' + s(x) \, y(x) = f(x), \quad x \in [a, b],$$

$$(1.105b) \quad y(a) = 0, \quad y(b) = 0.$$

Introducing the variable

$$z(x) := r(x) \, y'(x), \quad x \in I,$$

the second order ordinary differential equation (1.105a) can be written as a system of two first order ordinary differential equations

$$(1.106a) \quad y'(x) = z(x)/r(x), \quad x \in I,$$

$$(1.106b) \quad z'(x) = s(x) \, y(x) - f(x), \quad x \in I.$$

We partition the interval $I$ into an equidistant grid

$$I_h := \{x_k = a + k \, h \mid 0 \leq k \leq n + 1\}$$

of step size $h := (b - a)/(n + 1)$ and approximate the derivatives $y'(x)$ and $z'(x)$ in $x_{k+1/2} = x_k + h/2, 0 \leq k \leq n$, by the central difference quotients

$$\frac{1}{h} (y(x_{k+1}) - y(x_k)) \approx \frac{z(x_{k+1/2})}{r(x_{k+1/2})},$$

$$\frac{1}{h} (z(x_{k+1}) - z(x_k)) \approx s(x_{k+1/2}) \, y(x_{k+1/2}) - f(x_{k+1/2}).$$

Since we want to compute approximations only the grid points $x_k, 0 \leq k \leq n + 1$, we replace $z(x_{k+1/2})$ and $y(x_{k+1/2})$ by averaging

$$z(x_{k+1/2}) \approx \frac{1}{2} (z(x_k) + z(x_{k+1})), \quad y(x_{k+1/2}) \approx \frac{1}{2} (y(x_k) + y(x_{k+1})).$$

Setting

$$r_{k+1/2} := r(x_{k+1/2}), \quad s_{k+1/2} := s(x_{k+1/2}), \quad f_{k+1/2} := f(x_{k+1/2}),$$

we have the system

$$\frac{1}{h} (y(x_{k+1}) - y(x_k)) \approx \frac{z(x_{k+1/2})}{r(x_{k+1/2})},$$

$$\frac{1}{h} (z(x_{k+1}) - z(x_k)) \approx s(x_{k+1/2}) \, y(x_{k+1/2}) - f(x_{k+1/2}).$$

with boundary conditions $y(a) = 0$ and $y(b) = 0$. This system can be solved numerically for the unknowns $y(x_k)$ and $z(x_k)$, $0 \leq k \leq n$. This is the finite difference approximation to the original boundary value problem (1.105) on the interval $I$. The accuracy of the approximation depends on the size of the step $h$ and the smoothness of the functions $f, s, r$. Larger step sizes lead to lower accuracy, but faster computation. The method is easy to implement and can be adapted to various types of boundary value problems.
the finite difference approximation of (1.105a),(1.105b) amounts to the
computation of grid functions \( y_h, z_h \in C(I_h) \) as the solution of

\[
\begin{align*}
(1.107a) & \quad y_h(x_{k+1}) = y_h(x_k) + \frac{h}{2} r_{k+1/2} (z_h(x_k) + z_h(x_{k+1})), \\
(1.107b) & \quad z_h(x_{k+1}) = z_h(x_k) + \frac{h}{2} (y_h(x_k) + y_h(x_{k+1})), \\
(1.107c) & \quad y_h(a) = y_h(b) = 0.
\end{align*}
\]

Solving (1.107a) for \( z_h(x_{k+1}) \) yields

\[
(1.108) \quad z_h(x_{k+1}) = -z_h(x_k) + \frac{2 r_{k+1/2}}{h} (y_h(x_{k+1}) - y_h(x_k)).
\]

The corresponding equation for \( z_h(x_k) \) reads

\[
(1.109) \quad z_h(x_k) = -z_h(x_{k-1}) + \frac{2 r_{k-1/2}}{h} (y_h(x_k) - y_h(x_{k-1})).
\]

Inserting (1.109) into (1.108) results in

\[
(1.110) \quad z_h(x_{k+1}) - z_h(x_{k-1}) = - \frac{2 r_{k-1/2}}{h} (y_h(x_k) - y_h(x_{k-1})) + \frac{2 r_{k+1/2}}{h} (y_h(x_{k+1}) - y_h(x_k)).
\]

On the other hand, the addition of (1.107b) for \( k-1 \) and \( k \) gives

\[
\begin{align*}
(1.111a) & \quad z_h(x_k) = z_h(x_{k-1}) + \frac{h}{2} s_{k-1/2} (y_h(x_{k-1}) + y_h(x_k)) - h f_{k-1/2} + \frac{h}{2} s_{k+1/2} (y_h(x_k) + y_h(x_{k+1})) - h f_{k+1/2} \\
(1.111b) & \quad z_h(x_{k+1}) = z_h(x_k) + \frac{h}{2} s_{k+1/2} (y_h(x_k) + y_h(x_{k+1})) - h f_{k+1/2} \\
(1.111c) & \quad \frac{h}{2} s_{k-1/2} y_h(x_{k-1}) + \frac{h}{2} (s_{k-1/2} + s_{k+1/2}) y_h(x_k) + \frac{h}{2} s_{k+1/2} y_h(x_{k+1}) - (z_h(x_{k+1}) - z_h(x_{k-1})) = h (f_{k-1/2} + f_{k+1/2}).
\end{align*}
\]

Inserting (1.110) into yields

\[
(1.112) \quad \left( -\frac{2 r_{k-1/2}}{h} + \frac{h}{2} s_{k-1/2} \right) y_h(x_{k-1}) + \left( \frac{2 r_{k-1/2}}{h} + \frac{2 r_{k+1/2}}{h} + \frac{h}{2} s_{k-1/2} \right) y_h(x_k) + \left( -\frac{2 r_{k+1/2}}{h} + \frac{h}{2} s_{k+1/2} \right) y_h(x_{k+1}) = h \left( f_{k-1/2} + f_{k+1/2} \right).
\]
1.5.4 Galerkin approximations

We multiply (1.105a) with a function \( \varphi \in C^1_0(I) \) and integrate over \( I \):

\[
- \int_a^b (r(x) y'(x))' \varphi(x) \, dx + \int_a^b s(x) y(x) \varphi(x) \, dx = \int_a^b f(x) \varphi(x) \, dx.
\]

Partial integration of the first integral results in

\[
- \int_a^b (r(x) y'(x))' \varphi(x) \, dx = \int_a^b r(x) y'(x) \varphi'(x) \, dx - \left[ r(x) y'(x) \varphi(x) \right]_a^b = 0.
\]

Hence, we obtain

\[
\int_a^b r(x) y'(x) \varphi'(x) \, dx + \int_a^b s(x) y(x) \varphi(x) \, dx = \int_a^b f(x) \varphi(x) \, dx.
\]

In view of this variational equation, on \( C^1_0(I) \) we introduce a norm according to

\[
\| y \|_1 := \left( \int_a^b (y'(x)^2 + y(x)^2) \, dx \right)^{1/2}.
\]

The space \((C^1_0(I), \| \cdot \|_1)\) is not complete. We refer to \( H^1_0(I) \) as the closure of \((C^1_0(I), \| \cdot \|_1)\) with respect to the topology induced by \( \| \cdot \|_1 \) on \( C^1_0(I) \). Then, the so-called weak formulation of (1.105a),(1.105b) amounts to the computation of \( y \in H^1_0(I) \) such that for all \( \varphi \in H^1_0(I) \) it holds

\[
(1.113) \quad \int_a^b r(x) y'(x) \varphi'(x) \, dx + \int_a^b s(x) y(x) \varphi(x) \, dx = \int_a^b f(x) \varphi(x) \, dx.
\]

We choose a subspace \( S_h \subset H^1_0(I) \) with \( \dim S_h = n_h < \infty \). Then, the Galerkin approximation of (1.113) requires the computation of \( y_h \in S_h \) such that for all \( \varphi_h \in S_h \) the finite dimensional variational equation

\[
(1.114) \quad \int_a^b r(x) y_h(x) \varphi_h'(x) \, dx + \int_a^b s(x) y_h(x) \varphi_h(x) \, dx = \int_a^b f(x) \varphi_h(x) \, dx
\]

is satisfied.
We consider a partition
\[ I_h := \{ x_k := a + k \, h \mid 0 \leq k \leq n + 1 \}, \quad h := \frac{b - a}{n + 1}, \]
of \( I \) and choose \( S_h \) as the linear space \( S_{1,0}(I; I_h) \) of linear splines with respect to \( I_h \):
\[ S_{1,0}(I; I_h) := \{ v_h \in C_0(I) \mid v_h|_{[x_k,x_{k+1}]} \in P_1([x_k,x_{k+1}]), \ 0 \leq k \leq n \}. \]
The space \( S_{1,0}(I; I_h) \) is spanned by the basis functions \( \varphi^{(k)}_h, 1 \leq k \leq n \), with
\[ \varphi^{(k)}_h(x_j) = \delta_{jk}, \ 1 \leq j, k \leq n. \]
The solution \( y_h \in S_{1,0}(I; I_h) \) of (1.114) can be written as a linear combination of the basis functions according to
\[ y_h = \sum_{j=1}^{n} c_j \, \varphi^{(j)}_h. \]
Then, (1.114) holds true if and only if
\[ \sum_{j=1}^{n} \left( \int_a^b r(x) \, (\varphi^{(j)}_h)'(x) \, (\varphi^{(i)}_h)'(x) \, dx + \int_a^b s(x) \, (\varphi^{(j)}_h)(x) \, (\varphi^{(i)}_h)(x) \, dx \right) c_j =: a_{ij} =: b_i, \]
i.e., the Galerkin method requires the solution of the linear system
\[ (1.115) \quad A \, c = b. \]

**Definition 1.97. (Stiffness matrix and load vector)**
The matrix \( A = (a_{ij})_{i,j=1}^{n} \) and the vector \( b = (b_1, \ldots, b_n)^T \) are called the stiffness matrix and the load vector.

**Lemma 1.98. (Properties of the stiffness matrix)**
The stiffness matrix \( A \) is a symmetric, positive definite tridiagonal matrix.
Lemma 1.99. (Numerical quadrature)

If we compute the entries of the stiffness matrix $A$ and of the load vector $b$ by the quadrature formula

\[
\int_{x_k}^{x_{k+1}} g(x) \, dx \approx h \, g(x_k + \frac{h}{2}) =: h \, g_{k+1/2},
\]

we obtain

\[
a_{ii} = \frac{r_{k-1/2}}{h} + \frac{r_{k+1/2}}{h} + \frac{h}{4} \, s_{k-1/2} + \frac{h}{4} \, s_{k+1/2}, \quad 1 \leq i \leq n,
\]

\[
a_{i,i+1} = -\frac{r_{k+1/2}}{h} + \frac{h}{4} \, s_{k+1/2}, \quad 1 \leq i \leq n - 1,
\]

\[
a_{i-1,i} = -\frac{r_{k-1/2}}{h} + \frac{h}{4} \, s_{k-1/2}, \quad 2 \leq i \leq n,
\]

\[
b_i = \frac{h}{2} \, f_{k-1/2} + \frac{h}{2} \, f_{k+1/2}.
\]

Remark 1.100. (Equivalence with finite difference approximation)

If the entries of the stiffness matrix and the load vector are computed by numerical quadrature, the linear system $Ac = b$ coincides with (1.112).
**Literature**


2. Partial Differential Equations

2.1 Theoretical foundations and preliminaries

A partial differential equation is an equation which contains an unknown function in several independent variables along with certain partial derivatives of that function. The order of the differential equation corresponds to the highest partial derivative that occurs in the equation.

Let \( \Omega \) be a domain in \( \mathbb{R}^d \), \( d \in \mathbb{N}, d \geq 2 \), and let \( \alpha \) be a multi-index in \( \mathbb{N}_0^d \). For a function \( u \in C^m(\Omega), m \in \mathbb{N} \), we denote by \( D^\alpha u(x) := \frac{\partial^{\vert \alpha \vert} u(x)}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} \).

**Definition 2.1. (Partial differential equation of \( m \)-th order)**

Let \( F : \Omega \times \mathbb{R}^N \rightarrow \mathbb{R}, N := \text{card}\{ \alpha \in \mathbb{N}_0^d \mid \vert \alpha \vert \leq m \} \), and \( u \in C^m(\Omega) \). Then

\[
F(x, (D^\alpha u(x))_{\vert \alpha \vert \leq m}) = 0, \quad x \in \Omega
\]

is called a nonlinear partial differential equation of order \( m \).

A partial differential equation (2.1) is called quasilinear, if it is linear in the highest partial derivatives \( D^\alpha u, \vert \alpha \vert = m \), i.e., if there exist functions \( a_\alpha : \Omega \times \mathbb{R}^{N_1} \rightarrow \mathbb{R} \) and \( F_1 : \Omega \times \mathbb{R}^{N_1} \rightarrow \mathbb{R}, N_1 \in \mathbb{N} \), such that the function \( F \) in (2.1) has the representation

\[
F(x, (D^\alpha u(x))_{\vert \alpha \vert \leq m}) = \sum_{\vert \alpha \vert = m} a_\alpha(x, (D^\beta u(x))_{\vert \beta \leq m-1}) D^\alpha u(x) + F_1(x, (D^\beta u(x))_{\vert \beta \leq m-1}).
\]

A quasilinear partial differential equation is said to be semilinear, if the coefficient function \( a_\alpha, \vert \alpha \vert = m \), only depends on \( x_i, 1 \leq i \leq d \), but neither on \( u \) nor its partial derivatives.

A partial differential equation is called linear, if the function \( F \) in (2.1) is linear in \( u \) and its partial derivatives \( D^\alpha u, \vert \alpha \vert \leq m \), i.e., if there exist functions \( a_\alpha : \mathbb{R}^d \rightarrow \mathbb{R}, \vert \alpha \vert \leq m \), and \( f : \mathbb{R}^d \rightarrow \mathbb{R} \) such that

\[
Lu := L(\cdot, D)u = f \quad \text{in } \Omega \subset \mathbb{R}^d,
\]
where \( L(\cdot, D) \) denotes the linear partial differential operator
\[
L(x, D) u := \sum_{|\alpha| \leq m} a_\alpha(\cdot) \ D^\alpha.
\]
(2.4)

The differential operator
\[
L^H(\cdot, D) u := \sum_{|\alpha| = m} a_\alpha(\cdot) \ D^\alpha
\]
(2.5)
is called the main part of \( L(\cdot, D) \).

The classification of linear partial differential equations of \( m \)-th order is done by means of the symbol.

**Definition 2.2. (Symbol)**
The symbol of the linear partial differential operator \((2.4)\) is given by
\[
L(x, i\xi) := \sum_{|\alpha| \leq m} a_\alpha(x) (i\xi)^\alpha, \quad \xi^\alpha := \prod_{\nu=1}^d \xi_\nu^{\alpha_\nu},
\]
(2.6) where \( i \) stands for the imaginary unit. We refer to
\[
L^H(x, i\xi) := \sum_{|\alpha| = m} a_\alpha(x) (i\xi)^\alpha
\]
(2.7) as the main part of the symbol.

In the sequel we will restrict ourselves to the case \( m = 2 \). Given continuous functions \( a_{ij} : \mathbb{R}^d \to \mathbb{R}, 1 \leq i, j \leq d \), and \( b_i : \mathbb{R}^d \to \mathbb{R} \), as well as \( c : \mathbb{R}^d \to \mathbb{R} \), a linear partial differential equation of second order can be written in the form
\[
Lu = \sum_{i,j=1}^d a_{ij} \ u_{x_i x_j} + \sum_{i=1}^d b_i \ u_{x_i} + c \ u = f,
\]
(2.8) where \( u_{x_i} := \partial u / \partial x_i \) and \( u_{x_i x_j} := \partial^2 u / \partial x_i \partial x_j \). The main part \( L^H(x, i\xi) \) of the symbol \( L(x, i\xi) \) is given by
\[
L^H(x, i\xi) = - \sum_{i,j=1}^d a_{ij}(x) \ \xi_i \xi_j.
\]
(2.9) We define \( A(x) \in \mathbb{R}^{d \times d} \) as the matrix
\[
A(x) := (-a_{ij}(x))_{i,j=1}^d.
\]
(2.10) Hence, the main part \( L^H(x, i\xi) \) satisfies
\[
L^H(x, i\xi) = \xi^T A(x) \xi
\]
and thus represents a quadratic form in $\xi$.

**Definition 2.3. (Classification of linear second order partial differential equations)**

A linear second order partial differential equation (2.8) is called elliptic in $x \in \Omega$, if all eigenvalues of the matrix $A(x)$ have the same sign. It is said to be parabolic in $x \in \Omega$, if $A(x)$ is singular. Finally, it is called hyperbolic in $x \in \Omega$, if all but one eigenvalue of $A(x)$ have the same sign and one eigenvalue has the opposite sign. Note that eigenvalues are counted according to their multiplicity.

A linear second order partial differential equation (2.8) is called elliptic (parabolic, hyperbolic) in $\Omega$, if it is elliptic (parabolic, hyperbolic) for all $x \in \Omega$.

The main part of linear second order partial differential equations of the same type can be transformed by a principal axis transformation to a canonical form.

**Example 2.4. (Poisson and Laplace equation)**

The canonical form of an elliptic second order differential equation is the Poisson equation

\begin{equation}
-\Delta u = f \tag{2.11}
\end{equation}

with the Laplace operator

\begin{equation}
\Delta := \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \tag{2.12}
\end{equation}

In case $f \equiv 0$ (2.11) is said to be the Laplace equation. The symbol of $-\Delta$ is $\sum_{i=1}^{d} \xi_i^2$. Hence, $A = A(x)$ has the $d$-fold eigenvalue +1.

**Example 2.5. (Heat equation)**

The canonical form of a linear second order parabolic equation is the heat equation

\begin{equation}
ux_1 - \sum_{i=2}^{d} u_{x_ix_i} = f \tag{2.13}
\end{equation}

with the symbol $i\xi_1 + \sum_{i=2}^{d} \xi_i^2$. Here, $A = A(x)$ has the simple eigenvalue 0 and the $(d - 1)$-fold eigenvalue +1.

The heat equation represents the temporal and spatial temperature distribution of a heat conducting material.
Example 2.6. (Wave equation)
The canonical form of a linear second order hyperbolic differential equation is the wave equation

\( u_{x_1 x_1} - \sum_{i=2}^{d} u_{x_i x_i} = f. \)  

(2.14)

The symbol is \( -\xi^2_1 + \sum_{i=2}^{d} \xi^2_i. \) The matrix \( A = A(x) \) has the simple eigenvalue \(-1\) and the \((d-1)\)-fold eigenvalue \(+1.\)

The wave equation describes the temporal and spatial propagation of acoustic waves.

Example 2.7. (Tricomi equation)
In general, partial differential equations may change its type within a domain. An example is the Tricomi equation

\( x_2 u_{x_1 x_1} + u_{x_2 x_2} = 0. \)  

(2.15)

The symbol is \( -x_2 \xi^2_1 - \xi^2_2. \) Hence, the Tricomi equation is elliptic for \( x_2 > 0, \) parabolic for \( x_2 = 0, \) and hyperbolic for \( x_2 < 0. \)

The classification of linear second order partial differential equations is closely related to the notion of characteristics.

Definition 2.8. (Characteristics)
Let \( \phi \) be continuously differentiable in \( \mathbb{R}^d. \) Then, the \((d-1)\)-dimensional manifold

\( \mathcal{C} := \{x \in \mathbb{R}^d \mid \phi(x) = 0\} \)

is said to be characteristic in \( x^* \in \mathbb{R}^d, \) if it holds

\( \phi(x^*) = 0 \) and \( (\nabla \phi(x^*))^T A(x^*) \nabla \phi(x^*) = 0. \)

(2.16)

The manifold \( \mathcal{C} \) is called a characteristics, if it is characteristic in each of its points.

It follows from Definition 2.8 that elliptic partial differential equations do not possess real characteristics. For \( d = 2 \) the characteristics of the heat equation (2.13) are the straight lines \( x_2 = const. \) parallel to the \( x_1-\) Achse, whereas the characteristics of the wave equation (2.14) are the straight lines \( x_2 \pm x_1 = const.. \)

A classical existence and uniqueness result is the Theorem of Cauchy-Kovalevskaya for real analytic solutions of the so-called Cauchy problem. We consider the following system of quasilinear partial differential
equations of first order

\[(2.17) \quad (u_i)_x = \sum_{j=1}^{k} \sum_{|\alpha| \leq 1} a^{(ij)}_{\alpha}(x', u) D^\alpha u_j, \quad 1 \leq i \leq k, \]

where \(x' := (x_1, \cdots, x_{d-1})^T\) and \(u := (u_1, \cdots, u_k)^T\).

**Definition 2.9. (Real analytic solution)**

A function \(g : \mathbb{R}^d \to \mathbb{R}\) is called a real analytic function in \(x^* \in \mathbb{R}^d\), if there exists a neighborhood \(U(x^*)\) such that for all \(x \in U(x^*)\) the function \(g\) can be written as a Taylor series

\[g(x) = \sum_{\alpha \in \mathbb{N}_0^d} c_\alpha (x - x^*)^\alpha.\]

The function is said to be real analytic in \(\mathbb{R}^d\), if it is real analytic in each point \(x^* \in \mathbb{R}^d\).

If the coefficient functions \(a^{(ij)}_\alpha, |\alpha| \leq 1, 1 \leq i, j \leq k\), are real analytic in the origin \(p_0\) of \(\mathbb{R}^{d+k-1}\), the vector-valued function \(u = (u_1, \cdots, u_k)^T\) with \(u_i : \mathbb{R}^d \to \mathbb{R}, 1 \leq i \leq k\), is called a real analytic solution of \((2.17)\), if the functions \(u_i, 1 \leq i \leq k\), are real analytic in \(p_0\) and if \((2.17)\) is satisfied in \(p_0\).

If the functions \(a^{(ij)}_\alpha\) are real analytic in \(\mathbb{R}^d\), the function \(u\) is said to be a real analytic solution of \((2.17)\) in \(\mathbb{R}^d\), if the functions \(u_i, 1 \leq i \leq k\), are real analytic in \(\mathbb{R}^d\) and if \((2.17)\) is satisfied in each point \(x^* \in \mathbb{R}^d\).

The Cauchy problem associated with \((2.17)\) is to find a locally unique solution for given initial conditions on the \((d-1)\)-dimensional hyperplane

\[(2.18) \quad \mathcal{H} := \{x \in \mathbb{R}^d \mid x_d = 0\},\]

which is not a characteristics of \((2.17)\).

**Definition 2.10. (Cauchy problem)**

Let \(\mathcal{H}\) be given by \((2.18)\) and assume that \(g_i : \mathcal{H} \to \mathbb{R}, 1 \leq i \leq k\), are given functions. The Cauchy problem associated with \((2.17)\) is to compute a vector-valued function \(u = (u_1, \cdots, u_k)^T\) such that \((2.17)\) and

\[(2.19) \quad u_i(x) = g_i(x), \quad x \in \mathcal{H}, \quad 1 \leq i \leq k,\]

hold true.
The Theorem of Cauchy-Kovalevskaya provides conditions for the local existence and uniqueness of a real analytic solution of the Cauchy problem.

**Theorem 2.11. (Theorem of Cauchy-Kovalevskaya)**

Suppose that the functions $a_{ij}^{(\alpha)}$, $|\alpha| \leq 1$, $1 \leq i, j \leq k$, are real analytic in the origin of $\mathbb{R}^{d+k-1}$ and that the functions $g_i$, $1 \leq i \leq k$, are real analytic in the origin of the hyperplane $\mathcal{H}$. Then the Cauchy problem admits a locally unique real analytic solution.

*Proof.* The proof relies on the technique of real analytic majorants. For details we refer to Renardy and Rogers. □
2.2 Linear second order elliptic boundary value problems

2.2.1 Preliminaries

We consider linear second order elliptic differential equations in a bounded domain $\Omega \subset \mathbb{R}^d$

\[(2.20) \qquad Lu := -\sum_{i,j=1}^{d} a_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{d} b_i \frac{\partial u}{\partial x_i} + cu = f\]

under the following assumptions on the coefficient functions

\[(2.21a) \qquad \text{It holds } a_{ij}, b_i, c \in C(\Omega), \ 1 \leq i, j \leq d, \text{ and there exists a constant } C > 0, \text{ such that for all } 1 \leq i, j \leq d \text{ and } x \in \Omega \]

\[|a_{ij}(x)|, |b_i(x)|, |c(x)| \leq C,\]

\[(2.21b) \qquad \text{The functions } a_{ij} \text{ are symmetric, i.e., for all } 1 \leq i, j \leq d \text{ and } x \in \Omega \text{ it holds}\]

\[a_{ij}(x) = a_{ji}(x),\]

\[\text{and the matrix-valued function } (a_{ij})_{i,j=1}^{d} \text{ is uniformly positive definite in } \Omega, \text{ i.e., there exists a constant } \alpha > 0, \text{ such that for all } x \in \Omega \text{ and } \xi \in \mathbb{R}^d\]

\[\sum_{i,j=1}^{d} a_{ij}(x)\xi_i \xi_j \geq \alpha |\xi|^2.\]

For the sake of uniqueness of a solution of (2.20) we need boundary conditions on the boundary $\Gamma = \partial \Omega$.

\[\text{Definition 2.12. (Dirichlet, Neumann, and Robin problem)}\]

Under the assumptions (2.21a),(2.21b) let $L$ be the linear second order differential operator given by (2.20) and suppose further that $f \in C(\Omega)$ and $u^D \in C(\Gamma)$. Then, the boundary-value problem

\[(2.22a) \quad Lu(x) = f(x), \quad x \in \Omega,\]

\[(2.22b) \quad u(x) = u^D(x), \quad x \in \Gamma,\]

is called an inhomogeneous Dirichlet problem. If $u_D = 0$, it is said to be a homogeneous Dirichlet problem. A function $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is called a classical solution, if $u$ pointwise satisfies (2.22a),(2.22b).

In case of boundary conditions of the form

\[(2.23) \quad (\nu \cdot \nabla u)(x) = u^N(x), \quad x \in \Gamma,\]
where \( \nu \) is the exterior unit normal on \( \Gamma \) and \( u^N \in C(\Gamma) \), the boundary-value problem is referred to as a Neumann problem.

Boundary conditions of the form
\[
(\nu \cdot \nabla u)(x) + d(x)u(x) = u^R(x), \quad x \in \Gamma,
\]
where \( d, u^R \in C(\Gamma) \), are called Robin boundary conditions.

It is of course possible to prescribe different boundary conditions on different parts of the boundary.

If \( c(x) \geq 0, x \in \Omega \), in (2.20), the uniqueness of a solution of the Dirichlet problem (2.22a),(2.22b) follows from the maximum principle of Hopf.

**Theorem 2.13. (Maximum principle of Hopf)**

In addition the assumptions (2.21a),(2.21b) suppose that \( c(x) \geq 0, x \in \Omega \), and that the function \( u \in C^2(\Omega) \) satisfies the inequality
\[
Lu(x) \leq 0, \quad x \in \Omega.
\]

If there exists a point \( x_0 \in \Omega \) such that
\[
u(x_0) = \sup_{x \in \Omega} u(x) \geq 0,
\]
then it holds
\[
u(x) = \text{const.}, \quad x \in \Omega.
\]

**Proof.** We refer to Jost. \( \square \)

**Corollary 2.14. (Uniqueness of the Dirichlet problem)**

Under the assumptions of Theorem 2.13 the Dirichlet problem (2.22a), (2.22b) admits a unique classical solution.

**Proof.** Let \( u_1, u_2 \) be two classical solutions and \( z := u_1 - u_2 \). Then \( Lz(x) = 0, x \in \Omega \), and Theorem 2.13 implies \( z(x) = \text{const.}, x \in \Omega \). Since \( z(x) = 0, x \in \Gamma \), it follows that \( z \equiv 0 \). \( \square \)

For the Poisson equation (2.11) the existence of a solution of the Dirichlet problem can be shown by means of Green’s representation formula. We assume \( \Omega \) to be a domain whose boundary \( \Gamma \) admits a parametrization by a \( C^1 \) function (\( C^1 \)-domain).
Theorem 2.15. (First and second Green’s formula)
Let $\Omega \subset \mathbb{R}^d$ be a $C^1$-domain and assume that $u, v \in C^2(\overline{\Omega})$. Then, the equation
\[
\int_{\Omega} \Delta u \, v \, dx + \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Gamma} \nu \cdot \nabla u \, v \, d\sigma
\] (2.28)
is called Green’s first formula, whereas the equation
\[
\int_{\Omega} \left( \Delta u \, v - u \, \Delta v \right) \, dx = \int_{\Gamma} \left( \nu \cdot \nabla u \, v - u \, \nu \cdot \nabla v \right) \, d\sigma
\] (2.29)
is referred to as Green’s second formula.

Proof. The proof of (2.28) follows from the Gauss’ theorem
\[
\int_{\Omega} \nabla \cdot q \, dx = \int_{\Gamma} \nu \cdot q \, d\sigma
\]applied to the vector-valued function $q := v \nabla u$. Exchanging $u$ and $v$ in (2.28) and subtracting the resulting equation from (2.28) results in (2.29). \qed

Definition 2.16. (Fundamental solution)
A function $u \in C^2(\Omega)$ is called a harmonic function in $\Omega$, if $u$ satisfies the Laplace equation, i.e., $\Delta u(x) = 0, x \in \Omega$.

The function
\[
\Gamma(x, y) := \left\{ \begin{array}{ll}
\frac{1}{2\pi} \ln(|x - y|), & d = 2 \\
\frac{1}{d(2-d)|B_1|} |x - y|^{2-d}, & d > 2 \end{array} \right., x, y \in \mathbb{R}^d, x \neq y,
\] (2.30)
where $|B_1^d|$ is the volume of the unit ball in $\mathbb{R}^d$, is called the fundamental solution of the Laplace equation. As a function of $x$ it is a harmonic function in $\mathbb{R}^d \setminus \{y\}$.

The notion 'fundamental solution' is justified by Green’s representation formula.

Theorem 2.17. (Green’s representation formula)
Assume that $\Omega \subset \mathbb{R}^d$ is a $C^1$-domain and $\Gamma(\cdot, \cdot)$ the fundamental solution given by (2.30). Then, for $u \in C^2(\overline{\Omega})$ and $y \in \Omega$ it holds
\[
u(y) = \int_{\Omega} \Gamma(x, y) \, \Delta u(x) \, dx + \int_{\partial \Omega} \left( u(x) \, \nu \cdot \nabla \Gamma(x, y) - \Gamma(x, y) \, \nu \cdot \nabla u(x) \right) \, d\sigma(x),
\] (2.31)
where $\nabla_x$ stands for the gradient with respect to the differentiation in $x$.

Proof. The proof can be done by means of Green’s second formula (2.29). For details we refer to Jost. \hfill \Box

**Definition 2.18. (Green’s function)**
A function $G(x, y), x, y \in \overline{\Omega}, x \neq y$, is called Green’s function, if the following two conditions are satisfied:

(2.32a) $G(x, y) = 0$, $x \in \partial \Omega$,
(2.32b) The function $G(x, y) - \Gamma(x, y)$ is harmonic in $x \in \Omega$.

Green’s function enables an explicit representation of the solution of the inhomogeneous Dirichlet problem for the Poisson equation (2.11).

**Theorem 2.19. (Solution of the inhomogeneous Dirichlet problem)**
Under the assumptions of Theorem 2.17 let $G(x, y)$ be a Green’s function for $\Omega$. Then, the following statements hold true:

(i) If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is the classical solution of the inhomogeneous Dirichlet problem for the Poisson equation (2.11), for $y \in \Omega$ we have

(2.33) $u(y) = \int_{\Omega} G(x, y) f(x) \, dx + \int_{\partial \Omega} u^D(x) \, \nu \cdot \nabla_x G(x, y) \, d\sigma(x)$.

(ii) If $f$ is a Hölder continuous function in $\Omega$ and $u^D \in C(\partial \Omega)$, then the function $u$ given by (2.33) is the classical solution of the inhomogeneous Dirichlet problem for the Poisson equation.

Proof. The proof of (i) follows by the application of Green’s second formula (2.29) to the function $v(x) = G(x, y) - \Gamma(x, y)$. The proof of (ii) requires further results on the regularity of solutions of elliptic partial differential equations. We refer to Jost. \hfill \Box
2.2.2 Finite difference approximations

We consider the numerical solution of Poisson’s equation (2.11) with inhomogeneous Dirichlet boundary conditions by a finite difference method based on the approximation of the partial derivatives by difference quotients. To this end we assume \( \Omega \) to be a \( d \)-dimensional cube.

**Definition 2.20. (Grid point set)**

Let \( \Omega := (a,b)^d \) with \( a, b \in \mathbb{R}, a < b \), and \( h := (b - a)/(N + 1), N \in \mathbb{N} \).

Then the set

\[
\Omega_h := \{(x_{i_1}, \ldots, x_{i_d})^T \mid x_{i_j} = a + i_j h, \ 0 \leq i_j \leq N + 1, \ 1 \leq j \leq d\}
\]

is called a grid point set with step size \( h \). The elements of \( \Omega_h \) are called grid points. The set \( \Omega_h \) is a particular example of a uniform or equidistant grid where the grid points have the same distance to each other.

We further refer to \( \Omega_h := \Omega \cap \Omega_h \) as the set of interior grid points and to \( \Gamma_h := \Omega_h \setminus \Omega_h \) as the set of boundary grid points.

For the approximation of the partial derivatives \( \partial u / \partial x_i, \partial^2 u / \partial x_i^2, 1 \leq i \leq d \), we consider the following difference quotients:

**Definition 2.21. (Difference quotients)**

Let \( u : \Omega \to \mathbb{R} \) and let \( e_i, 1 \leq i \leq d \), be the \( i \)-th unit vector in \( \mathbb{R}^d \).

Then the difference quotients

\[
D_{h,i}^+ u(x) := h^{-1}(u(x + he_i) - u(x)),
\]

\[
D_{h,i}^- u(x) := h^{-1}(u(x) - u(x - he_i)),
\]

\[
D_{h,i}^1 u(x) := (2h)^{-1}(u(x + he_i) - u(x - he_i))
\]

are called the forward, the backward, and the central difference quotients for the first partial derivatives \( \partial u / \partial x_i, 1 \leq i \leq d \).

The difference quotients

\[
D_{h,i,i}^2 u(x) := h^{-2}(u(x + he_i) - 2u(x) + u(x - he_i))
\]

are called the central difference quotients for the second partial derivatives \( \partial^2 u / \partial x_i^2, 1 \leq i \leq d \).

It follows from Taylor expansion that for \( u \in C^2(\Omega) \) the difference quotients \( D_{h,i}^+ u \) provide an approximation of \( \partial u / \partial x_i \) of order \( O(h) \), whereas for \( u \in C^4(\Omega) \) the central difference quotients \( D_{h,i}^1 u \) bzw. \( D_{h,i}^2 u \) give rise to approximations of \( \partial^2 u / \partial x_i^2 \) of order \( O(h^2) \).
As far as approximations of the mixed second partial derivatives \( \partial^2 u / \partial x_i \partial x_j, i \neq j \), are concerned, we refer to Strikwerda.

A convenient tool to characterize difference quotients are difference stars (cf. Figure 9). Here, the grid points marked by \( \bullet \) are those grid points involved in the difference quotient. The number next to \( \bullet \) refers to the weight, i.e., the factor of the function value times \( h^{-2} \) in that grid point.

![Figure 9. Difference star for the Laplacian \( \Delta \) (d = 2)](image)

**Definition 2.22. (Discrete Laplace operator)**

The difference operator

\[
\Delta_h := \sum_{i=1}^{d} D_{h,i,i}^2
\]

is called the discrete Laplace operator. For \( u \in C^4(\Omega) \) the discrete Laplace operator \( \Delta_h u \) is an approximation of \( \Delta u \) of order \( O(h^2) \).

In case \( d = 2 \) the associated difference star is displayed in Figure 9. Since five grid points are involved, it is also referred to as the five-point approximation of the Laplacian.

For a bounded domain \( \Omega \subset \mathbb{R}^d \) with a curved boundary \( \Gamma \) the definition of interior grid points and boundary grid points has to be made more precise.

**Definition 2.23. (Domain with curved boundary)**

Let \( \Omega \subset \mathbb{R}^d \) be a bounded connected domain with boundary \( \Gamma \) and let \( \mathbb{R}^d_h \) be the uniform grid point set

\[
\mathbb{R}^d_h := \{ x = (x_1, \ldots, x_d)^T \mid x_i = ih , \ i \in \mathbb{Z} \}.
\]

Then \( \Omega_h := \mathbb{R}^d_h \cap \Omega \) is called the set of interior grid points. A point \( x \in \Gamma \) is called a boundary grid point, if there exists an interior grid
point \( x^* \in \Omega_h \) such that \( x = x^* + \alpha h e_i \) for some \( i \in \{1, \ldots, d\} \) and \( \alpha \in \mathbb{R} \) with \( |\alpha| \leq 1 \). The grid point \( x = x^* + \alpha h e_i \) is called an interior grid point close to the boundary. All interior grid points that are not close to the boundary are called grid points far from the boundary. The sets \( \Gamma_h, \Omega^I_h \), and \( \Omega^O_h \) are called the set of boundary grid points and the sets of interior grid points close to resp. far from the boundary (cf. Figure 10). We set \( \overline{\Omega}_h := \Omega_h \cup \Gamma_h \).

The connection between two neighboring grid points along a grid line is called an edge and the grid points are said to be the vertices of the edge. A grid point set is called discretely connected, if any two grid points in \( \overline{\Omega}_h \) are connected by a path consisting of edges with vertices in \( \overline{\Omega}_h \).

**Remark 2.24. (Grid points close to the boundary)**

We note that an interior grid point \( x \) can be close to the boundary, although \( x \pm h e_i \in \Omega_h, 1 \leq i \leq d \).

The definitions of difference quotients and the discrete Laplacian need a modification as well which is called Shortley-Weller approximation.

**Definition 2.25. (Shortley-Weller approximation)**

Let \( \overline{\Omega}_h \subset \mathbb{R}^d \) be a grid point set as in Definition 2.25 and let \( x \in \Omega_h \) with \( x \pm h^\pm e_i \in \overline{\Omega}_h, \( h^\pm_i = \alpha^\pm_i h_i, |\alpha^\pm_i| \leq 1, 1 \leq i \leq d \). Then the forward and backward difference quotient for the approximation of \( \partial u/\partial x_i \) in \( x \) are given by

\[
D^+_h,i u(x) := \frac{(u(x + h^+_i e_i) - u(x))}{h^+_i},
\]

\[
D^-_h,i u(x) := \frac{(u(x) - u(x - h^-_i e_i))}{h^-_i}.
\]
The central difference quotient for the second partial derivatives \( \partial^2 u/\partial x_i^2 \) is given by
\[
D_{h,i,i}^2 u(x) := \frac{2}{h_i^+ + h_i^-} \left( \frac{u(x + h_i^+ e_i) - u(x)}{h_i^+} + \frac{u(x - h_i^- e_i) - u(x)}{h_i^-} \right).
\]

For \( d = 2 \) and \( x = (x_1, x_2) \in \Omega_h \) with \( (x_1 \pm \alpha_i^+ h, x_2 \pm \alpha_2^+ h) \in \overline{\Omega}_h \), the discrete Laplacian \( \Delta_h \) has the representation
\[
\Delta_h u(x) := \frac{2}{h^2} \left( \frac{1}{\alpha_1^+ (\alpha_1^+ + \alpha_1^-)} u(x_1 + \alpha_1^+ h, x_2) + \frac{1}{\alpha_1^- (\alpha_1^- + \alpha_1^+)} u(x_1 - \alpha_1^- h, x_2) + \frac{1}{\alpha_2^+ (\alpha_2^+ + \alpha_2^-)} u(x_1, x_2 + \alpha_2^+ h) + \frac{1}{\alpha_2^- (\alpha_2^- + \alpha_2^+)} u(x_1, x_2 - \alpha_2^- h) - \right.
\]
\[
\left. \frac{1}{\alpha_1^+ \alpha_1^-} u(x_1, x_2) \right) \] (2.34)

Remark 2.26. (Miscellaneous)
(i) In grid points close to the boundary, the central difference quotient for the second partial derivatives and the discrete Laplacian are in general only approximations of order \( O(h) \).
(ii) The definition of the difference quotients and the definition of the discrete Laplacian can be generalized to non-equidistant grids. We refer to Strikwerda.

Definition 2.27. (Shortley-Weller approximation of the normal derivative)
Let \( \nu \cdot \nabla u \) be the normal derivative in \( x_{\Gamma} \in \Gamma_h \), where \( \nu \) stands for the outward normal unit vector in \( x_{\Gamma} \), and let \( x_{\Gamma}^* \in \Omega_h \) be the first intersection of the normal through \( x_{\Gamma} \) with a grid line (see Figure 10 (right)). Further, let \( x_i, 1 \leq i \leq 2 \), be those grid points on that grid line close to \( x_{\Gamma}^* \), where we suppose that the grid is sufficiently fine to ensure \( x_i \in \Omega_h, 1 \leq i \leq 2 \). We refer to \( \tilde{u}(x_{\Gamma}^*) \) as the linearly interpolated value with respect to \( u(x_1) \) and \( u(x_2) \). Then
\[
D_h^\nu u(x_{\Gamma}) := \frac{u(x_{\Gamma}) - \tilde{u}(x_{\Gamma}^*)}{|x_{\Gamma} - x_{\Gamma}^*|} \] (2.37)
is an approximation of the normal derivative in \( x_{\Gamma} \).

Definition 2.28. (Grid function)
Let \( \overline{\Omega}_h \) be a grid point set as in Definition 2.25. A function \( u_h : \overline{\Omega}_h \to \mathbb{R} \) is called a grid function. We refer to \( C(\overline{\Omega}_h) \) as the normed linear space
of grid functions equipped with the discrete analogue of the maximum norm

\[
\|u_h\|_h := \max_{x \in \Omega_h} |u_h(x)|.
\]  

(2.38)

We consider a difference approximation of Poisson’s equation (2.11) with inhomogeneous Dirichlet boundary conditions based on the use of the discrete Laplacian. Given grid functions \(f_h \in C(\Omega_h)\) and \(u_h^D \in C(\Gamma_h)\), we are looking for a grid function \(u_h \in C(\Omega_h)\) such that

\[
-\Delta_h u_h = f_h \quad \text{in } \Omega_h,
\]

(2.39a)

\[
u_h = u_h^D \quad \text{on } \Gamma_h.
\]

(2.39b)

The existence and uniqueness of a solution can be shown either by means of a discrete Green’s function and a discrete maximum principle or by the study of the resulting linear algebraic system. Here, we choose the latter approach and restrict ourselves to the case \(d = 2\) and \(\Omega = (a, b)^2\) as well as a uniform grid \(\Omega_h\). The unknown values of the grid function \(u_h\) the interior grid points \(x \in \Omega_h\) can be interpreted as the components of a vector. The structure of the associated linear algebraic system depends on the ordering of the grid points.

**Definition 2.29. (Lexicographic ordering of grid points)**

Let \(\Omega = (a, b)^2\), \(a, b \in \mathbb{R}, a < b\), and let \(\Omega_h\) be a uniform grid of step size \(h := (b - a)/(N + 1), N \in \mathbb{N}\), with the interior grid points \(x = (x_i, x_j)^T, 1 \leq i, j \leq N\), where \(x_i = a + ih\). The ordering of the interior grid points from bottom to top and from left to right according to

\[x_{(i-1)N+j} = (a + ih, a + jh)^T, \quad 1 \leq i, j \leq N,\]

is called the lexicographic ordering.

The ordering such that grid points with even \(i + j\) are counted first followed by those with odd \(i + j\) is said to be the checkerboard ordering.

For an ordering of the interior grid points according to Definition 2.29, the values of the grid function in the interior grid points can be identified with a vector with \(n_h := N^2\) components. For ease of notation we choose the same symbol as for the grid function, i.e., we define \(u_h \in \mathbb{R}^{n_h}\) by means of \(u_h, i := u_h(x_i), 1 \leq i \leq n_h\). Then (2.39a), (2.39a) is equivalent with a linear algebraic system

\[
A_h u_h = b_h
\]  

(2.40)
with a coefficient matrix $A_h \in \mathbb{R}^{n_h \times n_h}$ and a vector $b_h \in \mathbb{R}^{n_h}$, whose components are given by the values of $f_h$ and $u_D^h$.

In case of a lexicographic ordering, $A_h$ is a block tridiagonal matrix

$$A_h = h^{-2} \text{tridiag}(-I, T, -I),$$

where the diagonal blocks $T \in \mathbb{R}^{N \times N}$ are tridiagonal matrices of the form

$$T = \text{tridiag}(-1, 4, -1)$$

and the sub- and super-diagonal blocks are the negative $N \times N$ unit matrix.

The matrix $A_h$ is a symmetric, positive definite matrix with the eigenvalues

$$\lambda_{ij}(A_h) = 4h^{-2}\left(\sin^2\left(\frac{i\pi h}{2}\right) + \sin^2\left(\frac{j\pi h}{2}\right)\right), \quad 1 \leq i, j \leq N,$$

and the associated orthonormal eigenvectors $x^{(ij)} \in \mathbb{R}^{n_h}, 1 \leq i, j \leq N,$

with the components

$$x_{k\ell}^{(ij)} = \frac{h}{2} \sin\left(ikh\pi\right) \sin\left(jh\ell\pi\right), \quad 1 \leq k, \ell \leq N.$$

The properties of $A_h$ imply the unique solvability of the linear algebraic system (2.40) and hence the unique solvability of the difference approximation (2.39a),(2.39b). They also imply the convergence of $u_h$ to the solution $u$, provided that $f_h \rightarrow f$ and $u_D^h \rightarrow u_D$ as $h \rightarrow 0$.

**Definition 2.30. (Convergence of finite difference approximations)**

For $\Omega := (a, b)^2, a, b \in \mathbb{R}, a < b$, and a uniform grid $\Omega_h$ with step size $h > 0$ let $u : \Omega \rightarrow \mathbb{R}$ be the classical solution of Poisson’s equation (2.11) with inhomogeneous Dirichlet boundary conditions and let $u_h \in C(\Omega_h)$ be the solution of its finite difference approximation (2.39a),(2.39b). The grid function $e_h(x) = u(x) - u_h(x), x \in \Omega_h$, is called the global discretization error. The finite difference approximation (2.39a),(2.39b) is said to converge, if

$$\|e_h\|_h \rightarrow 0 \quad \text{as } h \rightarrow 0.$$

It is said to be convergent of order $p$, if there exists a constant $C > 0$, independent of $h$, such that for sufficiently small $h$ it holds

$$\|e_h\|_h \leq C h^p.$$

Sufficient conditions for the convergence are the consistency of the finite difference approximation with the boundary value problem and the stability of the finite difference approximation. The consistency is
measured by means of the local discretization error which results, if the solution \( u \) of the boundary value problem is inserted into the finite difference approximation.

**Definition 2.31. (Consistency and order of consistency)**

Let \( u \) be the classical solution of the boundary value problem and let \( f_h \) and \( u^D_h \) be the grid functions in (2.39a),(2.39b). Then the grid function \( \tau_h \in C(\Omega_h) \) given by

\[
\tau_h(x) := \begin{cases} \ -\Delta_h u(x) - f_h(x) & , \ x \in \Omega_h \\ \ u(x) - u^D_h(x) & , \ x \in \Gamma_h \end{cases}
\]

is called the local discretization error. The finite difference approximation (2.39a),(2.39b) is said to be consistent with the boundary value problem, if

\[
\| \tau_h \|_h \to 0 \quad \text{as} \quad h \to 0.
\]

It is said to be consistent of order \( p \), if there exists a constant \( C > 0 \), independent of \( h \), such that for sufficiently small \( h \) it holds

\[
\| \tau_h \|_h \leq C h^p.
\]

**Lemma 2.32. (Order of consistency for the finite difference approximation of Poisson’s equation)**

Assume that the solution \( u \) of Poisson’s equation (2.11) with inhomogeneous Dirichlet boundary conditions satisfies \( u \in C^4(\Omega) \) and that

\[
\max_{x \in \Omega_h} |f(x) - f_h(x)| = O(h^2), \quad \max_{x \in \Gamma_h} |u^D(x) - u^D_h(x)| = O(h^2).
\]

Then the finite difference approximation (2.39a),(2.39b) is consistent with Poisson’s equation (2.11) with inhomogeneous Dirichlet boundary conditions of order \( p = 2 \).

**Proof.** Under the assumptions on \( f_h \) and \( u^D_h \) the proof follows from the approximation property of the discrete Laplacian. \( \Box \)

**Remark 2.33. (General bounded domains)**

The result of Lemma 2.32 can be easily generalized to boundary value problems in higher dimensions.

For general bounded domains and the use of the Shortley-Weller approximation, in general the order of consistency is reduced to \( p = 1 \).
2.2.3 Sobolev spaces

We give a brief survey on Sobolev spaces and its associated dual spaces and trace spaces. For details we refer to Adams and Tartar.

Let $\Omega$ be a bounded domain in $\mathbb{R}^d$. We denote by $\overline{\Omega}$ the closure and by $\Gamma = \partial \Omega := \overline{\Omega} \setminus \Omega$ the boundary of $\Omega$. Moreover, let $\Omega_e := \mathbb{R}^d \setminus \overline{\Omega}$ the associated exterior domain.

We define $C^m(\Omega), m \in \mathbb{N}_0$, as the linear space of all continuous functions in $\Omega$ with continuous partial derivatives $D^\alpha u, |\alpha| \leq m$. We note that $C^m(\Omega)$ is a Banach space with respect to the norm

$$\|u\|_{C^m(\Omega)} := \max_{0 \leq |\alpha| \leq m} \sup_{x \in \Omega} |D^\alpha u(x)|.$$

Further, we define $C^{m,\alpha}(\Omega), m \in \mathbb{N}_0, 0 < \alpha < 1$, as the linear space of all functions in $C^m(\Omega)$ whose $m$-th partial derivatives are Hölder continuous, i.e., for all $\beta \in \mathbb{N}_0^d$ with $|\beta| = m$ there exist constants $\Gamma_\beta > 0$ such that for all $x, y \in \Omega$ it holds

$$|D^\beta u(x) - D^\beta u(y)| \leq \Gamma_\beta |x - y|^\alpha.$$

$C^{m,\alpha}(\Omega)$ is a Banach space with respect to the norm

$$\|u\|_{C^{m,\alpha}(\Omega)} := \|u\|_{C^m(\Omega)} + \max_{|\beta| = m} \sup_{x,y \in \Omega} \frac{|D^\beta u(x) - D^\beta u(y)|}{|x - y|^\alpha}.$$

We refer to $C^m(\Omega)$ as the subspace of $C^{m,\alpha}(\Omega)$ of all functions with compact support in $\Omega$. Moreover, $C^\infty(\Omega)$ stands for the linear space of all arbitrarily often continuously differentiable functions and $C^\infty_0(\Omega)$ is the subspace of $C^\infty(\Omega)$ of all functions with compact support in $\Omega$.

In the sequel, we restrict ourselves to Lipschitz-domains which are defined as follows:

**Definition 2.34. (Lipschitz domain)**

A bounded domain $\Omega \subset \mathbb{R}^d$ with boundary $\Gamma$ is called a Lipschitz domain, if there exist constants $\alpha > 0, \beta > 0$, and a finite number of local coordinate systems $(x_1^r, x_2^r, \ldots, x_d^r), 1 \leq r \leq R$, as well as local Lipschitz-continuous mappings

$$a_r : \{ (\hat{x}^r) = (x_2^r, \ldots, x_d^r) \in \mathbb{R}^{d-1} \mid |x_i^r| \leq \alpha , \ 2 \leq i \leq d \} \rightarrow \mathbb{R}$$

such that

$$\Gamma = \bigcup_{r=1}^R \{ (x_1^r, \hat{x}^r) \mid x_1^r = a_r(\hat{x}^r) , \ |\hat{x}^r| < \alpha \},$$

$$\{(x_1^r, \hat{x}^r) \mid a_r(\hat{x}^r) < x_1^r < a_r(\hat{x}^r) + \beta , \ |\hat{x}^r| < \alpha \} \subset \Omega , \ 1 \leq r \leq R,$$

$$\{(x_1^r, \hat{x}^r) \mid a_r(\hat{x}^r) - \beta < x_1^r < a_r(\hat{x}^r) , \ |\hat{x}^r| < \alpha \} \subset \Omega_e , \ 1 \leq r \leq R.$$
The geometric interpretation of the conditions in Definition 2.34 is that both \( \Omega \) and the exterior domain \( \Omega_e \) are located on exactly one side of the boundary \( \Gamma \).

If the mappings in Definition 2.34 are smoother, the domains are called \( C^m \)-domains of \( C^{m,\alpha} \)-domains, \( m \in \mathbb{N} \).

**Definition 2.35. (C\(^m\)-domain and C\(^{m,\alpha}\)-domain)**

A Lipschitz domain \( \Omega \subset \mathbb{R}^d \) is said to be a \( C^m \)-domain (\( C^{m,\alpha} \)-domain), if the functions \( a_r, 1 \leq r \leq R \), in Definition 2.34 are \( C^m \)-functions (\( C^{m,\alpha} \)-functions).

We denote by \( L^2(\Omega) \) the linear space of Lebesgue square integrable Functions in \( \Omega \). We note that \( L^2(\Omega) \) is a Hilbert space with respect to the inner product

\[
(v, w)_{0,\Omega} := \int_{\Omega} vw \, dx.
\]

The associated norm will be denoted by \( \| \cdot \|_{0,\Omega} \).

Sobolev spaces are based on the concept of weak (distributional) derivatives:

**Definition 2.36. (Weak derivative)**

Let \( u \in L^2(\Omega) \) and \( \alpha \in \mathbb{N}_0^d \). The function \( u \) is said to be differentiable in the weak sense, if there exists a function \( v \in L^2(\Omega) \) such that

\[
\int_{\Omega} u D^{\alpha} \varphi \, dx = (-1)^{|\alpha|} \int_{\Omega} v \varphi \, dx \quad , \quad \varphi \in C_0^\infty(\Omega).
\]

We set \( D^{\alpha}_w u := v \) and refer to \( D^{\alpha}_w u \) as the weak (distributional) derivative of \( u \).

**Example 2.37. (Weak derivative)**

Let \( d = 1 \) and \( \Omega := (-1, +1) \). The function \( u(x) := |x|, x \in \Omega \), is not differentiable in the origin. However, it has a weak derivative \( D^1_w u \), given by

\[
D^1_w u = \begin{cases} 
-1 & , \ x < 0 \\
+1 & , \ x > 0 
\end{cases}.
\]
For \( \varphi \in C^\infty_0(\Omega) \) partial integration yields
\[
\int_{-1}^{+1} u(x) D^1 \varphi(x) \, dx = \int_{-1}^{0} u(x) D^1 \varphi(x) \, dx + \int_{0}^{+1} u(x) D^1 \varphi(x) \, dx =
\]
\[
= - \int_{-1}^{0} D^1_w u(x) \varphi(x) \, dx + (u \varphi)|^0_{-1} - \int_{0}^{+1} D^1_w u(x) \varphi(x) \, dx + (u \varphi)|^1_0 =
\]
\[
= \int_{-1}^{+1} D^1_w u(x) \varphi(x) \, dx - [u(0)] \varphi(0),
\]
where \([u(0)] := u(0^+) - u(0^-)\) denotes the jump of \( u \) in \( x = 0 \). Since \( u \) is continuous, it follows that \([u(0)] = 0\) which allows to conclude.

**Definition 2.38. (Sobolev space)**

Let \( m \in \mathbb{N}_0 \). The linear space
\[
W^m(\Omega) := \{ u \in L^2(\Omega) \mid D^\alpha_w u \in L^2(\Omega), |\alpha| \leq m \}
\]
is called a Sobolev space. It is a Hilbert space with respect to the inner product
\[
(u, v)_{m, \Omega} := \sum_{|\alpha| \leq m} \int_{\Omega} D^\alpha_w u D^\alpha_w v \, dx.
\]
The associated norm will be denoted by \( \| \cdot \|_{m, \Omega} \). We further refer to \( | \cdot |_{m, \Omega} \) as the seminorm
\[
|u|_{m, \Omega} := \left( \sum_{|\alpha|=m} \int_{\Omega} (D^\alpha_w)^2 \, dx \right)^{1/2}.
\]

A natural question is whether \( C^m(\Omega) \) is dense in \( W^m(\Omega) \). To this end we consider the linear space
\[
C^{m,*}(\Omega) := \{ u \in C^m(\Omega) \mid \| u \|_{m,p,\Omega} < \infty \},
\]
which is a normed linear space with respect to the norm \( \| \cdot \|_{m,\Omega} \), but not complete. We denote by \( H^m(\Omega) \) its completion with respect to the \( \| \cdot \|_{m,\Omega} \)-norm
\[
H^m(\Omega) := \overline{C^{m,*}(\Omega)}_{\| \cdot \|_{m,\Omega}}.
\]
A famous result by Meyers and Serrin says that the spaces \( W^m(\Omega) \) and \( H^m(\Omega) \) coincide.

We define \( H^m_0(\Omega), m \in \mathbb{N}, \) as the completion of \( (C^m_0(\Omega), \| \cdot \|_{m,\Omega}) \).
Definition 2.39. (Sobolev spaces with negative integer indices)

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain and let $m$ be a negative integer. Then the Sobolev space $W^m(\Omega)$ is defined as the dual space of $W^{-m}(\Omega)$.

Definition 2.40. (Sobolev spaces with broken indices)

For $\Omega = \mathbb{R}^d$ we define the Sobolev space $H^s(\mathbb{R}^d)$ with broken index $s \in \mathbb{R}_+$ by means of the Fourier transform $\hat{v}$ of a function $v \in C_0^\infty(\mathbb{R}^d)$

$$\hat{v}(\xi) := (\frac{1}{2\pi})^{d/2} \int_{\mathbb{R}^d} \exp(-i\xi \cdot x) v(x) \, dx.$$ 

The corresponding Sobolev norm is given by

$$\|v\|_{s, \mathbb{R}^d} := \|(1 + |\cdot|^2)^{s/2} \hat{v}(\cdot)\|_{0, \mathbb{R}^d}.$$ 

We set

$$H^s(\mathbb{R}^d) := C_0^\infty(\mathbb{R}^d)^{\|\cdot\|_{s, \mathbb{R}^d}}.$$ 

If $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain, for $s = m + \lambda$, $m \in \mathbb{N}_0$, $0 \leq \lambda < 1$, we define $H^s(\Omega)$ by means of

$$\|v\|_{s, \Omega} := \left(\frac{1}{\text{diam}(\Omega)^2}\|v\|_{0, \Omega}^2 + |v|_{s, \Omega}^2\right)^{1/2},$$

where $|\cdot|_{s, \Omega}$ stands for the seminorm

$$|v|_{s, \Omega} := \sum_{1 \leq |\alpha| \leq m} \|D^\alpha v\|_{0, \Omega}^2 + \sum_{|\alpha| = m} \int_{\Omega} \int_{\Omega} \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x - y|^{d + 2\lambda}} \, dx \, dy.$$ 

For $\Sigma \subseteq \Gamma = \partial \Omega$ we define $H^s(\Sigma)$ according to

$$H^s(\Sigma) := \{v \in L^2(\Sigma) \mid \|v\|_{s, \Sigma} < \infty\},$$

equipped with the norm

$$\|v\|_{s, \Sigma} := \left(\frac{1}{\text{diam}(\Sigma)^2}\|v\|_{0, \Sigma}^2 + |v|_{s, \Sigma}^2\right)^{1/2},$$

where

$$|v|_{s, \Sigma} := \sum_{1 \leq |\alpha| \leq m} \|D^\alpha v\|_{0, \Sigma}^2 + \sum_{|\alpha| = m} \int_{\Sigma} \int_{\Sigma} \frac{|D^\alpha v(x) - D^\alpha v(y)|^2}{|x - y|^{d - 1 + 2\lambda}} \, d\sigma(x) \, d\sigma(y).$$
For a bounded Lipschitz domain \( \Omega \subset \mathbb{R}^d \) with boundary \( \Gamma \) and a function \( u \in C(\overline{\Omega}) \) the pointwise restriction of \( u \) to the boundary \( \Gamma \) is well defined. For \( d > 1 \) the Sobolev space \( H^1(\Omega) \) is not continuously embedded in \( C(\overline{\Omega}) \) and hence, the pointwise restriction to \( \Gamma \) does not make sense. However, the mapping \( u \mapsto u|_\Gamma \), which is well defined for \( C(\overline{\Omega}) \), can be extended to a continuous mapping \( H^1(\Omega) \rightarrow H^{1/2}(\Gamma) \).

**Theorem 2.41. (Trace theorem)**

Let \( \Omega \subset \mathbb{R}^d \) be a bounded Lipschitz domain with a \( C^{1,1} \)-boundary \( \Gamma \). The mapping

\[
  u \mapsto u|_\Gamma,
\]

which is well defined for \( u \in C^{1,1}(\overline{\Omega}) \), can be uniquely extended to a continuous surjective mapping

\[
  H^1(\Omega) \rightarrow H^{1/2}(\Gamma)
\]

which admits a continuous right inverse.

**Definition 2.42. (Trace, trace space)**

Under the assumptions of Theorem 2.41 the function \( u|_\Gamma \) is called the trace of \( u \) on \( \Gamma \). The space \( H^{1/2}(\Gamma) \) is called the trace space of \( H^1(\Omega) \). For functions \( v \in H^1(\Omega) \) the trace inequality

\[
  \|v|_\Gamma\|_{1/2,\Gamma} \leq C_\Gamma \|v\|_{1,\Omega}
\]

holds true with a constant \( C_\Gamma > 0 \). Moreover, \( H^{1/2}(\Gamma) \) is continuously embedded in \( L^2(\Gamma) \), i.e., there exists a constant \( C_0 > 0 \) such that for \( v \in H^{1/2}(\Gamma) \) it holds

\[
  \|v\|_{0,\Gamma} \leq C_0 \|v\|_{1/2,\Gamma}.
\]

**Remark 2.43. (Polyhedral domains)**

Theorem 2.41 is not immediately applicable to the important special case of polyhedral domains, since the boundary of such domains is no \( C^{1,1} \)-boundary. On the other hand, the boundary \( \Gamma \) of bounded polyhedral domains consist of a finite number of \( C^{1,1} \)-boundaries. For the extension of Theorem 2.41 to bounded polyhedral domains we refer to Grisvard.

**Remark 2.44. (Miscellaneous)**

Traces and trace spaces for functions in \( H^m(\Omega) \), \( m \in \mathbb{N}, m \geq 2 \), can be defined as well. Again, we refer to Grisvard.
2.2.4 Finite element approximations

For a bounded Lipschitz-domain $\Omega \subset \mathbb{R}^d$ with boundary $\Gamma = \partial \Omega$ and $f \in L^2(\Omega)$ we consider the boundary value problem for the Poisson equation with homogeneous Dirichlet boundary conditions

\begin{align}
\text{(2.50a)} & \quad -\Delta u = f \quad \text{in } \Omega, \\
\text{(2.50b)} & \quad u = 0 \quad \text{on } \Gamma,
\end{align}

Finite element approximations of (2.50a),(2.50b) are based on the concept of a weak solution. We multiply (2.50a) with a test function $v \in C^1_0(\Omega)$ and integrate over $\Omega$:

\begin{equation}
- \int_{\Omega} \Delta uv \, dx = \int_{\Omega} fv \, dx.
\end{equation}

By Green’s formula (2.28), for the integral on the left-hand side in (2.51) we find

\begin{equation}
- \int_{\Omega} \Delta uv \, dx = \int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\Gamma} \nu \cdot \nabla u \, v \, d\sigma(x).
\end{equation}

The boundary integral in (2.52) vanishes due to $v|_{\Gamma} = 0$. Hence, we obtain

\begin{equation}
\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} fv \, dx.
\end{equation}

We recall that the Sobolev space $H^1_0(\Omega)$ is the completion of $C^1_0(\Omega)$ with respect to the $\| \cdot \|_{1,\Omega}$-norm. This leads to the following definition of a weak solution.

**Definition 2.45.** (Weak solution of the Poisson equation with homogeneous Dirichlet boundary conditions)

A function $u \in H^1_0(\Omega)$ is called a weak solution of (2.50a),(2.50b), if for all $v \in H^1_0(\Omega)$ the variational equation

\begin{equation}
\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} fv \, dx
\end{equation}

is satisfied.

Next, we consider the Poisson equation with inhomogeneous Dirichlet boundary conditions

\begin{align}
\text{(2.55a)} & \quad -\Delta u = f \quad \text{in } \Omega, \\
\text{(2.55b)} & \quad u = g \quad \text{on } \Gamma,
\end{align}
where \( g \in H^{1/2}(\Gamma) \). We note that Theorem 2.41 implies the existence of a function \( \tilde{g} \in H^1(\Omega) \) such that \( \tilde{g}|_{\Gamma} = g \).

The weak solution of (2.55a),(2.55b) is defined as follows:

**Definition 2.46. (Weak solution of the Poisson equation with inhomogeneous Dirichlet boundary conditions)**
A function \( u \in H^1(\Omega) \) is called a weak solution of (2.50a),(2.50b), if \( u - \tilde{g} \in H^1_0(\Omega) \) and for all \( v \in H^1_0(\Omega) \) the variational equation

\[
\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} fv \, dx
\]

is satisfied.

Given functions \( f \in L^2(\Omega) \) and \( g \in L^2(\Gamma) \), we consider the Poisson equation with inhomogeneous Neumann boundary conditions

\[
\begin{align*}
(2.57a) \quad -\Delta u &= f \quad \text{in } \Omega, \\
(2.57b) \quad \nu \cdot \nabla u &= g \quad \text{on } \Gamma,
\end{align*}
\]

If we multiply (2.57a) with a test function \( v \in C^1(\Omega) \) and integrate over \( \Omega \), it follows from (2.51) and (2.52) that

\[
\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} fv \, dx + \int_{\Gamma} gv \, d\sigma(x).
\]

This gives rise to the following definition of a weak solution of (2.57a), (2.57b):

**Definition 2.47. (Weak solution of the Poisson equation with inhomogeneous Neumann boundary conditions)**
A function \( u \in H^1(\Omega) \) is called a weak solution of (2.57a), (2.57b), if for all \( v \in H^1(\Omega) \) the variational equation

\[
\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} fv \, dx + \int_{\Gamma} gv \, d\sigma(x)
\]

is satisfied.

For the proof of the existence and uniqueness of solutions of (2.54),(2.56), and (2.59) we introduce the notion of a bounded, \( V \)-elliptic bilinear form on a Hilbert space \( V \):
**Definition 2.48. (Bounded bilinear form)**

Let $V$ be a Hilbert space with norm $\| \cdot \|_V$. A bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ is said to be bounded, if there exists a constant $C > 0$ such that

$$|a(u, v)| \leq C \|u\|_V \|v\|_V, \quad u, v \in V. \quad (2.60)$$

**Definition 2.49. (V-elliptic bilinear form)**

Let $V$ be a Hilbert space with norm $\| \cdot \|_V$. A bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ is said to be $V$-elliptic, if there exists a constant $C > 0$ such that

$$a(u, u) \geq \alpha \|u\|_V^2, \quad u \in V. \quad (2.61)$$

Given a bounded, $V$-elliptic bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ and a bounded linear functional $\ell : V \to \mathbb{R}$, we consider the variational equation:

Find $u \in V$ such that for all $v \in V$ it holds

$$a(u, v) = \ell(v). \quad (2.62)$$

The existence and uniqueness of a solution of (2.62) follows from the Lemma of Lax and Milgram:

**Theorem 2.50. (Lemma of Lax and Milgram)**

Let $V$ be a Hilbert space, let $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a bounded, $V$-elliptic bilinear form, and let $\ell : V \to \mathbb{R}$ be a bounded linear functional. Then the variational equation (2.62) admits a unique solution $u \in V$.

**Proof.** We refer to Brenner/Scott. \qed

In order to apply the Lemma of Lax and Milgram to (2.54),(2.56), and (2.59), we further need the following two Poincaré-Friedrichs inequalities:

**Theorem 2.51. (Poincaré-Friedrichs inequalities)** Let $u \in H^1(\Omega)$. Then, there exist constants $C_1(\Omega) > 0$ and $C_2(\Omega) > 0$, depending only on the domain $\Omega$, such that

$$\|u\|_{0, \Omega} \leq C_1(\Omega) \left( |u|_{1, \Omega} + \int_F u \, d\sigma(x) \right), \quad (2.63a)$$

$$\|u\|_{0, \Omega} \leq C_2(\Omega) \left( |u|_{1, \Omega} + \int_{\Omega} u \, dx \right), \quad (2.63b)$$

**Proof.** We refer to Brenner/Scott. \qed
It follows from (2.63a) that on $H_0^1(\Omega)$ the seminorm $|\cdot|_{1,\Omega}$ is a norm which is equivalent to the $\|\cdot\|_{1,\Omega}$-norm
\begin{equation}
|u|^2_{1,\Omega} \leq \|u\|^2_{0,\Omega} + |u|^2_{1,\Omega} \leq (1 + C_1(\Omega)^2) |u|^2_{1,\Omega}.
\end{equation}

For (2.54) the Hilbert space is $V = H_0^1(\Omega)$, the bilinear form $a(\cdot, \cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is given by

$$a(u,v) := \int_{\Omega} \nabla u \cdot \nabla v \, dx,$$

and the linear functional $\ell : H_0^1(\Omega) \to \mathbb{R}$ reads

$$\ell(v) := \int_{\Omega} fv \, dx.$$

Obviously, we have

$$a(u,u) = |u|^2_{1,\Omega}, \quad |a(u,v)| \leq |u|_{1,\Omega} |v|_{1,\Omega}.$$

$$|\ell(v)| \leq \|f\|_{0,\Omega} \|v\|_{0,\Omega} \leq C_1(\Omega) \|f\|_{0,\Omega} |v|_{1,\Omega},$$

and hence, due to (2.64) we deduce the boundedness and $H_0^1(\Omega)$-ellipticity of $a(\cdot, \cdot)$ as well as the boundedness of the linear functional $\ell(\cdot)$. The Lemma of Lax and Milgram asserts the existence and uniqueness of a weak solution.

With $z := u - \tilde{g}$ the variational problem (2.56) can be reformulated as

\begin{equation}
\int_{\Omega} \nabla z \cdot \nabla v \, dx = \int_{\Omega} fv \, dx - \int_{\Omega} \nabla \tilde{g} \cdot \nabla v \, dx.
\end{equation}

The bilinear form $a(\cdot, \cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is again given by

$$a(u,v) := \int_{\Omega} \nabla u \cdot \nabla v \, dx,$$

whereas the linear functional $\ell : H_0^1(\Omega) \to \mathbb{R}$ is as follows:

$$\ell(v) := \int_{\Omega} fv \, dx - \int_{\Omega} \nabla \tilde{g} \cdot \nabla v \, dx.$$

Since the boundedness and $H_0^1(\Omega)$-ellipticity of $a(\cdot, \cdot)$ has been shown before for (2.54), we only have to verify the boundedness of the linear functional $\ell(\cdot)$:

$$|\ell(v)| \leq \|f\|_{0,\Omega} \|v\|_{0,\Omega} + |\tilde{g}|_{1,\Omega} |v|_{1,\Omega} \leq \left( C_1(\Omega) \|f\|_{0,\Omega} + |\tilde{g}|_{1,\Omega} \right) |v|_{1,\Omega}.$$
Hence, by the Lemma of Lax and Milgram we conclude that (2.65) has a unique solution which implies existence and uniqueness of a solution of (2.56).

In order to prove the existence and uniqueness of a solution of (2.59) we note that the functions \( f \) and \( g \) have to satisfy the compatibility condition

\[
\int_{\Omega} f \, dx + \int_{\Gamma} g \, d\sigma(x) = 0.
\]

We consider the quotient space \( H^1(\Omega) \setminus \mathbb{R} \) with the norm

\[
\|v\|_{H^1(\Omega) \setminus \mathbb{R}} = \inf_{a \in \mathbb{R}} \|v - a\|_{1,\Omega}.
\]

It follows from (2.63b) that on \( H^1(\Omega) \setminus \mathbb{R} \) the seminorm \( \cdot \rvert_{1,\Omega} \) and the norm \( \| \cdot \rvert_{H^1(\Omega) \setminus \mathbb{R}} \) are equivalent:

\[
|v|_{1,\Omega}^2 \leq \|v\|^2_{H^1(\Omega) \setminus \mathbb{R}} = \inf_{a \in \mathbb{R}} \|v - a\|^2_{1,\Omega} \leq \|v - \int_{\Omega} v \, dx\|^2_{1,\Omega} \leq
\|

\int_{\Omega} v \, dx\|^2_{0,\Omega} + |v|_{1,\Omega}^2 \leq (1 + C_2(\Omega)^2) \|v\|_{1,\Omega}^2.
\]

The bilinear form \( a(\cdot, \cdot) : (H^1(\Omega) \setminus \mathbb{R}) \times (H^1(\Omega) \setminus \mathbb{R}) \rightarrow \mathbb{R} \) and the linear functional \( \ell(\cdot) : (H^1(\Omega) \setminus \mathbb{R})H^1(\Omega) \setminus \mathbb{R} \rightarrow \mathbb{R} \) are given by

\[
a(u, v) := \int_{\Omega} \nabla u \cdot \nabla v \, dx,
\]

\[
\ell(v) := \int_{\Omega} f v \, dx + \int_{\Gamma} g v \, d\sigma(x).
\]

The boundedness and \((H^1(\Omega) \setminus \mathbb{R})\)-ellipticity of \( a(\cdot, \cdot) \) follow readily from (2.67). As far as the boundedness of \( \ell(\cdot) \) is concerned, in view of
\( (2.48), (2.49), (2.63b), \) and \( (2.66) \) we have
\[
|\ell(v)| = |\int_{\Omega} fv \, dx + \int_{\Gamma} gv \, d\sigma(x)| = \\
|\int_{\Omega} f(v - \int_{\Omega} v \, dx) \, dx + \int_{\Gamma} g(v - \int_{\Omega} v \, dx) \, d\sigma(x)| \leq \\
\|f\|_{0,\Omega} \|v - \int_{\Omega} v \, dx\|_{0,\Omega} + \|g\|_{0,\Gamma} \|v - \int_{\Omega} v \, dx\|_{0,\Gamma} \leq \\
(C_2(\Omega) \|f\|_{0,\Omega} + C_0 C_\Gamma (1 + C_2(\Omega)^2 \|g\|_{0,\Gamma})^{1/2}) |v|_{1,\Omega}.
\]

The Lemma of Lax and Milgram implies the existence and uniqueness of a solution \( u \in H^1(\Omega) \setminus \mathbb{R} \) of \( (2.59) \), i.e., the solution is unique up to a constant.

Let \( V \subseteq H^1(\Omega) \) be a Hilbert space, \( a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R} \) a bounded, \( V \)-elliptic bilinear form, and \( \ell : V \rightarrow \mathbb{R} \) a bounded linear functional. The Lemma of Lax and Milgram implies that the variational equation \( (2.62) \) admits a unique solution \( u \in V \).

The Galerkin method provides an approximation of \( u \in V \) by a function \( u_h \) in a finite dimensional subspace \( V_h \subset V \) with \( \dim V_h = n_h \) which solves the restriction of \( (2.62) \) to \( V_h \), i.e.,
\[
(2.68) \quad a(u_h, v_h) = \ell(v_h), \quad v_h \in V_h.
\]

Another application of the Lemma of Lax and Milgram shows the existence and uniqueness of a solution \( u_h \in V_h \) of \( (2.68) \). If the bilinear form is symmetric, i.e.,
\[
a(u, v) = a(v, u), \quad u, v \in V;
\]
the variational equation \( (2.68) \) is referred to as the Ritz-Galerkin method.

It is easily seen that the Galerkin method leads to a linear algebraic system, once a basis \( \{\varphi_h^{(i)}\}_{i=1}^{n_h} \) of \( V_h \) is provided. Then, the solution \( u_h \in V_h \) of \( (2.68) \) can be written as a linear combination of the basis functions according to
\[
u_h = \sum_{j=1}^{n_h} c_j \varphi_h^{(j)}, \quad c_j \in \mathbb{R}, \ 1 \leq j \leq n_h.
\]
Obviously, the variational equation (2.68) can be equivalently written as

\[
\sum_{j=1}^{n_h} a(\varphi_h^{(j)}, \varphi_h^{(i)}) c_j = \ell(\varphi_h^{(i)}), \quad 1 \leq i \leq n_h,
\]

which represents the linear algebraic system

\[
A_h c_h = b_h
\]

in the unknown vector \(c_h = (c_1, \ldots, c_{n_h})^T\). The stiffness matrix \(A_h \in \mathbb{R}^{n_h \times n_h}\) and the load vector \(b_h \in \mathbb{R}^{n_h}\) are given by

\[
A_h := \begin{pmatrix}
    a(\varphi_h^{(1)}, \varphi_h^{(1)}) & \cdots & a(\varphi_h^{(n_h)}, \varphi_h^{(1)}) \\
    \vdots & \ddots & \vdots \\
    a(\varphi_h^{(1)}, \varphi_h^{(n_h)}) & \cdots & a(\varphi_h^{(n_h)}, \varphi_h^{(n_h)})
\end{pmatrix},
\]

\[
b_h^{(i)} := \ell(\varphi_h^{(i)}), \quad 1 \leq i \leq n_h.
\]

We remark that the notions 'stiffness matrix' and 'load vector' stem from mechanical applications. Indeed, the variational equation (2.54) describes the deflection \(u\) of a clamped membrane under the influence of a load with density \(f\). The bilinear form \(a(\cdot, \cdot)\) is related to the stiffness of the membrane.

As far as the choice of appropriate finite dimensional subspaces \(V_h \subset V\) is concerned, there are essentially two aspects:

- the accuracy of the approximation \(u_h \in V_h\) of the solution \(u \in V\) of (2.62),
- the efficient numerical solution of the linear algebraic system (2.70).

The former aspect will be taken care of by Céa’s Lemma which says that under the assumptions of the Lemma of Lax and Milgram the accuracy of the approximation \(u_h \in V_h\) corresponds to the best approximation of the solution \(u \in V\) by a function in \(V_h\). Céa’s Lemma is based on the observation that the error \(u - u_h\) is \(a\)-orthogonal to \(V_h\), i.e.,

\[
a(u - u_h, v_h) = 0, \quad v_h \in V_h,
\]

a property, which is called Galerkin orthogonality.

**Definition 2.52. (Energy norm)**

If the bilinear form \(a(\cdot, \cdot)\) is symmetric, bounded, and \(V\)-elliptic, it defines an inner product \((\cdot, \cdot)_a := a(\cdot, \cdot)\) and an associated norm \(\|\cdot\|_a := a(\cdot, \cdot)^{1/2}\) on \(V\) which is called the energy norm.

In this case, the Galerkin orthogonality (2.72) says that the solution
$u_h \in V_h$ of (2.68) is the projection of the solution $u \in V$ of (2.62) onto $V_h$. Consequently, $u_h \in V_h$ is called the elliptic projection of $u \in V$ onto $V_h$.

**Theorem 2.53. (Céa’s Lemma)**

*Under the assumptions of the Lemma of Lax and Milgram let $u \in V$ and $u_h \in V_h$ be the unique solutions of (2.62) and (2.68). Then it holds*

$$
\|u - u_h\|_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V.
$$

*Proof.* Using the boundedness (2.60) and the $V$-ellipticity (2.61) of the bilinear form $a(\cdot, \cdot)$ as well as the Galerkin orthogonality (2.72), for an arbitrary $v_h \in V_h$ we have

$$
\alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \leq C \|u - u_h\|_V \|u - v_h\|_V,
$$

which readily allows to conclude. □

The efficient numerical solution of (2.70) depends on the structure of the stiffness matrix (2.71a). In particular, we are interested in a sparse matrix with a specific sparsity pattern which can be achieved by the choice of basis functions of $V_h \subset V$ of small support. In the finite element method, such basis functions are constructed on the basis of a triangulation of the computational domain $\Omega \subset \mathbb{R}^d$.

**Definition 2.54. (Triangulation)**

Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. A triangulation $\mathcal{T}_h(\Omega)$ of $\Omega$ is a partition of $\overline{\Omega}$ into a finite number of subsets $T$ such that

1. $\overline{\Omega} = \bigcup_{T \in \mathcal{T}_h(\Omega)} T$,
2. $T = \overline{T}$, int$(T) \neq \emptyset$, $T \in \mathcal{T}_h(\Omega)$,
3. int$(T_1) \cap$ int$(T_2) = \emptyset$ for all $T_1, T_2 \in \mathcal{T}_h(\Omega)$, $T_1 \neq T_2$,
4. $\partial T$, $T \in \mathcal{T}_h(\Omega)$ is Lipschitz continuous.

The index $h$ in $\mathcal{T}_h(\Omega)$ stands for the maximal diameter $diam(T)$ of the elements $T$ of the triangulation and hence, it is a measure for the granularity of the triangulation.
We distinguish two types of elements: simplicial elements and quadrilateral elements.

**Definition 2.55. (Simplex)**

A simplex $T$ in $\mathbb{R}^d$ is the convex hull of $d + 1$ points $a_j = (a_{ij})_{i=1}^d \in \mathbb{R}^d$:

$$T = \{ x = \sum_{j=1}^{d+1} \lambda_j \ a_j \mid 0 \leq \lambda_j \leq 1, \ \sum_{j=1}^{d+1} \lambda_j = 1 \}.$$ 

A simplex in $\mathbb{R}^2$ is a triangle, whereas a simplex in $\mathbb{R}^3$ is a tetrahedron. The points $a_j$, $1 \leq j \leq d + 1$, of the simplex $T$ are called vertices. An $m$-dimensional face of a simplex $T$, $0 \leq m \leq d - 1$, is a simplex in $\mathbb{R}^m$ whose vertices are also vertices of $T$. A one-dimensional face is called an edge.

For a subset $D \subseteq \Omega$ we denote by $\mathcal{F}_h(D)$ and $\mathcal{E}_h(D)$ the sets of the $(d - 1)$-dimensional faces and one-dimensional edges of $\mathcal{T}_h(\Omega)$ in $D$.

The simplex $\tilde{T}$ with the vertices $\tilde{a}_1 = (0, \ldots, 0)^T$ and $\tilde{a}_{i+1} = e_i, 1 \leq i \leq d$, is called the simplicial reference element.

The simplex $T$ is called non-degenerate, if any point $x \in \mathbb{R}^d$ can be uniquely represented according to

$$x = \sum_{j=1}^{d+1} \lambda_j \ a_j, \ \lambda_j \in \mathbb{R}, \ \sum_{j=1}^{d+1} \lambda_j = 1.$$
The non-degeneracy of a simplex is related to the unique solvability of the linear algebraic system

\[
\begin{pmatrix}
  a_{11} & \cdots & a_{1,d+1} \\
  \vdots & \ddots & \vdots \\
  a_{d1} & \cdots & a_{d,d+1} \\
  1 & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
  \lambda_1 \\
  \vdots \\
  \lambda_d \\
  \lambda_{d+1}
\end{pmatrix}
=:
\begin{pmatrix}
  x_1 \\
  \vdots \\
  x_d \\
  1
\end{pmatrix}.
\]

(2.74)

Obviously, the simplex is non-degenerate if and only if the matrix \(A\) is regular.

**Definition 2.56. (Barycentric coordinates)**

The barycentric coordinates \(\lambda_j, 1 \leq j \leq d + 1\), of a point \(x \in \mathbb{R}^d\) with respect to the \(d + 1\) vertices \(a_j\) of a non-degenerate simplex \(T\) are the components of the unique solution of the linear algebraic system (2.74). The center of gravity \(x(S)\) of a non-degenerate simplex \(T\) is the point with

\[
\lambda_j(x_S) = \frac{1}{d+1}, \quad 1 \leq j \leq d + 1.
\]

**Lemma 2.57. (Simplicial reference element)**

Each non-degenerate simplex \(T \subset \mathbb{R}^d\) is the image of the simplicial reference element \(\hat{T}\) under an affine transformation

\[
F_T : \mathbb{R}^d \rightarrow \mathbb{R}^d
\]

\[
\hat{x} \mapsto F_T(\hat{x}) = B_T \hat{x} + b_T
\]

with a regular matrix \(B_T \in \mathbb{R}^{d \times d}\) and a vector \(b_T \in \mathbb{R}^d\).

*Proof.* The proof follows from the regularity of the matrix \(A\) in (2.74). \(\square\)

**Definition 2.58. (Simplicial triangulation)**

A triangulation \(T_h(\Omega)\) of a polyhedral domain \(\Omega \subset \mathbb{R}^d\) is called a simplicial triangulation, if all its elements are non-degenerate simplices in \(\mathbb{R}^d\).

We now consider quadrilateral elements.
Definition 2.59. (Quadrilateral elements)
A quadrilateral element $T$ in $\mathbb{R}^d$ is the tensor product of $d$ intervals $[c_i, d_i], c_i \leq d_i, 1 \leq i \leq d$, i.e.,
\[ T = \prod_{i=1}^{d} [c_i, d_i] = \{ x = (x_1, ..., x_d)^T | c_i \leq x_i \leq d_i, 1 \leq i \leq d \}. \]

A quadrilateral element in $\mathbb{R}^2$ is a rectangle. A quadrilateral element in $\mathbb{R}^3$ is a cuboid.
A quadrilateral element $T$ is called non-degenerate, if $c_i < d_i, 1 \leq i \leq d$.
The quadrilateral element $\hat{T} := [0, 1]^d$ is called the quadrilateral reference element in $\mathbb{R}^d$.
The points $a_j = (a_{ji}, \cdots, a_{j_d})^T$, $a_{ji} = c_i$ or $a_{ji} = d_i$, $1 \leq i \leq d$, of a quadrilateral element $T$ are called vertices. An $m$-dimensional face of a quadrilateral element $T, 1 \leq m \leq d - 1$, is a quadrilateral element in $\mathbb{R}^m$ whose vertices are also vertices of $T$ A one-dimensional face is called an edge.

For $D \subseteq \Omega$ we denote by $F_h(D)$ and $E_h(D)$ the sets of the $(d - 1)$-dimensional faces and one-dimensional edges of $T_h(\Omega)$ in $D$.

\[ \begin{array}{c}
\begin{array}{c}
\includegraphics[width=0.4\textwidth]{quadrilateral_2d.png}
\end{array}
\end{array} \]

\[ \begin{array}{c}
\begin{array}{c}
\includegraphics[width=0.4\textwidth]{quadrilateral_3d.png}
\end{array}
\end{array} \]

FIGURE 12. Quadrilateral element in $\mathbb{R}^2$ (left) and in $\mathbb{R}^3$ (right)

Lemma 2.60. (Quadrilateral reference element)
Any non-degenerate quadrilateral element $T \subset \mathbb{R}^d$ is the image of the quadrilateral reference element $\hat{T}$ under a diagonally affine transformation
\[ F_T : \mathbb{R}^d \rightarrow \mathbb{R}^d \]
\[ \hat{x} \mapsto F_T(\hat{x}) = B_T \hat{x} + b_T \]
with a regular diagonal matrix $B_T \in \mathbb{R}^{d \times d}$ and a vector $b_T \in \mathbb{R}^d$.

Proof. The proof is done by the explicit computation of $B_T$ and $b_T$. □

Definition 2.61. (Quadrilateral triangulation)
A triangulation $\mathcal{T}_h(\Omega)$ of a domain $\Omega \subset \mathbb{R}^d$, which is the union of finitely many quadrilateral subdomains, is called a quadrilateral triangulation, if any of its elements $T$ is a non-degenerate quadrilateral element.

Given a $\mathcal{T}_h(\Omega)$, conforming finite element functions are defined locally on elements $T \in \mathcal{T}_h(\Omega)$ and composed in such a way that the resulting global function lives in $V$.

Definition 2.62. (Local trial functions, degrees of freedom, unisolvence)
Let $\mathcal{T}_h(\Omega)$ be a triangulation of $\Omega \subset \mathbb{R}^d$ and $T \in \mathcal{T}_h(\Omega)$. Moreover, let $P_T$ be a linear space of functions $p : T \to \mathbb{R}$ with $\dim P_T = n_T$. The elements of $P_T$ are called local trial functions.

For bounded linear functionals $\ell_i : P_T \to \mathbb{R}, 1 \leq i \leq n_T$, let

$$
\Sigma_T := \{ \ell_i(p) \mid p \in P_T, \ 1 \leq i \leq n_T \}. 
$$

(2.75)

The elements of $\Sigma_T$ are called degrees of freedom.

The triple $(T, P_T, \Sigma_T)$ is called a finite element. A finite element $(T, P_T, \Sigma_T)$ is called unisolvent, if any $p \in P_T$ is uniquely determined by the degrees of freedom in $\Sigma_T$.

The numerical evaluation of the elements of the stiffness matrix (2.71a) and of the components of the load vector (2.71b) is facilitated in the case of affine equivalent finite elements.

Definition 2.63. (Affine equivalent finite elements)
Let $\mathcal{T}_h(\Omega)$ be a triangulation of $\Omega \subset \mathbb{R}^d$, $T \in \mathcal{T}_h(\Omega)$ and $P_T, \Sigma_T$ as in Definition 2.62. Further, let $(\hat{T}, \hat{P}_T, \hat{\Sigma}_T)$ be the reference element. Then, the finite element $(T, P_T, \Sigma_T)$ is called affine equivalent, if there exists an invertible affine mapping $F_T : \mathbb{R}^d \to \mathbb{R}^d$ such that for all $T \in \mathcal{T}_h(\Omega)$ it holds

(i) $T = F_T(\hat{T}),$

(ii) $P_K = \{ p : T \to \mathbb{R} \mid p = \hat{p} \circ F_T^{-1}, \ \hat{p} \in \hat{P}_T \},$

(iii) $\Sigma_T = \{ \ell_i : P_T \to \mathbb{R} \mid \ell_i = \hat{\ell}_i \circ F_T^{-1}, \ \hat{\ell}_i \in \hat{\Sigma}_T, \ 1 \leq i \leq n_T \}.$
Definition 2.64. (Finite element space)
Let \( \mathcal{T}_h(\Omega) \) be a triangulation of \( \Omega \subset \mathbb{R}^d \), \( T \in \mathcal{T}_h(\Omega) \) and \( P_T, \Sigma_T \) as in Definition 2.62. Then
\[
V_h := \{ v_h : \Omega \to \mathbb{R} \mid v_h|_T \in P_T, \ T \in \mathcal{T}_h(\Omega) \}
\]
is called a finite element space. A finite element space \( V_h \) is said to be \( H^1 \)-conforming, if \( V_h \subset H^1(\Omega) \).

The following result provides sufficient conditions for the \( H^1 \)-conformity of a finite element space \( V_h \).

Theorem 2.65. (Sufficient conditions for \( H^1 \)-conformity)
Let \( V_h \) be a finite element space and suppose that
\begin{enumerate}[(i)]
  \item \( P_T \subset H^1(T), \ T \in \mathcal{T}_h(\Omega) \),
  \item \( V_h \subset C(\Omega) \).
\end{enumerate}
Then \( V_h \) is \( H^1 \)-conforming.

Proof. Let \( v_h \in V_h \). Obviously, we have \( v_h \in L^2(\Omega) \). It remains to be shown that \( v_h \) admits weak first derivatives \( w^\alpha_h \in L^2(\Omega) \), \( |\alpha| = 1 \), i.e.,
\[
\int_\Omega v_h D^\alpha z \, dx = (-1)^{|\alpha|} \int_\Omega w^\alpha_h z \, dx, \quad z \in C_0^\infty(\Omega). \tag{2.76}
\]
Since \( v_h|_T \in H^1(T) \), we apply Green’s formula elementwise and obtain
\[
\int_\Omega v_h D^\alpha z \, dx = \sum_{T \in \mathcal{T}_h(\Omega)} \int_T v_h D^\alpha z \, dx =
- \sum_{T \in \mathcal{T}_h(\Omega)} \int_T D^\alpha v_h z \, dx + \sum_{T \in \mathcal{T}_h(\Omega)} \int_T v_h \nu_\alpha D^\alpha z \, d\sigma =
- \sum_{T \in \mathcal{T}_h(\Omega)} \int_T D^\alpha v_h z \, dx + \sum_{F \in \mathcal{F}_h(\Omega)} \int_F [v_h] \nu_\alpha D^\alpha z \, d\sigma,
\]
where \([v_h]\) denotes the jump \([v_h] := v_h|_{T_1} - v_h|_{T_2} \) across \( F = T_1 \cap T_2, T_i \in \mathcal{T}_h(\Omega), 1 \leq i \leq 2 \). Due to \( v_h \in C(\overline{\Omega}) \) we have \([v_h] = 0\) in (2.76) which implies (2.76) with \( w^\alpha_h|_T := D^\alpha v_h|_T, T \in \mathcal{T}_h(\Omega) \). \( \square \)

Corollary 2.66. (Sufficient conditions for \( H_0^1 \)-conformity)
Let \( V_h \) be a finite element space and suppose that
\begin{enumerate}[(i)]
  \item \( P_T \subset H^1(T), \ T \in \mathcal{T}_h(\Omega) \),
  \item \( V_h \subset C_0(\Omega) \).
\end{enumerate}
Then $V_h$ is $H^1_0$-conforming.

In the sequel we will deal with the construction of simplicial Lagrangean finite elements. The local trial functions are chosen as follows: Let $T$ be a simplex in $\mathbb{R}^d$. For $k \in \mathbb{N}_0$ we define $P_k(T)$ as the linear space of all polynomials of degree \( \leq k \) in $T$, i.e., $p \in P_k(T)$, if
\[
p(x) = \sum_{|\alpha| \leq k} a_\alpha \, x^\alpha, \quad a_\alpha \in \mathbb{R}, \ |\alpha| \leq k,
\]
where
\[
x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}, \quad \alpha := (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d, \quad |\alpha| := \sum_{i=1}^d \alpha_i.
\]
We note that
\[
(2.78) \quad \dim P_k(T) = \binom{k+d}{d} = \frac{(k+d)!}{d! \, k!}.
\]
If $T$ has the vertices $a_i, 1 \leq i \leq d + 1$, then the set
\[
L_k(T) := \{ x = \sum_{i=1}^{d+1} \lambda_i \, a_i \mid \lambda_i \in \{0, \frac{1}{k}, \ldots, \frac{k-1}{k}, 1\}, \ \sum_{i=1}^{d+1} \lambda_i = 1 \}
\]
is called the principal lattice of order $k$ of $T$. We have
\[
(2.79) \quad \text{card}(L_k(T)) = \binom{k+d}{d}.
\]

![Figure 13. Nodal points • of a simplicial Lagrangean finite element of type (1) (left) and of type (2) (right)](image)

**Definition 2.67. (Nodal points)**
The elements of the principal lattice of order $k$ of $T$ are called nodal points. For $D \subseteq \mathbb{T}$ we denote by $N_h(D)$ the set of nodal points in $D$. 
If we define the set $\Sigma_T$ of degrees of freedom as the values of $p \in P_k(T)$ in the nodal points, we obtain a simplicial Lagrangean finite element.

**Definition 2.68. (Simplicial Lagrangean finite element)**
Let $T_h(\Omega)$ be a triangulation of $\Omega \subset \mathbb{R}^d$. Then the triple $(T, P_k(T), L_k(T))$, $T \in T_h(\Omega)$, is called a simplicial Lagrangean finite element of type $k$.

**Lemma 2.69. (Unisolvence of simplicial Lagrangean finite elements)**
A simplicial Lagrangean finite element of type $k$ is unisolvent and affine equivalent to the reference element $(\hat{T}, \hat{P}_T, \hat{\Sigma}_T)$ with $\hat{P}_T = P_k(\hat{T})$ und $\hat{\Sigma}_T = \{ \hat{p}(\hat{x}) \mid \hat{x} \in L_k(\hat{T}), \hat{p} \in P_k(\hat{T}) \}$.

**Proof.** The unisolvence is left as an exercise. For the proof of the affine equivalence let $F_T : \mathbb{R}^d \to \mathbb{R}^d$ be the invertible affine mapping with $T = F_T(\hat{T})$. Then we have $P_k(T) = \{ p = \hat{p} \circ F_T^{-1} \mid \hat{p} \in P_k(\hat{T}) \}$. Since $x \in L_k(T) \iff x = F_T(\hat{x}), \hat{x} \in L_k(\hat{T})$, it follows that (iii) from Definition 2.63 is satisfied. □

**Definition 2.70. (Simplicial Lagrangean finite element space)**
The finite element space $V_h$ composed by simplicial Lagrangean finite elements of type $k$ is called simplicial Lagrangean finite element space. It will be denoted by $S_k(\Omega, T_h(\Omega))$. Moreover, we define $S_{k,0}(\Omega, T_h(\Omega)) := \{ v_h \in S_k(\Omega, T_h(\Omega)) \mid v_h|_\Gamma = 0 \}$.

In order to ensure $H^1$-conformity of $S_k(\Omega, T_h(\Omega))$ we have to assume that the simplicial triangulation is geometrically conforming.

**Definition 2.71. (Geometrically conforming simplicial triangulation)**
A simplicial triangulation $T_h(\Omega)$ is said to be geometrically conforming, if the intersection of two different elements of the triangulation is either empty or consists of a common face, a common edge, or a common vertex.

**Lemma 2.72. (Geometrical conformity)**
Let $T_h(\Omega)$ be a geometrically conforming simplicial triangulation. Then it holds
$$S_k(\Omega, T_h(\Omega)) \subset H^1(\Omega), \quad S_{k,0}(\Omega, T_h(\Omega)) \subset H^1_0(\Omega).$$
Proof. Taking \( P_k(T) \subset H^1(T) \), \( T \in \mathcal{T}_h(\Omega) \), and Theorem 2.65 into account, we only have to show that \( S_k(\Omega, \mathcal{T}_h(\Omega)) \subset C(\overline{\Omega}) \). We provide the proof exemplarily for \( d = 2 \) and \( k = 2 \): Let \( T_i \in \mathcal{T}_h(\Omega) \), \( 1 \leq i \leq 2 \), be two neighboring elements such that \( E = T_1 \cap T_2 \in \mathcal{E}_h(\Omega) \) with nodal points \( d_j \in N_h(E) \), \( 1 \leq j \leq 3 \). Further, let \( p_i \in P_2(K_i) \), \( 1 \leq i \leq 2 \). Then \( p_i|_E \in P_2(E) \). Since \( p_1(d_j) = p_2(d_j) \), \( 1 \leq j \leq 3 \), we obtain \( p_1|_E \equiv p_2|_E \). □

It remains to specify the global trial functions which are called nodal basis functions:

**Definition 2.73. (Nodal basis functions I)**

Let \( N_h(\Omega) = \{ x_j \mid 1 \leq j \leq n_h \} \). The function \( \varphi^{(i)}_h \in S_k(\Omega, \mathcal{T}_h(\Omega)) \) given by

\[
\varphi^{(i)}_h(x_j) = \delta_{ij}, \quad 1 \leq i, j \leq n_h,
\]

is called nodal basis function associated with the nodal point \( x_i \). The set of these functions is a basis of the simplicial finite element space \( S_k(\Omega, \mathcal{T}_h(\Omega)) \).

We now consider the construction of quadrilateral Lagrangean finite elements. Let \( T := \prod_{i=1}^d [c_i, d_i] \subset \mathbb{R}^d \) be a quadrilateral element. For \( k \in \mathbb{N}_0 \) we denote by \( Q_k(T) \) the linear space of all polynomials of degree \( \leq k \) in each of the \( d \) variables \( x_i \), \( 1 \leq i \leq d \), i.e., \( p \in Q_k(T) \) is of the form

\[
p(x) = \sum_{\alpha_i \leq k} a_{\alpha_1 \cdots \alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}.
\]

It follows that

\[
\dim Q_k(T) = (k + 1)^d.
\]

**Definition 2.74. (Quadrilateral Lagrangean finite element)**

Let \( L_{[k]}(T) \) be the set

\[
L_{[k]}(T) := \{ x = (x_{i_1}, \ldots, x_{i_d})^T \mid x_{i_\ell} := c_\ell + \frac{i_\ell}{k}(d_\ell - c_\ell), \ i_\ell \in \{0, 1, \ldots, k\}, \ 1 \leq \ell \leq d \}
\]

and let \( \Sigma(T) \) be given by

\[
\Sigma_T := \{ p(x) \mid x \in L_{[k]}(T) \}, \ p \in Q_k(T).
\]

The triple \((T, Q_k(T), \Sigma_T)\) is called a Quadrilateral Lagrangean finite element of type \([k]\). It will be denoted by \( S_{[k]}(K) \). The points \( x \in L_{[k]}(T) \) are called nodal points.
Lemma 2.75. (Quadrilateral Lagrangean finite elements)
The quadrilateral Lagrangean finite element of type \([k]\) is based on a
tensor product like polynomial interpolation of Lagrangean type. The
polynomial which interpolate in \(x \in L_{[k]}(T)\) have the representation
\[
p(x) = \sum_{y \in L_{[k]}(T)} L(x) p(y).
\]
Here, \(L \in Q_k(T)\) is given by
\[
L(x) = \prod_{\ell=1}^d L_{i_\ell}(x),
\]
where \(L_{i_\ell}, 1 \leq i \leq d,\) are the one-dimensional Lagrangean fundamental
polynomials.

Proof. Let \(p \in Q_k(T)\) and let \(e \subset T\) be an edge of \(T\) given by
\[
e := [c_j, d_j] \times \prod_{j \neq \ell=1}^d \{g_\ell\}, \quad g_\ell \in \{c_\ell, d_\ell\}.
\]
Then \(p|_e \in P_k(e)\) is uniquely determined by its values in \(x_i := c_j +
\frac{1}{k}(d_j - c_j) \in e, i \in \{0, 1, \ldots, k\},\) and has the Lagrangean representation
\[
p|_e(x) = \sum_{i=0}^k L_i(x) p|_e(x_i), \quad x \in e.
\]
\[\square\]

Lemma 2.76. (Unisolvence and affine equivalence of quadrilateral Lagrangean finite elements)
The quadrilateral Lagrangean finite element of type \([k]\) is unisolvent. The
quadrilateral Lagrangean finite elements of type \([k]\) are affine
equivalent to the reference element \((\hat{T}, \hat{P}_T, \hat{\Sigma}_T)\) with \(\hat{P}_T = Q_k(\hat{T})\) and
\(\hat{\Sigma}_T = \{\hat{p}(\hat{x}) \mid \hat{x} \in L_{[k]}(\hat{T}), \hat{p} \in \hat{Q}_T\}..\)

Proof. The proof of the unisolvence follows easily from Lemma 2.76. The
proof of the affine equivalence is left as an exercise. \[\square\]

Definition 2.77. (Geometrically conforming quadrilateral triangulation)
A quadrilateral triangulation \(\mathcal{T}_h(\Omega)\) is said to be geometrically con-
forming, if the intersection of two different elements of the triangula-
tion either is empty or consists of a common face, a common edge, or
a common vertex.
Definition 2.78. (Quadrilateral Lagrangean finite element space)
Let $\Omega \subset \mathbb{R}^d$ be the union of finitely many quadrilateral subdomains in $\mathbb{R}^d$ and let $\mathcal{T}_h(\Omega)$ be a geometrically conforming quadrilateral triangulation. The finite element space $V_h$ consisting of quadrilateral Lagrangean finite elements of type $k$ is called a quadrilateral Lagrangean finite element space. It will be denoted by $S_{[k]}(\Omega, \mathcal{T}_h(\Omega))$. We further define
\[ S_{[k],0}(\Omega, \mathcal{T}_h(\Omega)) := \{ v_h \in S_{[k]}(\Omega, \mathcal{T}_h(\Omega)) \mid v_h|_{\Gamma} = 0 \}. \]

Lemma 2.79. (H$^1$-conformity of quadrilateral Lagrangean finite element spaces)
Let $\mathcal{T}_h(\Omega)$ be a geometrically conforming quadrilateral triangulation of $\Omega$. Then it holds
\[ S_{[k]}(\Omega, \mathcal{T}_h(\Omega)) \subset H^1(\Omega), \]
\[ S_{[k],0}(\Omega, \mathcal{T}_h(\Omega)) \subset H^1_0(\Omega). \]

Definition 2.80. (Nodal basis functions II)
Let $\mathcal{N}_h(\Omega) = \{ x \in L_{[k]}(T) \mid T \in \mathcal{T}_h(\Omega) \}$ with $\text{card}(\mathcal{N}_h(\Omega)) = n_h$. The function $\varphi_h^{(i)} \in S_{[k]}(\Omega, \mathcal{T}_h(\Omega))$ given by
\[ \varphi_h^{(i)}(x_j) = \delta_{ij}, \quad 1 \leq i, j \leq n_h, \]
is called the nodal basis function associated with the nodal point $x_i \in \mathcal{N}_h(\Omega)$. The set of these basis functions is a basis of $S_{[k]}(\Omega, \mathcal{T}_h(\Omega))$.

The accuracy of the finite element approximation $u_h \in V_h$ of the solution $u \in V$ of (2.62) will be measured by the global discretization error $u - u_h$ on the basis of interpolation in Sobolev spaces.
Definition 2.81. (Local and global interpolation operator)
Let $\mathcal{T}_h(\Omega)$ be a geometrically conforming simplicial triangulation of $\Omega \subset \mathbb{R}^d$ and let $(T, P_T, \Sigma_T), T \in \mathcal{T}_h(\Omega)$, be a simplicial Lagrangean finite element of type $k$, where $P_T = P_k(T)$ with $\dim P_T = n_k$ and $\Sigma_T = \{p(x_i) \mid x_i \in L_k(T), 1 \leq i \leq n_k\}$. Further, let $\varphi_T^{(i)}, 1 \leq i \leq n_k$, be the nodal basis functions associated with the nodal points $x_i$. The operator $I_T : \mathcal{V} \cap C(T) \rightarrow P_T$ given by

$$I_T v := \sum_{i=1}^{n_k} v(x_i) \varphi_T^{(i)}, \quad v \in \mathcal{V} \cap C(T)$$

(2.84)

is called the local interpolation operator. If $V_h$ is the associated simplicial Lagrangean finite element space, the operator $I_h : \mathcal{V} \cap C(\Omega) \rightarrow V_h$ given by

$$I_h v|_T := I_T v|_T, \quad T \in \mathcal{T}_h(\Omega)$$

(2.85)

is called the global interpolation operator.

We now assume that the solution $u \in \mathcal{V}$ of (2.62) satisfies $u \in \mathcal{V} \cap H^{k+1}(\Omega)$ with $k \in \mathbb{N}$ sufficiently large such that $H^{k+1}(\Omega) \subset C(\Omega)$. Then $I_h u$ is well defined and Céa’s Lemma (cf. Lemma 2.81) implies

$$\|u - u_h\|_{1,\Omega} \leq \frac{C}{\alpha} \|u - I_h u\|_{1,\Omega}.$$  

(2.86)

In order to estimate the global interpolation error $\|u - I_h u\|_{1,\Omega}$, taking advantage of the definition of $I_h$ and of $P_T = P_k(T) \subset H^1(T), T \in \mathcal{T}_h(\Omega)$, we obtain

$$\|u - I_h u\|_{1,\Omega} = \left( \sum_{T \in \mathcal{T}_h(\Omega)} \|u - I_T u\|^2_{1,T} \right)^{1/2},$$

(2.87)

i.e., we obtain an estimate of the global interpolation error by means of estimates of the local interpolation errors $\|u - I_T u\|_{1,T}, T \in \mathcal{T}_h(\Omega)$.

Lemma 2.82. (Estimate of the local interpolation operator)
Let $(\hat{T}, P_k(\hat{T}), \hat{\Sigma}_{\hat{T}})$ be the simplicial Lagrangean reference element of type $(k)$ with $k \in \mathbb{N}$ such that $H^{k+1}(\hat{T})$ is continuously embedded in $C(\hat{T})$. Further, let $(T, P_k(T), \Sigma_T)$ be a simplicial Lagrangean finite element of type $(k)$ which is affine equivalent to $(\hat{T}, P_k(\hat{T}), \hat{\Sigma}_{\hat{T}})$ with $\rho(T)$ denoting the radius of the largest ball that can be inscribed in $T$. Then there exists a constant $C > 0$ which only depends on the reference element such that

$$|v - I_T v|_{1,T} \leq C \frac{h_{k+1}^T}{\rho_T} |v|_{k+1,T}, \quad v \in H^{k+1}(T).$$  

(2.88)
Proof. We refer to Ciarlet. □

Now, let $\mathcal{H}$ be a null sequence of positive real numbers and let $\{\mathcal{T}_h(\Omega)\}_{\mathcal{H}}$ be a family of geometrically conforming simplicial triangulations of $\Omega$. The estimate (2.88) shows that the estimate (2.87) of the global interpolation error depends in such a way on the triangulations that the right-hand side in (2.87) is unbounded, if $h_T/\rho_T \to \infty$, $T \in \mathcal{T}_h(\Omega)$ as $h \to 0$. In order to exclude this case, we have to impose a further condition on $\{\mathcal{T}_h(\Omega)\}_{\mathcal{H}}$.

**Definition 2.83. (Shape regularity)**
The family $\{\mathcal{T}_h(\Omega)\}_{\mathcal{H}}$ of geometrically conforming simplicial triangulations is called shape regular, if there exists a constant $\sigma > 0$, independent of $h_T$, $T \in \mathcal{T}_h(\Omega)$, such that for all $T \in \mathcal{T}_h(\Omega), h \in \mathcal{H}$ it holds

$$ \frac{h_T}{\rho_T} \leq \sigma. \quad (2.89) $$

We are now in a position to provide an a priori error estimate of the global discretization error in the $\| \cdot \|_{1,\Omega}$-norm.

**Theorem 2.84. (A priori error estimate in the $H^1$-norm I)**
Let $\{\mathcal{T}_h(\Omega)\}_{\mathcal{H}}$ be a shape regular family of geometrically conforming simplicial triangulations of $\Omega$ and let $\{V_h\}_{\mathcal{H}}$ be the associated family of finite element spaces based on simplicial Lagrangean finite elements of type $k$ with $k$ such that the embeddings $H^{k+1}(\Omega) \to C(\overline{\Omega})$ and $H^{k+1}(\hat{T}) \to C(\hat{T})$ are continuous. If $u \in V$ is the solution of (2.62) with $u \in V \cap H^{k+1}(\Omega)$ and if $u_h \in V_h, h \in \mathcal{H}$, are the solutions of (2.68), then there exists a constant $C > 0$, depending only on the local geometry of the triangulations, such that

$$ \|u - u_h\|_{1,\Omega} \leq C h^k |u|_{k+1,\Omega}. \quad (2.90) $$

Proof. The proof follows readily from (2.86),(2.87),(2.88), and (2.89). □

For so-called $(k + 1)$-regular variational equations we obtain upper bounds that only depend on $h$ and on the data of the problem.

**Definition 2.85. (k-regularity)**
The variational equation (2.62) with

$$ \ell(v) = (f,v)_{0,\Omega}, \quad f \in L^2(\Omega) $$

is called \( k \)-regular, \( k \geq 2 \), if its solution satisfies \( u \in V \cap H^{k+1}(\Omega) \) and if there exists a constant \( C > 0 \), independent of \( h \), such that
\[
\|u\|_{k,\Omega} \leq C \|f\|_{0,\Omega}.
\]

(2.91)

**Corollary 2.86. (A priori error estimate in the \( H^1 \)-norm II)**

Under the assumptions of Theorem 2.84 suppose that the solution \( u \) of (2.62) is \((k + 1)\)-regular. Then, there exists a constant \( C > 0 \), depending only on the local geometry of the triangulation, such that
\[
\|u - u_h\|_{1,\Omega} \leq C h^k \|f\|_{0,\Omega}.
\]

(2.92)

Since \( \|v\|_{0,\Omega} \leq \|v\|_{1,\Omega} \), \( v \in V \), the estimate (2.92) implies \( \|u - u_h\|_{0,\Omega} = O(h^k) \). However, the estimation \( \|I_h u - u\|_{0,\Omega} = O(h^{k+1}) \) of the interpolation error suggests that this estimate is not optimal. In fact, under certain assumptions the optimal order \( O(h^{k+1}) \) can be shown. The theoretical background of an optimal a priori error estimate in the \( L^2 \)-norm is given by the Lemma of Aubin and Nitsche, also known as Nitsche’s trick.

**Theorem 2.87. (Lemma of Aubin and Nitsche)**

Let \( V \) be a Hilbert space with inner product \((\cdot,\cdot)_{V}\), let \( a(\cdot,\cdot) : V \times V \to \mathbb{R} \) be a bounded, \( V \)-elliptic bilinear form, and let \( \ell : V \to \mathbb{R} \) be a bounded linear functional. Let \( H \) be another Hilbert space with inner product \((\cdot,\cdot)_H\) such that \( V \) is continuously and densely embedded in \( H \). Further, let \( u \in V \) and \( u_h \in V_h \) be the unique solutions of (2.62) and (2.68). Then it holds

(i) For each \( g \in H \) the adjoint variational equation
\[
a(v, z_g) = (g, v)_H, \quad v \in V
\]

admits a unique solution \( z_g \in V \).

(ii) There exists a constant \( C > 0 \) such that the global discretization error \( u - u_h \) satisfies
\[
\|u - u_h\|_H \leq C \|u - u_h\|_V \left( \sup_{g \in H} \frac{1}{\|g\|_{H}} \inf_{\varphi_h \in V_h} \|z_g - \varphi_h\|_V \right).
\]

(2.94)

**Proof.** We refer to Ciarlet. \( \square \)

The following a priori estimate of the global discretization error in the \( L^2 \)-norm is an immediate consequence of the Lemma of Aubin and Nitsche.
Theorem 2.88. (A priori error estimate in the $L^2$-norm)
Let the assumptions of Theorem 2.90 be satisfied and suppose that the adjoint variational equation (2.93) is $(k+1)$-regular. Then there exists a constant $C > 0$, depending only on the local geometry of the triangulations, such that

$$
\|u - u_h\|_{0,\Omega} \leq C h^{k+1} \|u\|_{k+1,\Omega}.
$$

(2.95)

Proof. The proof follows from the Lemma of Aubin and Nitsche with $H = L^2(\Omega)$. In particular, it follows from Corollary 2.86 and the $(k + 1)$-regularity of the adjoint variational equation that

$$
\inf_{\varphi_h \in V_h} \|z_g - \varphi_h\|_{1,\Omega} \leq \|z_g - I_h z_g\|_{1,\Omega} \leq C h \|z_g\|_{k+1,\Omega} \leq C h \|g\|_{0,\Omega}.
$$

Hence, from (2.94) we deduce

$$
\|u - u_h\|_{0,\Omega} \leq C h \|u - u_h\|_{1,\Omega}.
$$

Theorem 2.84 provides a further estimate of the right-hand side which allows to conclude. □

Remark 2.89. (A priori error estimate in the $L^\infty$-norm)
For an optimal a priori error estimate of the global discretization error in the $L^\infty$-norm we refer to Ciarlet.
2.3 The heat equation

2.3.1 Preliminaries

We consider an initial-boundary value problem for the heat equation (2.6) in \( Q := \Omega \times (0, T) \), where \( \Omega \) is a bounded domain in \( \mathbb{R}^d \) with boundary \( \Gamma := \partial \Omega \), and \( T \in \mathbb{R}_+ \). We set \( \Sigma := \Gamma \times (0, T) \). We further assume given functions

\[
\begin{align*}
(2.96) & \quad f \in C(Q), \quad u^D \in C(\Gamma \times (0, T)), \quad u^0 \in C(\overline{\Omega}), \\
(2.97) & \quad u^D(x, 0) = u^0(x), \quad x \in \Gamma.
\end{align*}
\]

which satisfy the compatibility condition

\[
(2.97)
\]

The initial-boundary value problem amounts to the computation of a function \( u : Q \rightarrow \mathbb{R} \) such that

\[
\begin{align*}
(2.98a) & \quad u_t(x, t) - \Delta u(x, t) = f(x, t), \quad (x, t) \in Q, \\
(2.98b) & \quad u(x, t) = u^D(x, t), \quad (x, t) \in \Sigma, \\
(2.98c) & \quad u(x, 0) = u^0(x), \quad x \in \Omega.
\end{align*}
\]

Definition 2.90. (Classical solution of the initial-boundary value problem for the heat equation)

Under the assumptions (2.96) and (2.97), a function \( u \in C(\overline{Q}) \) with \( u(x, \cdot) \in C^1((0, T)) \) for \( x \in \Omega \), and \( u(\cdot, t) \in C^2(\Omega) \) for \( t \in (0, T) \), is called a classical solution of (2.98a)-(2.98c), if \( u \) pointwise satisfies the equations (2.98a)-(2.98c). The initial-boundary value problem is said to be well posed, if the solution depends continuously on the data \( f, u^D, u^0 \).

In the sequel we summarize the most important existence and uniqueness results for classical solutions of the initial-boundary value problem (2.98a)-(2.98c). For proofs and further results we refer to Evans.

The uniqueness of a classical solution follows from the weak maximum principle.

Theorem 2.91. (Weak maximum principle)

We suppose that (2.96) and (2.97) hold true. Further, we refer to \( \partial_R Q \) as the parabolic boundary \( \partial_R Q := (\overline{\Omega} \times \{0\}) \cup (\Gamma \times [0, T]) \). If the function \( u \in C(\overline{Q}) \) with \( u(x, \cdot) \in C^1((0, T)) \) for \( x \in \Omega \) and \( u(\cdot, t) \in C^2(\Omega) \) for \( t \in (0, T) \) satisfies the inequality

\[
(2.99) \quad \frac{\partial u}{\partial t}(x, t) - \Delta u(x, t) \leq 0, \quad (x, t) \in Q;
\]
then it holds
\[
\max_{(x,t) \in Q} u(x,t) = \max_{(x,t) \in \partial R} u(x,t).
\]

If \(u_1\) and \(u_2\) are two classical solutions, an application of Theorem 2.91 to \(u_1 - u_2\) and \(u_2 - u_1\) yields \(u_1 = u_2\), i.e., it holds:

**Corollary 2.92. (Uniqueness of a classical solution)**
Under the assumptions (2.96) and (2.97), the initial-boundary value problem (2.98a)-(2.98c) admits a unique classical solution.

**Remark 2.93. (Growth condition)**
In case \(\Omega = \mathbb{R}^d\) the boundary condition (2.98b) has to be replaced by the growth condition
\[
(2.101) \quad u(x,t) = O(\exp(-\lambda|x|^2)) \quad \text{for} \quad |x| \to \infty \quad (\lambda > 0)
\]
The existence of a classical solution follows from an explicit representation with respect to the data \(f, u_0, u_0\) by means of a kernel function of the heat equation.

**Definition 2.94. (Kernel function of the heat equation)**
Let \(\Omega \subseteq \mathbb{R}^d\). A function \(K : \Omega \times \Omega \times \mathbb{R}_+ \to \mathbb{R}\) is called a kernel function of the heat equation, if it holds
\[
(i) \quad \frac{\partial}{\partial t}K(x,y,t) - \Delta_x K(x,y,t) = 0, \quad (x,y,t) \in \Omega \times \Omega \times \mathbb{R}_+,
\]
\[
(ii) \quad K(x,y,t) = 0, \quad (x,y,t) \in \Gamma \times \Omega \times \mathbb{R}_+,
\]
\[
(iii) \quad \lim_{t \to 0} \int_\Omega K(x,y,t)v(x) \, dx = v(y), \quad y \in \Omega, \quad v \in C(\Omega).
\]
The existence of a kernel function can be shown by means of a fundamental solution of the heat equation
\[
(2.102) \quad \Lambda(x,y,t) := (4\pi t)^{-d/2} \exp\left(-\frac{|x-y|^2}{4t}\right), \quad (x,y,t) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}_+.
\]
For a bounded function \(u_0 \in C(\mathbb{R}^d)\) let \(u : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}\) be the function defined by the convolution
\[
u(x,t) = \Lambda(x,\cdot,t) * u_0(\cdot) = \int_{\mathbb{R}^d} \Lambda(x,y,t)u_0(y) \, dy.
\]
Then $u \in C^\infty(\mathbb{R}^d \times \mathbb{R}_+)$ and $u$ solves the heat equation in $\mathbb{R}^d \times \mathbb{R}_+$ with the initial condition $u(x,0) = u^0(x), x \in \Omega$. In fact, computing the partial derivatives $\partial \Lambda / \partial t$ and $\partial^2 \Lambda / \partial x_i \partial x_j, 1 \leq i, j \leq d$, one easily verifies that $\Lambda$ (as a function of $x$) satisfies the heat equation in $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$. It follows that

$$\frac{\partial}{\partial t} u(x,t) = \left( \frac{\partial}{\partial t} \Lambda(x, \cdot, t) \right) * u^0(\cdot) = (\Delta_x \Lambda(x, \cdot, t)) * u^0(\cdot) = \Delta_x u(x, t).$$

Since $\Lambda \in C^\infty(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}_+)\), differentiation shows that $u \in C^\infty(\mathbb{R}^d \times \mathbb{R}_+)$. In case of a bounded $C^2$-domain $\Omega \subset \mathbb{R}^d$, the fundamental solution provides an explicit representation of the solution of the initial-boundary value problem (2.98a)-(2.98c) with homogeneous initial conditions $u^0 \equiv 0$. This allows to deduce the existence of a kernel function.

**Theorem 2.95. (Kernel function for a bounded spatial domain)**

Let $\Omega \subset \mathbb{R}^d$ be a bounded $C^2$-domain. Then the heat equation admits a kernel function $K \in C^\infty(\Omega \times \Omega \times \mathbb{R}_+)$ such that $K(x,y,t) > 0$ and $K(x,y,t) = K(y,x,t), (x,y,t) \in \Omega \times \Omega \times \mathbb{R}_+$.

The kernel function $K$ from Theorem 2.95 gives rise to the following representation of the unique classical solution of the initial-boundary value problem (2.98a)-(2.98c):

**Theorem 2.96. (Classical solution of the initial-boundary value problem for the heat equation)**

Let $\Omega \subset \mathbb{R}^d$ be a bounded $C^2$-domain and suppose that the conditions (2.96) and (2.97) are satisfied. Then the initial-boundary value problem (2.98a)-(2.98c) admits a unique classical solution $u \in C(\Omega) \cap C^\infty(Q)$ which has the explicit representation

$$u(x,t) = \int_0^t \int_\Omega K(x,y,t-s)f(y,s) \, dy \, ds + \int_\Omega K(x,y,t)u^0(y) \, dy - \int_0^t \int_{\partial \Omega} \nu_{\Gamma} \cdot \nabla_y K(x,y,t-s)u^D(y,s) \, d\sigma(y) \, ds.$$

**Remark 2.97. (Compatibility condition)**

If the compatibility condition (2.97) is not satisfied, one can show the existence of a unique classical solution $u \in C((\Omega \times \mathbb{R}_+) \setminus (\Gamma \times \{0\})) \cap C^\infty(Q)$ with the same representation as in Theorem 2.96.
Theorem 2.96 shows the well posedness of the initial-boundary value problem (2.98a)-(2.98c) and reflects the smoothing property of the parabolic differential operator. The solution satisfies $u \in C^\infty(Q)$ in $Q$, even if the initial and boundary data are only continuous functions.

**Remark 2.98. (General initial-boundary value problems)**

As far as more general initial-boundary value problems for linear second order parabolic differential equations are concerned, we refer to Evans.
2.3.2 Method of lines and Rothe’s method

The numerical solution of the initial-boundary value problem (2.98a)-(2.98c) can be done on the basis of a semi-discretization in space or in time. The semi-discretization in space is called the method of lines. We consider an equidistant grid point set

$$\Omega_h := \{ x = (x_1, \ldots, x_d)^T \}$$

of step size $h > 0$ and refer to $\Omega_h$ and $\Gamma_h$ as the sets of interior and boundary grid points. Assume $f_h(t) \in C(\Omega_h)$, $t \in (0, T)$, $u^D_h \in C(\Gamma_h)$, and $u^0_h \in C(\Omega_h)$ to be approximations of $f$, $u_D$, and $u_0$. Then the method of lines requires the solution of the following initial value problem for a system of first order ordinary differential equations

$$\begin{align*}
(u_h)_t - \Delta_h u_h &= f_h(t), \quad t \in (0, T), \\
u_h(0) &= u^0_h
\end{align*}$$

in the unknowns $u_h(x), x \in \Omega_h$. Typically, this system is stiff so that the numerical integration of (2.103a),(2.103b) requires the use of implicit or semi-implicit methods (cf. Chapter 1).

The semi-discretization in time is called Rothe’s method. We consider an equidistant partition

$$\mathcal{T}_k := \{ t_m := mk, \quad k := T/(M + 1), \quad M \in \mathbb{N} \}$$

of the time interval $[0, T]$ of step size $k > 0$ and discretize (2.98a) implicitly in time, e.g., by means of the implicit Euler method. Then, for each time step we have to solve a boundary value problem for a linear second order elliptic differential equation

$$\begin{align*}
(I - k\Delta u(t_m)) &= u(t_{m-1}) + kf(t_m) \quad \text{in } \Omega, \\
u(\cdot, t_m) &= u^D(t_m) \quad \text{on } \Gamma
\end{align*}$$

For the numerical solution of (2.105a),(2.105b) we can use the techniques presented in subsections 2.2.2 and 2.2.3.
2.3.3 Finite difference approximations

We consider finite difference approximations in space and time with respect to a grid point set
\[ Q_{h,k} := \Omega_h \times \bar{T}_k. \]

For \( D_{h,k} \subseteq Q_{h,k} \) we refer to \( C(D_{h,k}) \) as the linear space of grid functions in \( D_{h,k} \).

For grid functions \( u_{h,k} \in C(\Omega_{h,k}) \) we define the forward difference quotient
\[ D^+_k u_{h,k}(x,t) := k^{-1} \left( u_{h,k}(x,t+k) - u_{h,k}(x,t) \right), \]
where \( x \in \Omega_h, t \in \overline{T}_k \setminus \{T\} \), and the backward difference quotient
\[ D^-_k u_{h,k}(x,t) := k^{-1} \left( u_{h,k}(x,t) - u_{h,k}(x,t-k) \right), \]
where \( x \in \Omega_h, t \in \overline{T}_k \setminus \{0\} \). Moreover, let \( f_{h,k} \in C(\Omega_{h,k}) \), \( u^D \in C(\Gamma_h \times \overline{T}_k) \), and \( u^0_h \in C(\Omega_h) \) be approximations of \( f, u_d \), and \( u_0 \).

**Definition 2.99. (Explicit Euler method)**
Using the forward difference quotient, we obtain the explicit Euler method
\[ u_{h,k}(x,t+k) = (I + k\Delta_h)u_{h,k}(x,t) + kf_{h,k}(x,t), \]
x \( \in \Omega_h, t \in \overline{T}_k \setminus \{T\} \),
\[ u_{h,k}(x,t+k) = u^D(x,t+k), \quad x \in \Gamma_h, t \in \overline{T}_k \setminus \{T\} \),
\[ u_{h,k}(x,0) = u^0_h(x), \quad x \in \Omega_h. \]

**Definition 2.100. (Implicit Euler method)**
The use of the backward difference quotient gives rise to the implicit Euler method
\[ (I - k\Delta_h)u_{h,k}(x,t) = u_{h,k}(x,t-k) + kf_{h,k}(x,t), \]
x \( \in \Omega_h, t \in \overline{T}_k \setminus \{0\} \),
\[ u_{h,k}(x,t) = u^D(x,t), \quad x \in \Gamma_h, t \in \overline{T}_k \setminus \{0\} \),
\[ u_{h,k}(x,0) = u^0_h(x), \quad x \in \Omega_h. \]

If \( u_{h,k}(x,t) \) is known in \( x \in \Omega_h, t = t_m, 0 \leq m \leq M \), for the explicit Euler method (2.109a)-(2.109c) we obtain \( u_{h,k}(x,t) \) in \( x \in \Omega_h, t = t_{m+1} \), by the evaluation of the right-hand side in (2.109a). However, knowing \( u_{h,k}(x,t) \) in \( x \in \Omega_h, t = t_{m-1}, 1 \leq m \leq M \), for the implicit
Euler method (2.110a)-(2.110c) the computation of \( u_{h,k}(x, t_m), x \in \Omega_h \), requires the solution of a linear algebraic system.

**Definition 2.101. (Θ-Method)**

We multiply (2.110a) (with \( t + k \) instead of \( t \)) with some \( \Theta \in [0, 1] \), multiply (2.109a) with \( 1 - \Theta \), and add the resulting equations. Denoting by \( L_{h,k}^{\Theta} \) the difference operator

\[
L_{h,k}^{\Theta} u_{h,k} := \Theta \Delta_h u_{h,k}(\cdot, t + k) + (1 - \Theta) \Delta_h u_{h,k}(\cdot, t)
\]

and by \( f_{h,k}^{\Theta} \) the grid function

\[
f_{h,k}^{\Theta} := \Theta f_{h,k}(\cdot, t + k) + (1 - \Theta) f_{h,k}(\cdot, t),
\]

we obtain the so-called Θ-method:

\[
\begin{align*}
(2.111a) \quad & \quad k^{-1}(u_{h,k}(\cdot, t + k) - u_{h,k}(\cdot, t)) - \\
& = L_{h,k}^{\Theta} u_{h,k}, \ x \in \Omega_h, \ t \in I_k \setminus \{T\}, \\
(2.111b) \quad & \quad u_{h,k}(x, t) = u_{h}^D(x, t), \ x \in \Gamma_h, \ t \in I_k \setminus \{0\}, \\
(2.111c) \quad & \quad u_{h,k}(x, 0) = u_{h}^0(x), \ x \in \Omega_h.
\end{align*}
\]

For \( \Theta = 0 \) we recover the explicit Euler method (2.109a)-(2.109c), whereas \( \Theta = 1 \) yields the implicit Euler method (2.110a)-(2.110c). The implicit scheme for \( \Theta = 1/2 \) is called the Crank-Nicolson method.

For the definition of the convergence of the finite difference approximations we measure the global discretization error \( e_{h,k} := u|\Omega_{h,k} - u_{h,k} \) in a suitable norm \( \| \cdot \|_{h,k} \) of the linear space \( C(\Omega_{h,k}) \) of grid functions in \( \Omega_{h,k} \). An example is the discrete maximum norm

\[
\|u_{h,k}\|_{h,k} := \max_{(x,t) \in \Omega_{h,k}} |u_{h,k}(x,t)|.
\]

**Definition 2.102. (Convergence of the Θ-method)**

Let \( u \) be the classical solution of the initial-boundary value problem for the heat equation and let \( u_{h,k} \in C(\Omega_{h,k}) \) be the solution of the Θ-method. The Θ-method is said to be convergent, if

\[
\|u - u_{h,k}\|_{h,k} \to 0 \quad \text{as} \ h, k \to 0.
\]

The Θ-method is said to be convergent of order \( p_1 \) in \( h \) and \( p_2 \) in \( k \), if there exists a constant \( C > 0 \), independent of \( h \) and \( k \), such that for sufficiently small \( h, k \) it holds

\[
\|u - u_{h,k}\|_{h,k} \leq C (h^{p_1} + k^{p_2}).
\]
As in the numerical solution of initial value problems for ordinary differential equations (cf. Chapter 1), sufficient conditions for the convergence are given by the consistency of the Θ-method with the given initial-boundary value problem and by the stability of the Θ-method.

**Definition 2.103. Consistency of the Θ-method**

Under the assumptions of Definition 2.102 the grid function

\[
\tau_{h,k}(x,t) := \begin{cases} 
D^+_h u(x,t) - L^{\Theta}_{h,k} u(x,t) - f_{h,k}(x,t), & (x,t) \in \Omega_h \times (\bar{I}_k \setminus \{0\}) \\
u(x,t) - u^D_{h,k}(x,t), & (x,t) \in \Gamma_h \times (\bar{I}_k \setminus \{0\}) \\
u(x,0) - u^0_h(x), & x \in \Omega_h
\end{cases}
\]

is called the local discretization error. The Θ-method is said to be consistent with the initial-boundary value problem, if

\[\|\tau_{h,k}\|_{h,k} \to 0 \quad \text{as} \quad h, k \to 0.\]

The Θ-method is said to be consistent of order \(p_1\) in \(h\) and \(p_2\) in \(k\), if there exists a constant \(C > 0\), independent of \(h\) and \(k\), such that for sufficiently small \(h, k\) it holds

\[\|\tau_{h,k}\|_{h,k} \leq C (h^{p_1} + k^{p_2}).\]

Under the assumption of smooth data and a sufficiently smooth classical solution, the consistency and the order of consistency of the Θ-method can be derived by multidimensional Taylor expansion.

**Theorem 2.104. Order of consistency of the Θ-method**

We consider the Θ-method (2.111a)-(2.111c) with \(f_{h,k} = f|_{Q_{h,k}}, u^D_h = u^D|_{\bar{\Gamma}_h \times \bar{I}_k}\), and \(u^0_h = u^0|_{\partial \Omega_h}\). We assume that the classical solution satisfies \(u \in C^2([0,T]; C^4(\Omega))\). Then it follows by multidimensional Taylor expansion that for \(\Theta \neq 1/2\) the order of consistency of the Θ-method is \(p_1 = 2\) in \(h\) and \(p_2 = 1\) in \(k\). Note that for the explicit Euler method the Taylor expansion is around \((x,t)\), whereas for the implicit methods the Taylor expansion is around \((x, t + k)\). If the classical solution satisfies \(u \in C^3([0,T]; C^4(\Omega))\), the Crank-Nicolson method \((\Theta = 1/2)\) has the order of consistency \(p_1 = 2\) in \(h\) and \(p_2 = 2\) in \(k\). In this case, the Taylor expansion is around \((x, t + k/2)\).

The stability of the Θ-method means continuous dependence of the solution \(u_{h,k}\) on the data \(f_{h,k}, u^D_{h,k}\), and \(u^0_h\).
**Definition 2.105. (Stability of the Θ-method)**

Let \( u_{h,k} \) and \( z_{h,k} \) be the solutions of the Θ-method (2.111a)-(2.111c) with respect to the ‘unperturbed’ data \( f_{h,k}, u_{h,k}^D, u_0^h \), and the ‘perturbed’ data \( f_{h,k} + \delta f_{h,k}, u_{h,k}^D + \delta u_{h,k}^D, u_0^h + \delta u_0^h \) with grid functions \( \delta g_{h,k} \in C(Q_{h,k}), \delta u_{h,k}^D \in C(\Gamma_h \times \bar{I}_k), \delta u_0^h \in C(\Omega_h) \). Then the Θ-method is said to be stable for \((h,k) \in \Lambda \subseteq \mathbb{R}^+ \times \mathbb{R}^+ \), if for each \( \epsilon > 0 \) there exists \( \delta = \delta(\epsilon, h, k) > 0 \) such that

\[
\| u_{h,k} - z_{h,k} \|_{h,k} < \epsilon
\]

for all perturbations \( \delta u_0^h \) and \((\delta g_{h,k}, \delta u_{h,k}^D)\) satisfying

\[
\| \delta u_0^h \|_h + \| (\delta f_{h,k}, \delta u_{h,k}^D) \|_{h,k} < \delta.
\]

The set \( \Lambda \) is called the stability region. If \( \Lambda = \mathbb{R}_+ ^+ \times \mathbb{R}_+ ^+ \), the Θ-method is said to be unconditionally stable, otherwise it is called conditionally stable.

As in case of finite difference approximations of elliptic boundary value problems, the stability of the consistent Θ-method is sufficient for the convergence.

**Theorem 2.106. (Sufficient condition for convergence)**

Assume that the Θ-method is consistent with the initial-boundary value problem. If the Θ-method is stable, it is convergent and the order of convergence corresponds to the order of consistency.

**Proof.** The solution \( u \) of the initial-boundary value problem satisfies the difference equations

\[
D_k^+ u(x,t) - L_h^0 u(x,t) = f_{h,k}^0(x,t) + \tau_{h,k}(x,t), \quad x \in \Omega_h, \quad t \in \bar{I}_k \setminus \{0\},
\]

\[
u(x,t) = u_{h,k}^D(x,t) + \tau_{h,k}(x,t), \quad x \in \Gamma_h, \quad t \in \bar{I}_k \setminus \{0\},
\]

\[
u(x,0) = u_0^h(x) + \tau_{h,k}(x,0), \quad x \in \Omega_h.
\]

Setting \( \delta f_{h,k}(x,t) = \tau_{h,k}(x,t), x \in \Omega_h, t \in \bar{I}_k \setminus \{0\}, \delta u_{h,k}^D = \tau_{h,k}(x,t), x \in \Gamma_h, t \in \bar{I}_k \setminus \{0\}, \) and \( \delta u_0^h(x) = \tau_{h,k}(x,0), x \in \Omega_h, \) and taking the consistency \( \| \tau_{h,k} \|_{h,k} \to 0 \) as \( h, k \to 0 \) into account, the assertion follows from the Definition 2.105.

**Remark 2.107. (Stability)**

We note that in the literature there is a wide variety of stability notions for finite difference approximations of parabolic initial-boundary value problems (cf., e.g., Thomas).

Theorem 2.106 is sometimes referred to as the Theorem of Lax. Under
the same assumptions, it can be shown that the stability is also necessary for the convergence of the Θ-method. The equivalence of the stability and the convergence of consistent finite difference approximations of parabolic initial-boundary value problems is called the equivalence theorem of Lax and Richtmyer (for a proof we refer to Strikwerda).

Since the Θ-method corresponds to a linear algebraic system, the stability can be established by arguments from numerical linear algebra. To this end, we assume $\Omega = (a, b)^2$, $a, b \in \mathbb{R}$, $a < b$. If $\Omega_h = \{x_1, \cdots, x_{n_h}\}^T$ and if $u_h^{(m)}$ denotes the vector $u_h^{(m)} = (u_{h,k}(x_1, t_m), \cdots, u_{h,k}(x_{n_h}, t_m))^T$, $0 \leq m \leq M + 1$, the Θ-method is equivalent to the linear algebraic system

\[
(I_h + kB_h(\Theta))u_h^{(m)} = (I_h - kC_h(\Theta))u_h^{(m-1)} + b_h^{(m)},
\]

where $B_h(\Theta) = \Theta A_h, C_h(\Theta) = (1 - \Theta)A_h$ with the block tridiagonal matrix $A_h \in \mathbb{R}^{n_h \times n_h}$ from (2.41) and the vector $b_h^{(m)} \in \mathbb{R}^{n_h}$. If $\delta u_h^{(0)} \in \mathbb{R}^{n_h}$ and $\delta b_h^{(m)} \in \mathbb{R}^{n_h}, 1 \leq m \leq M + 1$, are perturbations in the initial condition and the right-hand side, and if $z_h^{(m)}$ is the solution of (2.112) with respect to the perturbed data, by induction on $m$ it can be shown that

\[
\|u_h^{(m)} - z_h^{(m)}\| \leq \|D_h(\Theta)\|^m \|\delta u_h^{(0)}\| + \sum_{\ell=0}^{m-1} \|D_h(\Theta)\|^{m-\ell-1} \|(I_h + kB_h(\Theta))^{-1}\| \|\delta b_h^{(\ell)}\|,
\]

where $D_h(\Theta) := (I_h + kB_h(\Theta))^{-1}(I_h - kC_h(\Theta))$ and the matrix norm is the norm associated with the Euclidean vector norm. The estimate (2.113) implies that for stability it suffices to show $\rho(D_h(\Theta)) < 1$, where $\rho(D_h(\Theta))$ is the spectral radius of $D_h(\Theta)$. The knowledge of the eigenvalues of $A_h$ (cf. (2.43)) enables to compute the eigenvalues of $D_h(\Theta)$. For $\Theta < 1/2$, the step size restriction $k \leq h^2(2 - 4\Theta)$ implies $\rho(D_h(\Theta)) < 1$, whereas for $\Theta \geq 1/2$ we have $\rho(D_h(\Theta)) < 1$ without conditions on the step sizes. Hence, the Θ-method is conditionally stable for $\Theta < 1/2$ and unconditionally stable for $\Theta \geq 1/2$.

For further results on the stability of finite difference approximations of parabolic initial-boundary value problems we refer to Gustafsson/Kreiss/Oliger, Richtmyer/Morton, Strikwerda, and Thomas.
2.3.4 Finite element approximations

Finite element approximations of initial-boundary value problems for the heat equation

\[
\begin{align*}
\frac{\partial u}{\partial t} - \Delta u &= f \quad \text{in } Q, \\
u &= 0 \quad \text{on } \Sigma, \\
u(\cdot, 0) &= u_0 \quad \text{in } \Omega,
\end{align*}
\]

are based on the concept of a weak solution. Here, \( \Omega \subset \mathbb{R}^d \) is supposed to be a bounded Lipschitz-domain with boundary \( \Gamma = \partial \Omega \). Moreover, we assume

\[
f \in L^2((0, T); L^2(\Omega)), \quad u_0 \in L^2(\Omega).
\]

We note that for Hilbert spaces \( H \) with norm \( \| \cdot \|_H \) and \( V \) with norm \( \| \cdot \|_V \) such that \( V \) is continuously and densely embedded in \( H \), we refer to \( \langle \cdot, \cdot \rangle_{V^*, V} \) the dual product of \( V^* \) and \( V \). We refer to \( L^2((0, T); H) \) as the Hilbert space with norm

\[
\| u \|_{L^2((0, T); H)} := \left( \int_0^T \| u(t) \|_H^2 \, dt \right)^{1/2},
\]

and define \( L^2((0, T); V) \) and \( \| \cdot \|_{L^2((0, T); V)} \) analogously. Moreover, we introduce \( H^1((0, T); V^*) \) as the Hilbert space with the norm

\[
\| u \|_{H^1((0, T); V^*)} := \left( \int_0^T \left( \| u(t) \|_{V^*}^2 + \| u_t(t) \|_{V^*}^2 \right) dt \right)^{1/2}.
\]

**Definition 2.108. (Weak solution of the initial-boundary value problem)**

Under the assumptions (2.115) a function \( u \in H^1((0, T); H^{-1}(\Omega)) \cap L^2((0, T); H_0^1(\Omega)) \) is called a weak solution of the initial-boundary value problem (2.114a)-(2.114c), if for all \( v \in H_0^1(\Omega) \) it holds

\[
\begin{align*}
\langle \frac{\partial u}{\partial t}, v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)} + (\nabla u, \nabla v)_0,\Omega &= (f, v)_0,\Omega \quad \text{for almost all } t \in (0, T), \\
(u(\cdot, 0), v)_0,\Omega &= (u_0, v)_0,\Omega.
\end{align*}
\]

**Remark 2.109. (Pointwise continuity in t)**

It can be shown that the space \( H^1((0, T); H^{-1}(\Omega)) \cap L^2((0, T); H_0^1(\Omega)) \)
Numerical Differential Equations 137

is continuously embedded in \( C([0,T]; L^2(\Omega)) \). Hence, \( u(\cdot, 0) \) in (2.116b) is well defined. For a proof we refer to Renardy/Rogers.

**Theorem 2.110. (Existence and uniqueness of a weak solution)**

*Under the assumptions (2.115) the initial-boundary value problem (2.114a)-(2.114c) admits a unique weak solution.*

**Proof.** We refer to Renardy/Rogers. \( \square \)

The standard finite element method for the computation of a weak solution relies on a discretization in the spatial variables by restricting (2.116a), (2.116b) to a finite element space \( V_h \subset H^1_0(\Omega) \) with \( \dim V_h = n_h \). We are looking for \( u_h \in V_h \) such that for all \( v_h \in V_h \) it holds

\[
\frac{\partial u_h}{\partial t}, v_h \rangle_{0,\Omega} + \langle \nabla u_h, \nabla v_h \rangle_{0,\Omega} = \langle f(t), v_h \rangle_{0,\Omega},
\]

(2.117b) \( u_h(0), v_h \rangle_{0,\Omega} = \langle u_0, v_h \rangle_{0,\Omega} \).

The finite element approximation (2.117a),(2.117b) represents an initial value problem for a system of linear first order ordinary differential equations. If \( (\varphi^{(i)}_h)_{i=1}^{n_h} \) is a basis of \( V_h \), the solution \( u_h(t) \in V_h \), \( t \in [0,T] \), is of the form \( u_h(t) = \sum_{j=1}^{n_h} y_j(t) \varphi^{(j)}_h \). Hence, for \( y(t) = (y_1(t), \ldots, y_{n_h}(t))^T \) we obtain

(2.118a) \( M_h \frac{dy(t)}{dt} + A_h y(t) = b_h(t), \quad t \in (0,T], \)
(2.118b) \( y(0) = y_0 \).

Here, \( M_h \in \mathbb{R}^{n_h \times n_h} \) stands for the mass matrix

\[
M_h := \begin{pmatrix}
(\varphi^{(1)}_h, \varphi^{(1)}_h)_{0,\Omega} & \cdots & (\varphi^{(n_h)}_h, \varphi^{(1)}_h)_{0,\Omega} \\
\vdots & \ddots & \vdots \\
(\varphi^{(1)}_h, \varphi^{(n_h)}_h)_{0,\Omega} & \cdots & (\varphi^{(n_h)}_h, \varphi^{(n_h)}_h)_{0,\Omega}
\end{pmatrix}
\]

(2.119) and \( A_h \in \mathbb{R}^{n_h \times n_h} \) refers to the stiffness matrix

\[
A_h := \begin{pmatrix}
(\nabla \varphi^{(1)}_h, \nabla \varphi^{(1)}_h) & \cdots & (\nabla \varphi^{(n_h)}_h, \nabla \varphi^{(1)}_h) \\
\vdots & \ddots & \vdots \\
(\nabla \varphi^{(1)}_h, \nabla \varphi^{(n_h)}_h) & \cdots & (\nabla \varphi^{(n_h)}_h, \nabla \varphi^{(n_h)}_h)
\end{pmatrix}
\]

(2.120)
whereas the vectors $b_h(t) \in \mathbb{R}^{n_h}, t \in [0, T]$, and $y_0 \in \mathbb{R}^{n_h}, y_0 = (y_1^0, \ldots, y_{n_h}^0)^T$ are given by

\begin{equation}
(2.121a)
    b_h(t) := ((f(t), \varphi_h^{(1)}), \cdots, (f(t), \varphi_h^{(n_h)}))_0, \Omega, \Theta = 1
\end{equation}

\begin{equation}
(2.121b)
    y_j^0 = (u_0, \varphi_h^{(j)}), 0, \Omega, 1 \leq j \leq n_h.
\end{equation}

For the discretization in time we consider a partition $\bar{I}_k$ of the time interval $[0, T]$ according to (2.104) and approximate the time derivative by the forward difference quotient, by the backward difference quotient, or by a convex combination of both as in the $\Theta$-method (cf. Definition 2.101). We are looking for functions $u_{h,k}^{(m)} \in V_h, 0 \leq m \leq M + 1$, such that

\begin{equation}
(2.122a)
    k^{-1}M_h(u_{h,k}^{(m+1)} - u_{h,k}^{(m)}) + \Theta A_h u_{h,k}^{(m+1)} + \\
    + (1 - \Theta)A_h u_{h,k}^{(m)} = b_h^0(t_m, t_{m+1}), 0 \leq m \leq M,
\end{equation}

\begin{equation}
(2.122b)
    u_{h,0} = u_h^0,
\end{equation}

where $b_h^0(t_m, t_{m+1}) := \Theta b_h(t_{m+1}) + (1 - \Theta)b_h(t_m)$ and $u_h^0 = \sum_{j=1}^{n_h} y_j^0 \varphi_h^{(j)}$.

As far as a priori estimates of the global discretization error $e_{h,k}(t) := u(t) - u_{h,k}(t), t \in \bar{I}_k$, are concerned, we assume that for $v \in H^{r+1}(\Omega), r \in \mathbb{N}$, the finite element spaces $V_h$ satisfy the approximation property

\begin{equation}
(2.123)
    \inf_{v_h \in V_h} \left( \|v - v_h\|_{0, \Omega} + h \|v - v_h\|_{1, \Omega} \right) \leq Ch^{r+1},
\end{equation}

where $C > 0$ is a constant independent of $h$. If we use the backward difference quotient in time, i.e., $\Theta = 1$ in (2.122a), we have the following a priori error estimate in the $H^1$-seminorm which is a norm on $H^1_0(\Omega)$.

**Theorem 2.111.** (A priori error estimate in the $H^1$-seminorm for the implicit Euler method)

Let $u \in H^1((0, T); H^{r+1}(\Omega)) \cap H^2((0, T); L^2(\Omega)), r \in \mathbb{N}$, the weak solution of the initial-boundary value problem and let $u_{h,k}$ be the solution of (2.122a),(2.122b) for $\Theta = 1$. We further assume (2.123) and $|u^0 - u_{h,k}^0|_{1, \Omega} = O(h^r)$. Then, there exists a constant $C > 0$, independent of $h, k$, such that for all $0 \leq m \leq M + 1$ it holds

\begin{equation}
(2.124)
    |u(t_m) - u_{h,k}(t_m)|_{1, \Omega} \leq C(h^r + k).
\end{equation}

**Proof.** We refer to Thomée. \hfill \□

Under stronger regularity assumptions on the weak solution, for the Crank-Nicolson method, i.e., $\Theta = 1/2$, we obtain convergence of order 2 in $k$. 
Theorem 2.112. (A priori error estimate in the $H^1$-seminorm for the Crank-Nicolson method)
Assume that the weak solution satisfies
\[ u \in H^1((0,T); H^{r+1}(\Omega)) \cap H^2((0,T); H^2(\Omega)), \quad \Delta u_{tt} \in L^2((0,T); L^2(\Omega)). \]
Further, let $u_{h,k}$ be the solution of (2.122a),(2.122b) for $\Theta = 1/2$ and suppose that (2.123) and $|u^0 - u_h^0|_{1,\Omega} = O(h^r)$ hold true. Then, there exists a constant $C > 0$, independent of $h,k$, such that for all $0 \leq m \leq M + 1$ it holds
\[ |u(t_m) - u_{h,k}(t_m)|_{1,\Omega} \leq C(h^r + k^2). \]
\[ (2.125) \]
Proof. We refer to Thomée.

We recall Theorem 2.87 which provides a quasi-optimal a priori error estimate in the $L^2$-norm for $(r+1)$-regular elliptic boundary value problems. This result also holds true for the heat equation.

Theorem 2.113. (A priori error estimates in the $L^2$-norm for the $\Theta$-method)
In addition to the assumptions of Theorem 2.111 and Theorem 2.112 suppose that the boundary value problem for the Poisson equation (2.11) is $(r+1)$-regular. Then there exists a constant $C > 0$, independent of $h,k$, such that for all $0 \leq m \leq M + 1$ it holds
\[ \|u(t_m) - u_{h,k}(t_m)\|_{0,\Omega} \leq C(h^{r+1} + k^p) \]
with $p = 1$ for $\Theta = 1$ and $p = 2$ for $\Theta = 1/2$.
\[ (2.126) \]
Proof. We refer to Thomée.

Remark 2.114. (A priori error estimate in the $L^\infty$-norm)
For a priori error estimates in the $L^\infty$-norm we refer to Thomée.
2.4 The wave equation

2.4.1 Preliminaries

We consider the Cauchy problem for the spatially one-dimensional wave equation

\[ \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0, \quad (x, t) \in \mathbb{R} \times \mathbb{R}, \]  
\[(2.127a)\]

\[ u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x), \quad x \in \mathbb{R}, \]  
\[(2.127b)\]

where \(u_0\) and \(u_1\) are given functions in \(\mathbb{R}\).

Introducing the coordinate transformation

\[ \xi = t + x, \quad \eta = x - t, \]  
\[(2.128)\]

and the function

\[ v(\xi, \eta) := u\left(\frac{1}{2}(\xi + \eta), \frac{1}{2}(\xi - \eta)\right), \]  
\[(2.129)\]

it is easy to see that \(v\) satisfies the partial differential equation

\[ \frac{\partial^2 v}{\partial \xi \partial \eta} = 0, \quad (\xi, \eta) \in \mathbb{R} \times \mathbb{R}. \]  
\[(2.130)\]

Integration with respect to \(\eta\) implies the existence of a function \(\tilde{g}_1 = \tilde{g}_1(\xi)\) such that \(\frac{\partial v}{\partial \xi}(\xi, \eta) = \tilde{g}_1(\xi)\) and a further integration with respect to \(\xi\) implies the existence of functions \(g_1 = g_1(\xi)\) and \(g_2 = g_2(\eta)\) such that \(v(\xi, \eta) = g_1(\xi) + g_2(\eta)\). Hence, it follows from (2.128) and (2.129) that any solution of (2.127a) is of the form

\[ u(x, t) = g_1(x - t) + g_2(x - t). \]  
\[(2.131)\]

This result enables us to provide the following representation of the solution of the Cauchy problem (2.127a),(2.127b).

**Theorem 2.115. (Solution of the Cauchy problem for the spatially one-dimensional wave equation)**

Suppose that \(u_0 \in C^2(\mathbb{R}), u_1 \in C^1(\mathbb{R})\). Then, the unique solution of the Cauchy problem for the spatially one-dimensional wave equation (2.127a),(2.127b) is given by

\[ u(x, t) = \frac{1}{2} \int_{x-t}^{x+t} u_1(s) \, ds + \frac{\partial}{\partial t} \left( \frac{1}{2} \int_{x-t}^{x+t} u_0(s) \, ds \right). \]  
\[(2.132)\]

**Proof.** According to (2.131) the solution \(u\) satisfies

\[ u(x, t) = g_1(x - t) + g_2(x - t), \quad \frac{\partial u}{\partial t}(x, t) = g'_1(x + t) - g'_2(x - t). \]
Setting $t = 0$, we obtain the system of equations
\[ g_1(x) + g_2(x) = u_0(x), \quad g_1'(x) - g_2'(x) = u_1(x), \]
which has the unique solution
\[ g_1(x) = \frac{1}{2} \int_0^x u_1(s) \, ds + \frac{1}{2} \left( g_1(0) - g_2(0) \right) + \frac{1}{2} u_0(x), \]
\[ g_2(x) = \frac{1}{2} u_0(x) - \frac{1}{2} \int_0^x u_1(s) \, ds - \frac{1}{2} \left( g_1(0) - g_2(0) \right). \]
The assertion is an immediate consequence of these two equations. □

\begin{figure}
\centering
\begin{tikzpicture}
\draw[->, thick] (-3,0) -- (3,0) node[right] {$x$};
\draw[->, thick] (0,-3) -- (0,3) node[above] {$t$};
\draw (0,0) -- (3,0) -- (0,3) -- cycle;
\node at (0,0) {$x_0$};
\node at (3,0) {$x_0 + t_0$};
\node at (0,3) {$x_0 - t_0$};
\node at ($(0,0) + (0.5,0.5)$) {$D$};
\node at ($(0,0) + (0.5,0.5) + (2,2)$) {$(x_0, t_0)$};
\end{tikzpicture}
\caption{Domain of dependence and domain of influence.}
\end{figure}

**Definition 2.116.** (Domain of dependence and domain of influence)
It follows from (2.132) that for $(x_0, t_0) \in \mathbb{R}^2$ the solution $u$ in
\[ D = D(x_0, t_0) := \{(x, t) \in \mathbb{R}^2 \mid 0 \leq t \leq t_0, \ (t - t_0)^2 - (x - x_0)^2 \geq 0\} \]
is uniquely determined by the initial values in the interval
\[ I = I(x_0, t_0) := [x_0 - t_0, x_0 + t_0]. \]
$D$ is called the domain of dependence and $I$ is said to be the domain of influence (cf. Figure 15).
Remark 2.117. (Finite speed of propagation)
It also follows from (2.132) and Figure 15 that the Cauchy data propagate with finite speed. This is in contrast to the heat equation where the Cauchy data propagate with infinite speed.

The representation formula (2.132) for the solution of the Cauchy problem for the spatially one-dimensional wave equation can be generalized to higher dimensions $d > 1$, i.e., for the Cauchy problem

\[(2.133a) \quad \frac{\partial^2 u}{\partial t^2} - \Delta u = 0, \quad (x, t) \in \mathbb{R}^d \times \mathbb{R}_+,
\]
\[(2.133b) \quad u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x), \quad x \in \mathbb{R}^d,
\]
where $u_0$ and $u_1$ are given functions in $\mathbb{R}^d$.

However, we have to distinguish between $d$ even and $d$ odd. We refer to $S^{d-1}$ as the unit sphere in $\mathbb{R}^d$ and to $|S^{d-1}|$ as the area of $S^{d-1}$ which is given by

\[|S^{d-1}| = \frac{2\pi^{d/2}}{\Gamma(d/2)},\]

where $\Gamma$ stands for the Gamma function $\Gamma(x) := \int_0^\infty s^{x-1} \exp(-s) \, ds$.

Moreover, for a map $g$ in $\mathbb{R}^d$ we introduce the function

\[M(x, t, g) := \frac{1}{|S^{d-1}|} \int_{S^{d-1}} g(x + t\xi) \, d\sigma(\xi).
\]

Theorem 2.118. (Solution of the Cauchy problem for the spatially $d$-dimensional wave equation: $d$ odd)
Assume $d = 2m + 1, m \geq 1$, and $u_0, u_1 \in C^{(d+3)/2}(\mathbb{R}^d)$. Then, the unique solution of the Cauchy problem (2.133a),(2.133b) is given by

\[u(x, t) = \left( \prod_{i=1}^{m-1} (2i + 1) \right)^{-1} \left( \frac{\partial}{\partial t} \left( \frac{1}{t} \frac{\partial}{\partial t} \right)^{(d-3)/2} (t^{d-2} M(x, t, u_0)) + \right)
\]
\[(2.134) \quad \left( \frac{1}{t} \frac{\partial}{\partial t} \right)^{(d-3)/2} (t^{d-2} M(x, t, u_1)).\]

Proof. We refer to Evans. \qed
Theorem 2.119. (Solution of the Cauchy problem for the spatially \( d \)-dimensional wave equation: \( d \) even)

Assume \( d = 2m, m \geq 1 \), and \( u_0, u_1 \in C^{(d+4)/2}(\mathbb{R}^d) \). Then, the unique solution of the Cauchy problem (2.133a),(2.133b) is given by

\[
 u(x, t) = \left( \max(1, \prod_{i=1}^{m-1} 2i) \right)^{-1} \cdot \\
 \cdot \left( \frac{1}{t} \frac{\partial}{\partial t} \right)^{(d-2)/2} \left( \int_0^t \frac{s^{d-1}}{\sqrt{t^2 - s^2}} M(x, s, u_0) \, ds \right) + \\
\left( \frac{1}{t} \frac{\partial}{\partial t} \right)^{(d-2)/2} \left( \int_0^t \frac{s^{d-1}}{\sqrt{t^2 - s^2}} M(x, s, u_1) \, ds \right).
\]

(2.135)

\( \Box \)

Proof. We refer to Evans.
2.4.2 Finite difference approximations

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with boundary $\Gamma = \partial \Omega$. For $T > 0$ we set $Q := \Omega \times (0, T)$ and $\Sigma := \Gamma \times (0, T)$. Given functions $f : Q \to \mathbb{R}$ and $u_0, u_1 : \Omega \to \mathbb{R}$, as well as $u_D : \Gamma \times (0, T) \to \mathbb{R}$, we consider the following initial-boundary value problem for the wave equation

\begin{align*}
(2.136a) & \quad \frac{\partial^2 u}{\partial t^2} - \Delta u = f, \quad (x, t) \in Q, \\
(2.136b) & \quad u(x, t) = u_D(x, t), \quad (x, t) \in \Sigma, \\
(2.136c) & \quad u(x, 0) = u_0(x), \quad \frac{\partial u}{\partial t}(x, 0) = u_1(x), \quad x \in \Omega,
\end{align*}

For the finite difference approximation of (2.136a)-(2.136c) we consider an equidistant space-time grid $Q_{h,k}$ with spatial step size $h$ and time step size $k$ (cf. (2.106)). We approximate the Laplacian $\Delta$ by the discrete Laplacian $\Delta_h$ (cf. (2.34)) and the time derivative $\partial^2 u/\partial t^2$ by the central difference quotient

$$D_k^2u(x, t) := \frac{u(x, t + k) - 2u(x, t) + u(x, t - k)}{k^2}.$$ 

For the approximation of the second initial condition $\partial u(x, 0)/\partial t = u_1(x), x \in \Omega$, we introduce the artificial time level $t_{-1} = -k$ and use the central difference quotient

$$D_k^c u(x, 0) := \frac{u(x, k) - u(x, -k)}{2k}.$$ 

The finite difference approximation of (2.136a)-(2.136c) amounts to the computation of a grid function $u_{h,k} : \tilde{Q}_{h,k} \to \mathbb{R}$ on the extended grid $\tilde{Q}_{h,k}$ such that

\begin{align*}
(2.137a) & \quad D_k^2u_{h,k}(x, t) - \Delta_h u_{h,k}(x, t) = f(x, t), \quad (x, t) \in Q_{h,k} \cap Q, \\
(2.137b) & \quad u_{h,k}(x, t) = u_D(x, t), \quad (x, t) \in Q_{h,k} \cap \Sigma, \\
(2.137c) & \quad u_{h,k}(x, 0) = u_0(x), \quad D_k^c u_{h,k}(x, 0) = u_1(x), \quad x \in Q_{h,k} \cap \overline{\Omega}.
\end{align*}

From the second equation in (2.137c) we deduce

\begin{align*}
(2.138) & \quad u_{h,k}(x, -k) = u_{h,k}(x, k) - 2ku_1(x).
\end{align*}
If we assume (2.137a) to hold true on the time level \( t_0 = 0 \) as well and solve for \( u_{h,k}(x, k) \), we obtain

\[
(2.139) \quad u_{h,k}(x, k) = \sum_{i=1}^{d} \left( \left( \frac{k}{h} \right)^2 u_0(x + h e_i) + 2(1 - \left( \frac{k}{h} \right)^2) u_0(x) + \left( \frac{k}{h} \right)^2 u_0(x - h e_i) \right) - u_{h,k}(x, -k) + k^2 f(x, 0).
\]

Combining (2.138) and (2.139) finally yields

\[
(2.140) \quad u_{h,k}(x, k) = \frac{1}{2} \sum_{i=1}^{d} \left( \left( \frac{k}{h} \right)^2 u_0(x + h e_i) + 2(1 - \left( \frac{k}{h} \right)^2) u_0(x) + \left( \frac{k}{h} \right)^2 u_0(x - h e_i) \right) + k u_1(x) + k^2 f(x, 0).
\]

Using the first equation in (2.137a) and (2.140) as initial conditions on the time levels \( t_0 = 0 \) and \( t_1 = k \), we thus obtain an explicit finite difference scheme where we can explicitly compute \( u_{h,k}(x, t_k), 2 \leq k \leq M \).

The domain of dependence of the solution \( u_{h,k}(x, t_m) \) of the finite difference approximation are all grid points in

\[
(2.141) \quad \prod_{i=1}^{d} [x - m h e_i, x + m h e_i],
\]

whereas the domain of dependence of the exact solution \( u(x, t_m) \) is given by

\[
(2.142) \quad \prod_{i=1}^{d} [x - m k e_i, x + m k e_i].
\]

For stability reasons we have to assume that the domain of dependence of the exact solution is contained in the domain of dependence of the solution of the finite difference approximation, i.e., it follows from (2.141) and (2.142) that we have to require

\[
\prod_{i=1}^{d} [x - m k e_i, x + m k e_i] \subseteq \prod_{i=1}^{d} [x - m k e_i, x + m k e_i]
\]

which holds true, if the step sizes \( h \) and \( k \) satisfy

\[
(2.143) \quad \frac{k}{h} \leq 1.
\]

The stability condition (2.143) is called the Courant-Friedrichs-Levy (CFL) condition.
For sufficiently smooth exact solutions $u$ we obtain the following convergence result:

**Theorem 2.120.** (Convergence of the finite difference approximation of the initial-boundary value problem for the wave equation)

Assume that the exact solution $u$ of (2.136a)-(2.136c) satisfies $u \in C^4([0,T]; C^2(\Omega)) \cap C([0,T], C^4(\Omega))$ and that the finite difference approximation fulfills the CFL condition (2.143). Then there exist constants $C_1 > 0$ and $C_2 > 0$, independent of $h, k$, such that

$$(2.144) \quad |u_{h,k}(x,t) - u(x,t)| \leq C_1 h^2 + C_2 k^2, \quad (x,t) \in Q_{h,k}.$$
2.4.3 Finite element approximations

We consider the initial-boundary value problem (2.136a)-(2.136c) with $u_D \equiv 0$. We multiply (2.136a) by a function $v \in H^1(Q)$ with $v(\cdot, t) \in H^1_0(\Omega)$ for almost all $t \in (0, T)$, and $v(x, T) = 0, x \in \Omega$, and integrate over $Q$:

\begin{equation}
\int_Q \frac{\partial^2 u}{\partial t^2} \, v \, dx \, dt - \int_Q \Delta u \, v \, dx \, dt = \int_Q f \, v \, dx \, dt.
\end{equation}

Partial integration of the first integral on the left-hand side in (2.145) yields

\begin{equation}
\int_Q \frac{\partial^2 u}{\partial t^2} \, v \, dx \, dt = \int_0^T \int_\Omega \frac{\partial^2 u}{\partial t^2} \, v \, dx \, dt - \int_0^T \int_\Omega \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} \, dx \, dt + \int_\Omega \frac{\partial u(x, T)}{\partial t} v(x, T) \, dx - \int_\Omega \frac{\partial u(x, 0)}{\partial t} v(x, 0) \, dx.
\end{equation}

On the other hand, for the second integral on the left-hand side in (2.145) Green’s formula (2.28) implies

\begin{equation}
\int_Q \Delta u \, v \, dx \, dt = \int_0^T \int_\Omega \Delta u \, v \, dx \, dt = - \int_0^T \int_\Omega \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} \, dx \, dt + \int_\Omega \nu \cdot \nabla u \cdot \nabla v \, dx \, dt - \int_\Gamma \nu \cdot \nabla u \cdot v \, d\sigma(x) \, dt.
\end{equation}

The equations (2.146),(2.147) motivate the definition of a weak solution of (2.136a)-(2.136c).

**Definition 2.121. (Weak solution of the initial-boundary value problem for the wave equation)**

Assume that $u_0, u_1 \in L^2(\Omega), u_D \equiv 0$, and $f \in L^2(Q)$. A function $u \in H^1(Q)$ with $u(\cdot, t) = 0 \in H^1_0(\Omega)$ for almost all $t \in (0, T)$, and $u(\cdot, 0) = u_0$ is called a weak solution of the initial-boundary value problem (2.136a)-(2.136c), if for all $v \in H^1(Q)$ with $v(\cdot, t) \in H^1_0(\Omega)$ for almost all $t \in (0, T)$, and $v(x, T) = 0, x \in \Omega$, and
it holds
\begin{equation}
\int_Q \frac{\partial u}{\partial t} \frac{\partial v}{\partial t} \, dx \, dt - \int_Q \nabla \cdot \nabla v \, dx \, dt = - \int_Q f v \, dx \, dt - \int_\Omega u_1 v(\cdot, 0) \, dx.
\end{equation}

\textbf{Theorem 2.122. (Existence and uniqueness of a weak solution)}

Under the assumptions of Definition 2.121 the initial-boundary value problem (2.136a)-(2.136c) admits a unique weak solution.

\textit{Proof.} The proof can be done by a Galerkin approximation in the space variables, the solution of the initial-value problem of the resulting system of ordinary differential equations, and appropriate a priori estimates. For details we refer to Evans. \hfill \Box

For the semi-discretization in space of (2.136a)-(2.136c) we choose \( V_h \subset H^1_0(\Omega) \), \( \dim V_h = N_h \), as the finite element space of continuous, piecewise linear finite elements with respect to a simplicial triangulation of the spatial domain \( \Omega \). We further choose \( u_{h,0}, u_{h,1} \in V_h \). Then the spatial semi-discretization of (2.136a)-(2.136c) requires the computation of a function \( u_h \in H^2((0,T), V_h) \) such that for all \( v_h \in V_h \) it holds
\begin{align}
\int_\Omega \frac{\partial^2 u_h}{\partial t^2} \, v_h \, dx + \int_\Omega \nabla u_h \cdot \nabla v_h \, dx &= \int_\Omega f v_h \, dx, \tag{2.149a} \\
u_h(\cdot, 0) &= u_{h,0}, \quad \frac{\partial u_h(\cdot, 0)}{\partial t} = u_{h,1}. \tag{2.149b}
\end{align}

We note that (2.149a),(2.149b) represents an initial-value problem for a system of linear second order ordinary differential equations which has a unique solution.

For the global discretization error \( u - u_h \) of the semi-discretized initial-boundary value problem we have the following a priori estimate:

\textbf{Theorem 2.123. (A priori error estimate for the semi-discretized wave equation)}

Assume that the weak solution \( u \) of (2.136a)-(2.136c) satisfies \( u \in C([0,T], H^2(\Omega) \cap H^1_0(\Omega)), \partial u/\partial t \in C([0,T], H^2(\Omega)), \) and \( \partial^2 u/\partial t^2 \in L^1((0,T), H^2(\Omega)) \). Further, let \( u_{h,0} \) and \( u_{h,1} \) be the elliptic projections of \( u_0 \) and \( u_1 \) onto \( V_h \) and let \( u_h \in H^2((0,T), V_h) \) be the unique solution of
the spatially semi-discretized problem (2.149a), (2.149b). Then it holds

\[
\begin{align*}
(2.150a) & \quad \sup_{t \in (0,T)} \| u - u_h \|_{0,\Omega} \leq C_1 h^2, \\
(2.150b) & \quad \sup_{t \in (0,T)} \| \nabla (u - u_h) \|_{0,\Omega} \leq C_2 h + C_3 h^2.
\end{align*}
\]

The constants \( C_i, 1 \leq i \leq 3 \), are given by

\[
\begin{align*}
C_1 & := c \left( \sup_{t \in (0,T)} \| \partial_t u \|_{2,\Omega} + \int_0^T \| \partial_{tt} u \|_{2,\Omega} \, dt \right), \\
C_2 & := c \sup_{t \in (0,T)} \| u \|_{2,\Omega}, \\
C_3 & := c \int_0^T \| \partial_{tt} u \|_{2,\Omega} \, dt,
\end{align*}
\]

where \( c \) is a positive constant independent of \( u \).

Proof. We refer to Johnson. \( \square \)

We choose \( \{ \varphi_h^{(i)} \}_{i=1}^{n_h} \) as the nodal basis of \( V_h \). Then \( u_h(\cdot, t), t \in [0,T] \), can be written as a linear combination of the basis functions:

\[
u_h(\cdot, t) = \sum_{j=1}^{n_h} u_j(t) \varphi_h^{(j)}.
\]

We refer to \( u_h(t) \) as the vector-valued function \( u_h(t) := (u_1(t), \cdots, u_{n_h}(t))^T \) and note that the elliptic projections \( u_{h,0}, u_{h,1} \) of \( u_0, u_1 \) onto \( V_h \) can be identified with vectors \( u_{h,0} = (u_{h,0,1}, \cdots, u_{h,0,n_h})^T, u_{h,1} = (u_{h,1,1}, \cdots, u_{h,1,n_h})^T \). We further denote by \( M_h \in \mathbb{R}^{n_h \times n_h}, A_h \in \mathbb{R}^{n_h \times n_h}, b_h(t) \in \mathbb{R}^{n_h}, t \in [0,T], \) the mass matrix, the stiffness matrix, and the load vector. Then (2.149a), (2.149b) can be written as an initial-value problem for a system of linear second order ordinary differential equations

\[
\begin{align*}
(2.151a) & \quad M_h \frac{\partial^2 u_h}{\partial t^2} + A_h u_h = b_h, \quad t \in (0,T], \\
(2.151b) & \quad u_h(0) = u_{h,0}, \quad \frac{\partial u_h}{\partial t}(0) = u_{h,1}.
\end{align*}
\]

As shown in Chapter 1, by introducing \( v_h(t) = \partial u_h(t)/\partial t \) as a new variable, the initial-value problem (2.151a), (2.151b) is equivalent to the following initial-value problem for a system of first order ordinary
differential equations

\begin{align}
(2.152a) \quad \frac{\partial u_h}{\partial t} - v_h &= 0, \quad t \in (0, T], \\
(2.152b) \quad M_h \frac{\partial v_h}{\partial t} + A_h u_h &= b_h, \quad t \in (0, T], \\
(2.152c) \quad u_h(0) &= u_{h,0}, \quad v_h(0) = u_{h,1}.
\end{align}

For the time discretization of (2.152a)-(2.152c) we choose an equidistant partial

\[ I_k := \{0 = t_0 < t_1 < \cdots < t_M := T\} \]

of the time interval \([0, T]\) of step size \(k := T/M\) and denote by \(u_{h,m}, v_{h,m}, 0 \leq m \leq M\), approximations of \(u_h(t_m), v_h(t_m)\). We discretize in time by the Crank-Nicolson method and obtain

\begin{align}
(2.153a) \quad \frac{v_{h,m+1} - v_{h,m}}{k} - \frac{1}{2} (v_{h,m+1}^2 + v_{h,m}^2) &= 0, \\
(2.153b) \quad M_h \frac{v_{h,m+1} - v_{h,m}}{k} + \frac{1}{2} A_h (u_{h,m+1}^2 + u_{h,m}^2) &= \frac{1}{2} (b_h(t_{m+1}) + b_h(t_m)), \\
(2.153c) \quad u_{h,0} &= u_{h,0}, \quad v_{h,0} = u_{h,1}.
\end{align}

The system (2.153a)-(2.153c) can be decoupled resulting in

\begin{align}
(2.154a) \quad (M_h + \frac{1}{4} k^2 A_h) u_{h,m+1} &= M_h u_{h,m} + \frac{1}{2} k A_h v_{h,m} - \frac{1}{4} k^2 A_h u_{h,m}^2 + \frac{1}{4} k^2 (b_h(t_{m+1}) + b_h(t_m)), \\
(2.154b) \quad M_h v_{h,m+1} &= M_h v_{h,m} - \frac{1}{2} k A_h (u_{h,m+1}^2 + u_{h,m}^2), \\
(2.154c) \quad u_{h,0} &= u_{h,0}, \quad v_{h,0} = u_{h,1}.
\end{align}

Both equations (2.154a) and (2.154b) can be solved by the preconditioned conjugate gradient method with the symmetric Gauss-Seidel iteration as a preconditioner.

The following result provides an a priori error estimate for the global discretization error \(u_{h,m} - u(\cdot, t_m)\):

**Theorem 2.124. (A priori error estimate for the fully discretized wave equation)**

Assume that for some \(r \in \mathbb{N}\) the weak solution \(u\) of (2.136a)-(2.136c) satisfies \(u \in C^3([0, T]; L^2(\Omega)) \cap C^2([0, T]; H^r(\Omega) \cap H_0^1(\Omega)) \cap C([0, T]; H^{r+1}(\Omega) \cap H_0^1(\Omega))\) and that \(u_{h,m} \in V_h, v_{h,m} \in V_h, 0 \leq m \leq M\), is the solution of the Crank-Nicolson approximation (2.154a)-(2.154c), where \(V_h\) is the Lagrangean finite element space composed by simplicial Lagrangean finite elements of type \(r\) with respect to a shape regular family
of geometrically conforming simplicial triangulations. Then it holds
\[
\|u_h^m - u(\cdot, t_m)\|_{0, \Omega} \leq C_1^m k^2 + C_2^m h^{r+1}, \quad 1 \leq m \leq M,
\]
where the constants $C_1^m$ and $C_2^m$ are given by
\[
C_1^m := \sup_{0 \leq t \leq t_m} \|\frac{\partial^3 u(\cdot, t)}{\partial t^3}\|_{0, \Omega} + \sup_{0 \leq t \leq t_m} \|\frac{\partial^2 \nabla u(\cdot, t)}{\partial t^2}\|_{0, \Omega},
\]
\[
C_2^m := \sup_{0 \leq t \leq t_m} \|\frac{\partial^2 \nabla^r u(\cdot, t)}{\partial t^2}\|_{0, \Omega} + \sup_{0 \leq t \leq t_m} \|\nabla^{r+1} u (\cdot, t)\|_{0, \Omega}.
\]

Proof. We refer to Johnson. \qed
Literature

Appendix: Exercises

**Exercise 1 (Non-Uniqueness of Initial-Value Problems)**

For given $x_0 < x_1 \in \mathbb{R}, y_0 \in \mathbb{R}$ and $\alpha > 0$ consider the rectangle
\[ R := \{(x,y) \in \mathbb{R}^2 \mid x \in [x_0, x_1], \ |y - y_0| \leq \alpha\} \]
and assume $f : R \to \mathbb{R}$ to be a continuous function satisfying
\[ \sup_{(x,y) \in R} |f(x,y)| \leq \frac{\alpha}{|x_1 - x_0|}. \]

Let $\Phi_1$ and $\Phi_2$ be two solutions of the initial-value problem
\[ y'(x) = f(x,y(x)), \ x \geq x_0, \ y(x_0) = y_0, \]
such that $\Phi_1(x_1) < \Phi_2(x_1)$.

(i) Show that for any $\Phi_1(x_1) < \eta < \Phi_2(x_1)$ there exists a solution $\tilde{\Phi}$ of the initial-value problem satisfying $\tilde{\Phi}(x_1) = \eta$.

(ii) Compute all solutions of the initial-value problem
\[ y'(x) = -\frac{1}{y(x)}\sqrt{1 - y^2(x)}, \ x \geq 0, \ y(0) = 1, \]
and verify (i) for $x_0 = 0, x_1 = 1$ and $y_0 = 1$. What is the reason for the non-uniqueness of the initial-value problem?

[Hint: Prove the existence of a solution $\kappa \in C^1([x_0, x_1])$ of the final-time problem
\[ (*) \quad y'(x) = f(x,y(x)), \ x \in [x_0, x_1], \ y(x_1) = \eta, \]
and consider
\[ \tilde{\Phi}(x) = \begin{cases} \Phi_\mu(x), & x \leq \xi, \\ \kappa(x), & x > \xi, \end{cases} \]
where $\xi := \inf \{x \in [x_0, x_1] \mid \Phi_1(x) < \kappa(x) < \Phi_2(x)\}$ and $\kappa(\xi) = \Phi_\mu(\xi)$ for $\mu \in \{1, 2\}$.]
Exercise 2 (Ordinary Differential Equation of Order m)

Given \( I := [x_0, x_1] \subset \mathbb{R} \), a function \( f : I \times \mathbb{R}^m \rightarrow \mathbb{R}, m \in \mathbb{N}, m \geq 2 \), and \( y^{(i)}_0 \in \mathbb{R}, 0 \leq i \leq m - 1 \), consider the initial-value problem

\[
y^{(m)}(x) = f(x, y(x), y'(x), \ldots, y^{(m-1)}(x)) , \quad y^{(i)}(x_0) = y^{(i)}_0 , \quad 0 \leq i \leq m - 1 .
\]

(i) Show that the initial-value problem can be transformed to a first order system

\[
Y'(x) = F(x, Y(x)) , \quad x \geq x_0 , \quad Y(x_0) = Y_0 .
\]

(ii) Establish conditions on \( f \) which guarantee the unique solvability of the initial-value problem.

(iii) Given \( a_i \in \mathbb{R} , 0 \leq i \leq 4 , a_4 \neq 0 \), reformulate the initial-value problem

\[
a_4 x^4 y^{(4)}(x) + a_3 x^3 y^{(3)}(x) + a_2 x^2 y^{(2)}(x) + a_1 x y'(x) + a_0 y(x) = 0 , \quad y(0) = y'(0) = y^{(2)}(0) = y^{(3)}(0) = 0
\]

for the (implicit) fourth order Euler equation as a first order system.

Exercise 3 (Condition of Initial-Value Problems)

The condition of an initial-value problem can be interpreted as a measure that indicates how sensitive the solution of the problem

\[
y'(x) = f(x, y(x)) , \quad x \in I := [a, b] \subset \mathbb{R} , \quad y(a) = \alpha \in (0, 1)^n
\]

depends on the initial data \( \alpha \). A suitable tool for the sensitivity analysis is given by the differential analysis of the condition:

Taking the derivative of the solution with respect to the initial data, we obtain a Jacobi matrix of absolute componentwise condition numbers which is called the Wronski matrix. Condition numbers \( \sigma(a, b) \) with respect to suitable vector norms \( \| \cdot \| \) in \( \mathbb{R}^n \) are obtained by the corresponding matrix norms of the Wronski matrix

\[
\sigma(a, b) := \max_{x \in [a, b]} \| \frac{\partial y(x)}{\partial \alpha} \| .
\]
Analyze the condition number of the initial-value problem for the logistic differential equation

\[ y'(x) = \lambda y(x) (1 - y(x)), \quad \lambda > 0, \]
\[ y(0) = \alpha, \]

which models the limited growth of populations caused by a shortage of resources.

(i) Determine analytically the exact solution of the initial-value problem.

(ii) Compute the condition \( \sigma(0, x) \) on the interval \([0, x], x > 0\).

(iii) For \( \lambda = 10 \) and \( \alpha = 0.01 \) display both the exact solution \( y(x) \) and the condition \( \sigma(0, x) \) on the interval \([0, 1]\) by using a common coordinate system with differently scaled ordinate.

**Exercise 4** (*Initial value problem with discontinuous right-hand side*)

Consider the initial value problem

\[ y'(x) = -\text{sgn}(y(x)), \quad x \geq -1, \]
\[ y(-1) = 1. \]

(i) Compute the exact solution of the differential equation. Here, exact solution means a piecewise continuously differentiable function which satisfies the differential equation almost everywhere.

(ii) For \( x_i := -1 + ih, i \in \mathbb{N}_0, \) and \( 0 < h < 1 \) consider the explicit Euler method

\[ y_{i+1} = y_i - h \text{sgn}(y_i), \quad y_0 = 1. \]

Show that for \( n \in \mathbb{N} \) with \( nh \leq 1 < (n+1)h \) and all \( k \in \mathbb{N} \) there holds

\[ y_{n+2k} = y_n, \quad y_{n+2k+1} = y_{n+1}. \]

(iii) Give a reason for the oscillating behavior of the approximations. How can you avoid the oscillations?

**Exercise 5** (*Midpoint rule applied to a model problem*)

For the numerical integration of the model problem

\[ y'(x) = \lambda y(x), \quad x \geq a, \quad y(a) = \alpha \]
consider the explicit midpoint rule with an initial Euler step using an equidistant grid $x_i := a + ih, h > 0$:

\begin{align*}
y_{i+1} &= y_{i-1} + 2h\lambda y_i, \ i \geq 1, \\
y_1 &= y_0 + h\lambda y_0, \\
y_0 &= \alpha.
\end{align*}

(i) Determine functions $g_n(\lambda h), 1 \leq n \leq 2$, such that the sequences

\begin{equation}
y_i^{(\alpha)} := g_i(\lambda h)y_0
\end{equation}

satisfy the difference equation associated with the explicit midpoint rule.

(ii) Assume $\lambda \ll 0$. Does the approximate solution show the same qualitative behavior as the exact solution?

(iii) Determine $\mu_k, 1 \leq k \leq 2$, depending on $\lambda$ and $h$, such that

\begin{equation}
y_i = (\mu_1 g_1^{(\lambda h)} + \mu_2 g_2^{(\lambda h)}) \alpha.
\end{equation}

Derive a modification of the Euler initial step to improve the quality of the method in case $\lambda \ll 0$.

**Exercise 6 (Lady Windermere’s fan)**

Assume that $f : I \times D \to D, I := [a, b] \subset \mathbb{R}, D \subset \mathbb{R}^m$, is continuous on $I \times D$ and satisfies the Lipschitz condition

\begin{equation}
\|f(x, y_1) - f(x, y_2)\| \leq L \|y_1 - y_2\|, \ x \in I, \ y_1, y_2 \in D, \ L > 0.
\end{equation}

For the numerical integration of the initial value problem

\begin{align*}
y'(x) &= f(x, y(x)), \ x \in I := [a, b] \subset \mathbb{R}, \\
y(a) &= \alpha \in D
\end{align*}

consider the explicit one-step method

\begin{align*}
y_{i+1} &= y_i + h_i \Phi(x_i, y_i, h_i; f), \\
y_0 &= \alpha,
\end{align*}

where $a = x_0 < x_1 < \ldots < x_N = b$ is a not necessarily equidistant partition of $I$ with step sizes $h_i := x_{i+1} - x_i, 0 \leq i \leq N - 1$.

Prove the following assertion: If the one-step method is consistent of order $p \in \mathbb{N}$, i.e.,

\begin{equation}
\|r_h(x, y)\| \leq C h_{\text{max}}^p, \ x \in I, \ y \in D,
\end{equation}

\begin{proof}

\end{proof}
where $h_{\text{max}} := \max_{0 \leq i \leq n-1} (x_{i+1} - x_i)$ and $C \in \mathbb{R}_+$ is independent of $h_{\text{max}}$, then there exists a constant $C' \in \mathbb{R}_+$, independent of $h_{\text{max}}$, such that

$$\|y_N - y(b)\| \leq C' \exp(L(b - a) - 1) \frac{h_{\text{max}}}{L}.$$  

[Hint: Use the technique known as Lady Windermere’s fan which is illustrated in the figure below:

![Diagram of Lady Windermere's fan](https://example.com/diagram)

Construct $N$ initial value problems of the form

$$y'(x) = f(x, y(x)) \ , \ x \geq x_i \ , \ 0 \leq i \leq N - 1 ,$$

$$y(x_i) = y_i$$

with the solutions $y(x; x_i, y_i)$. The quantities $e_i := \|y(x_i; x_{i-1}, y_{i-1}) - y_i\|$ are available by means of the local discretization error. Interpreting $y(x; x_i, y_i)$ as the solution of the perturbed initial value problem

$$y'(x) = f(x, y(x)) \ , \ x \geq x_i \ , \ 0 \leq i \leq N - 1 ,$$

$$y(x_{i+1}) = y_{i+1} + \frac{h_i}{\text{perturbation}} \tau_{hi}(x_i, y_i),$$

the error propagation can be accessed by Gronwall’s lemma.]

**Exercise 7 (Maximum consistency order of s-stage Runge Kutta methods)**

Show that the application of $s$-stage Runge Kutta methods to the initial-value problem

$$y'(x) = \lambda \, y(x) \ , \ y(x_0) = y_0$$

gives rise to increment functions of the form

$$k_j(x_k, y_k) = p_j(\lambda h) \, y_k \ , \ 1 \leq j \leq s ,$$
where \( p_j \) are polynomials of degree at most \( j \). Use this result for the determination of the maximum consistency order of \( s \)-stage Runge Kutta methods.

**Exercise 8 (Runge Kutta methods without memory)**

(i) Compute all Runge Kutta methods of order 2 having the Butcher scheme

\[
\begin{array}{c|ccc}
  c_1 & 0 & \\
  c_2 & c_2 & 0 & c_3 \\
  c_3 & 0 & 0 & 1
\end{array}
\]

(ii) In algorithmic notation derive an efficient implementation with respect to storage.

(iii) Compute the polynomials \( p_j \), \( 1 \leq j \leq 3 \), as introduced in Exercise 7.

**Exercise 9 (Asymptotic expansion for the explicit Euler method)**

Gragg’s theorem tells us that the coefficients in the asymptotic expansion of the global discretization error are solutions of inhomogeneous linear initial-value problems. These initial-value problems should be derived for a specific problem: The discretization of the initial-value problem \( y'(x) = y(x) \), \( y(a) = \alpha \) by the explicit Euler scheme with constant step size \( h \) gives rise to the difference equation

\[ y_h(x + h) = y_h(x) + h \, y_h(x) , \quad y_h(a) = \alpha . \]

Consider the asymptotic expansion

\[ y_h(x) = \sum_{k=0}^{\infty} e_k(x) \, h^k \]

and compute the staggered system of differential equations for the determination of the coefficients \( e_k, k \in \mathbb{N}_0 \), as well as the initial conditions. Compute \( e_0 \) and \( e_1 \) explicitly.

**Exercise 10 (Consistency of linear multi-step methods)**

Assume \( f \in C^p(I \times D) \), \( I := [a, b] \subset \mathbb{R}, D \subset \mathbb{R}^d \) and \( \alpha \in \mathbb{R}^d \). Consider the initial-value problem

\[
\begin{align*}
(IVP_1) \quad & y'(x) = f(x, y(x)) , \quad x \in I , \\
(IVP_2) \quad & y(a) = \alpha .
\end{align*}
\]

For a given equidistant partition \( x_j := a + jh, h := (b - a)/N, N \in \mathbb{N} \), of \( I \) and given real numbers \( \alpha_k, \beta_k, 0 \leq k \leq m, m \in \mathbb{N} \), with \( \alpha_m \neq 0 \)
and vectors \( \alpha^{(j)} \in \mathbb{R}^d \), \( 0 \leq j \leq m - 1 \), a linear multi-step method is of the form

\[
(MSM_1) \quad \frac{1}{h} \sum_{k=0}^{m} \alpha_k \, y_{j+k} = \sum_{k=0}^{m} \beta_k \, f(x_{j+k}, y_{j+k}) \quad , \quad 0 \leq j \leq N - m ,
\]

\[
(MSM_2) \quad y_j = \alpha^{(j)} \quad , \quad 0 \leq j \leq m - 1 .
\]

Denoting by \( \tau_h \) the local discretization error, suppose that

\[
(\ast) \quad \tau_h(x_j) = O(h^p) \quad , \quad 0 \leq j \leq m - 1 .
\]

Show that the linear multi-step method \((MSM)_1, (MSM)_2\) is consistent with the initial-value problem \((IVP)_1, (IVP)_2\) of order \( p \) if and only if it is consistent of order \( p \) with

\[
(IPV'_1) \quad y'(x) = y(x) \quad , \quad x \in I ,
\]

\[
(IPV'_2) \quad y(a) = 1 .
\]

**Exercise 11** (Stability of multi-step methods)

In class, the stability of multi-step methods has been investigated by means of the Lipschitz stability of operators on the linear space of grid functions. This approach can be generalized as follows:

Let \((X, \| \cdot \|_X)\) and \((Y, \| \cdot \|_Y)\) be normed spaces and let \((A_h)_h\) be a family of continuous operators \( A_h : X \rightarrow Y \). This family is called Lipschitz stable, if there exist positive numbers \( \delta, \eta \) such that for all \( x, z \in X \) and all \( h \) satisfying

\[
(LS)_1 \quad \| A_h x - A_h z \|_Y \leq \delta
\]

there holds

\[
(LS)_2 \quad \| x - z \|_X \leq \eta \| A_h x - A_h z \|_Y .
\]

(i) For Lipschitz stable families of continuous linear operators \( A_h : X \rightarrow Y \) derive a simpler definition of stability by taking advantage of the linearity.

(ii) Let \((A_h)_h\) be a Lipschitz stable family of continuous linear operators which are additionally assumed to be surjective. Show that for all \( h \) the operator \( A_h \) is invertible with \( \| A_h^{-1} \| \leq C, C > 0 \), and derive an upper bound for the constant \( C \).

(Hint: Show injectivity first and then use the definition of the operator norm \( \| A_h \| = \sup \limits_{0 \neq x \in X} \frac{\| A_h x \|_Y}{\| x \|_X} \).)

(iii) Let \( H \) be a Hilbert space, i.e., a complete inner product space with inner product \( < \cdot, \cdot > : H \times H \rightarrow \mathbb{R} \), and let \((A_h)_h\) be a family of
continuous endomorphisms $A_h : H \rightarrow H$. Further, let $a_h : H \times H \rightarrow \mathbb{R}$ be the bilinear form given by $a_h(x, y) := \langle A_h x, y \rangle$.

Prove the equivalence of the following two statements:

(*) For all $h$ the operator $A_h$ is invertible such that $\|A_h^{-1}\| \leq C$, where $C$ is a positive constant independent of $h$.

(**) There holds

$$\sup_{\|y\|=1} |a_h(x, y)| > \frac{1}{C} \|x\|, \quad x \in H,$$

$$\sup_{\|x\|=1} |a_h(x, y)| > 0, \quad 0 \neq y \in H.$$

(Hint: For $(*) \implies (**)$ use $\|x\| = \sup\{<x, y> | \|y\| = 1\}, x \in H$. For $(**) \implies (*)$ verify first that $\|A_h x\| \geq C^{-1} \|x\|, x \in H$. Then, using an arbitrary Cauchy sequence, show that the range of each operator $A_h$ is closed. Finally, prove surjectivity by contradiction, i.e., assume range$(A_h)^\perp \neq \{0\}$.)

**Exercise 12** (Affine-Invariance of Runge-Kutta methods)

Consider the initial-value problem

$$y'(x) = f(x, y(x)), \quad y(a) = \alpha,$$

where $f : \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m, a \in \mathbb{R}$ and $\alpha \in \mathbb{R}^m$. After a transformation $\hat{y} = Ay$ with a regular matrix $A \in \mathbb{R}^{m \times m}$ the initial-value takes the form

$$\hat{y}'(x) = A f(x, A^{-1} \hat{y}(x)), \quad \hat{y}(a) = A \alpha.$$

Show that any Runge-Kutta method inherits this property, i.e., it is affine invariant in the following sense: If the application of the Runge Kutta method to $(*)$ results in a grid function $y_k, k \in \mathbb{N}_0$, then the application of the method to $(**)$ results in the grid function $Ay_k, k \in \mathbb{N}_0$.

**Exercise 13** (Maximal Consistency Order of s-Stage Runge Kutta Methods)

Show that the application of an $s$-stage Runge Kutta method to the initial-value problem

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

yields increment functions of the form

$$k_j(x_k, y_k) = p_j(\lambda h) y_k, \quad 1 \leq j \leq s,$$
where \( p_j \) are polynomials of degree at most \( j \). Use this result to determine the maximal consistency order of \( s \)-stage Runge Kutta methods.

**Exercise 14 (Cauchy’s \( \theta \)-Method)**

Consider the initial-value problem
\[
y'(x) = f(x, y(x)) \quad x \geq a \quad y(a) = \alpha ,
\]
where \( f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m \) and \( \alpha \in \mathbb{R}^m \).

Show that the only consistent linear one-step method for the numerical solution of (*) are given by the one-parameter family
\[
y_h(x_{k+1}) = y_h(x_k) + h \left((1 - \theta)f(x_k, y_h(x_k)) + \theta f(x_{k+1}, y_h(x_{k+1}))\right).
\]

(i) Which values of \( \theta \in \mathbb{R} \) provide the consistency order \( p = 2 \)?

(ii) Which methods result for \( \theta = 0, \theta = 0.5, \) and \( \theta = 1.0 \)?

**Exercise 15 (Runge Kutta Fehlberg 1(2) Method)**

Consider the initial-value problem
\[
y'(x) = f(x, y(x)) \quad y(a) = \alpha .
\]

For the numerical solution use Runge Kutta methods of order 1 and 2
\[
\begin{align*}
y_h(a + h) & = y_h(a) + h (b_1 k_1 + b_2 k_2) , \\
y(a + h) - y_h(a + h) & = O(h^2) \\
\hat{y}_h(a + h) & = \hat{y}_h(a) + h (\hat{b}_1 k_1 + \hat{b}_2 k_2 + \hat{b}_3 k_3) , \\
y(a + h) - \hat{y}_h(a + h) & = O(h^3)
\end{align*}
\]

with step size \( h \) and the increments
\[
k_1 = f(a, \alpha) , \\
k_i = f(a + c_i h, \alpha + h \sum_{j=1}^{i-1} a_{ij} k_j) , \quad 2 \leq i \leq 3 .
\]

(i) Derive order conditions for the coefficients \( c_i, a_{ij}, b_i \) and \( \hat{b}_i \). Observe that here \( c_i = \sum_j a_{ij} \) is not requested.

(ii) How many function evaluations per step are required by (2.156),(2.157) and (2.158), (2.159)?

(iii) The embedded method uses Fehlberg’s trick in the sense that \( y_h \) is used in the following step: One takes advantage in the choice of the parameters to use the increment \( k_3 \) as the increment \( k_1 \) in the following
step which results in a reduction of the computational work. Derive conditions for the coefficients $c_i, a_{31},$ and $a_{32}$ such that the applicability of Fehlberg’s trick is guaranteed.

(iv) Consider the local discretization error of the method (2.158),(2.159) including Fehlberg’s trick from (iii) and the simplification $\hat{b}_2 = b_2.$ Compute the coefficient of the lowest power in $h.$ What are the consequences of the choice $c_2 = \frac{1}{2}, b_2 = 1$ taking into account that embedded methods are used for step size control?

(v) The potential advantages of RKF-1(2) methods are not completely exhausted by Fehlberg’s trick: Compute the parameters such that a Runge Kutta Fehlberg method of order 1 is obtained which only requires one function evaluation per step.

(vi) Sketch the method from (v) in algorithmic notation with the following step size control: an integration step has to be repeated with the computed smaller step size, if the estimated error $EST$ exceeds a user specific tolerance $TOL.$ Otherwise, the following integration step is performed with the suggested step size. In order to prevent large oscillations in the step size use

$$h_{\text{new}} := h \min(fac_{\text{max}}, \max(fac_{\text{min}}, fac(TOL/EST)^{1/(p+1)}))$$

with a safety factor $fac := 0.9$ and the limiters

$$fac_{\text{min}} := 0.5, \quad fac_{\text{max}} := 2.0.$$

**Exercise 16 (BDF)**

BDF rely on the idea to compute the approximation $y_{j+m}$ in $x_{j+m}$ as the value of the interpolating polynomial $p \in P_m$ with

$$p(x_\ell) = y_\ell, \quad j \leq \ell \leq j + m - 1,$$

$$p'(x_{j+m}) = f(x_{j+m}, y_{j+m}),$$
i.e., the polynomial interpolates the previously computed approximations in $x_j, ..., x_{j+m-1}$ and solves the differential equation in $x_{j+m}.$

(i) Show that for constant step size $h := x_{k+1} - x_k$ BDF is a linear $m$-step method of the form

$$\frac{1}{h} \sum_{k=0}^{m} \alpha_k y_{j+k} = \sum_{k=0}^{m} \beta_k f(x_{j+k}, y_{j+k}), \quad 0 \leq j \leq N.$$

Compute the coefficients $\alpha_k, \beta_k$ from the Lagrange fundamental polynomials $L_i \in P_m$ with $L_i(x_{j+k}) = \delta_{ik}, \quad 0 \leq i, k \leq m.$
(ii) Derive a representation of BDF on the basis of Newton’s representation of the interpolating polynomials. In particular, show: If
\[ \nabla^0 f_j := f_j , \quad 0 \leq j \leq m , \]
\[ \nabla^p f_j := \nabla^{p-1} f_j - \nabla^{p-1} f_{j-1} , \quad p \leq j \leq m , \quad 1 \leq p \leq m \]
are the backward differences with respect to the nodes \((x_j, y_j)\), \(0 \leq j \leq m\), where \(x_j = x_0 + jh\), \(0 \leq j \leq m\), then the interpolating polynomial \(p_q \in P_q\) with \(p_q(x_j) = f_j\), \(m - q \leq j \leq m\), is given by
\[ p_q(x) = \sum_{j=0}^{q} (-1)^j \binom{q}{j} \nabla^j f_m . \]
Here \(s := (x - x_m)/h\).

(iii) In addition to part (ii) show that implicit BDF with constant step size \(h\) have the equivalent representation
\[ \sum_{k=1}^{m} \frac{1}{k} y_{j+m} = f(x_{j+m}, y_{j+m}) . \]

(iv) For an equidistant grid of step size \(h\) give an explicit representation of BDF with \(m = 3\).

(v) Using the representation in (i) and
\[ \tau_h(x) = \frac{1}{h} \sum_{k=0}^{m} \alpha_k y(x - (m - k) h) - \sum_{k=0}^{m} \beta_k y'(x - (m - k) h) , \]
compute the local discretization error.

(vi) Compute the order of consistency.

**Exercise 17** (Stability of BDF)

Dahlquist has shown that there are no A-stable linear multi-step methods of order \(> 2\). On the other hand, an \(m\)-step BDF has the order of consistency \(m\). Hence, for \(m \geq 3\) BDF can not be A-stable. In this exercise we prove this fact based on the definition of stability.

(i) Derive the formulas for BDF with \(m = 2\) and \(m = 3\) in the standard representation of linear multi-step methods.

(Hint: Use the representation from Exercise 16 (i).)

(ii) Show that BDF with \(m = 2\) is A-stable, but not the one with \(m = 3\).

(Hint: Use Dahlquist’s root condition.)
Exercise 18 (Adams-Bashforth Method)

Consider the scalar initial-value problem
\[ y'(x) = f(x, y(x)) \quad , \quad y(a) = \alpha \ . \]

Show that the Adams-Bashforth m-step method has the representation
\[ y_{j+m} = y_{j+m-1} + \int_{x_{j+m-1}}^{x_{j+m}} p(x) \, dx \ , \]
where \( p \in \mathcal{P}_{m-1} \) is the polynomial with \( p(x_k) = f(x_k, y_k), j \leq k \leq j+m-1 \). In the sequel assume an equidistant grid of step size \( h > 0 \).

(i) Prove the stability of the m-step Adams-Bashforth method.

(ii) Compute the consistency of the m-step Adams-Bashforth method.

(iii) Using part (ii) of Exercise 16 prove the following representation of the m-step Adams-Bashforth method on the basis of backward differences
\[ y_{j+m} = y_{j+m-1} + h \sum_{k=0}^{m-1} \gamma_k \nabla^k f_{j+m-1} \quad , \quad 0 \leq j \leq N - m \ , \]
where
\[ \gamma_k = (-1)^k \int_0^1 (-s^k) \, ds \ . \]

(iv) Prove the following property of the coefficients \( \gamma_k \) of part (iii)
\[ \sum_{\ell=0}^{k} \frac{1}{k - \ell + 1} \gamma_\ell = 1 \ . \]

(Hint: Consider a power series with coefficients \( \gamma_k \) and use the Cauchy product of power series.)

Exercise 19 (Stability of semi-implicit methods)

Solve the stiff, autonomous, scalar initial value problem
\[ y'(x) = f(y(x)) \quad , \quad y(a) = \alpha \]
by means of a semi-implicit numerical integrator using a single Newton step for the nonlinear equation which determines the approximation \( y_{k+1} \). As initial guess choose the approximation \( y_k \) known from the previous step.
(i) A simplified version of the semi-implicit Euler method is given by the difference equation

\[ y_{k+1} = y_k + (1 - h a)^{-1} h f(y_k), \quad k \in \mathbb{N}_0 \]

with a suitable approximation \( a \) of the derivative \( f_y \). Apply the method to the test problem

\[ (* \quad y' = \lambda y, \quad y(0) = 1 \quad (\lambda \in \mathbb{C}, \ Re\lambda < 0) \]

and set \( z := \lambda h, \ z_0 := ah \), where \( z_0 \) is supposed to have a negative real. Prove the A-stability of the method for \( z = z_0 \).

(ii) For \( z \neq z_0 \), compute the boundary \( \partial G \) of the stability region

\[ G(z_0) := \{ z \in \mathbb{C} \mid |R(z, z_0)| \leq 1 \} \]

(\( R \) denotes the stability function). For \( z_0 = -1 + i \), display \( \partial G \) in the complex plane.

(iii) Apply the semi-implicit midpoint rule (without initial and final step)

\[ y_{k+1} = (1 - h a)^{-1} ((1 + h a) y_{k-1} + 2 h f(y_k)) \]

to the scalar test equation \( (*) \) and use \( z, z_0 \) as in part (i). Verify the A-stability for \( z = z_0 \).

(iv) For \( z \neq z_0 \), use the ansatz \( y_k = \xi^k \) to derive the associated characteristic polynomial. Denote by \( \xi_1(z, z_0) \) and \( \xi_2(z, z_0) \) the zeroes of this polynomial. Determine the boundary \( \partial G \) of the stability region

\[ G(z_0) := \{ z \in \mathbb{C} \mid |\xi_i(z, z_0)| \leq 1, \ 1 \leq i \leq 2 \} \]

and study the special case \( z_0 \in \mathbb{R} \).

**Exercise 20 (Method of lines)**

The spatial and temporal distribution of the temperature \( u(x, t) \) in a homogeneous slab of length 1m the endpoints of which are kept at the constant temperature \( 0^\circ C \) can be described by the heat equation

\[ \frac{\partial u}{\partial t}(x, t) = \frac{\partial^2 u}{\partial x^2}(x, t), \quad (x, t) \in \mathbb{R}_+ \times (0, 1), \]

\[ u(x, 0) = u_0(x), \quad x \in (0, 1), \]

\[ u(0, t) = u(1, t) = 0, \quad t \in \mathbb{R}_+, \]

where \( u_0(0) = u_0(1) = 0 \).

Discretize in space by means of symmetric difference quotients with respect to an equidistant grid \( (x_i)_{i=0}^n, \ x_i := ih, \ 0 \leq i \leq n, \ h := 1/n: \)

\[ \frac{\partial^2 u}{\partial x^2}(x_i, t) \approx \frac{u(x_i + h, t) - 2u(x_i, t) + u(x_i - h, t)}{h^2}. \]
This leads to a system of first order ordinary differential equations for
the time-dependent grid function \((u_i(t))_{i=1}^{n-1}\) with \(u_i(t) \approx u(x_i, t)\), \(1 \leq i \leq n - 1\), where \(u_0 = u_n = 0\). The initial values are given by \(u_i(0) := u_0(x_i)\).

(i) Derive the resulting initial value problem for a linear autonomous
system of the form

\[ u'(t) = Au(t) , \]

where \(A \in \mathbb{R}^{(n-1) \times (n-1)}\) and \(u(t) := (u_i(t))_{i=1}^{n-1} \in \mathbb{R}^{n-1}\).

(ii) Show that the grid functions \(e_k := (\sin h \pi ik)_{i=1}^{n-1}, 1 \leq k \leq n - 1\),
form an orthonormal system of eigenvectors of \(A\) from part (i).

(iii) Why is the system of ordinary differential equations a stiff system?

(iv) Use an algorithmic notation to formulate an efficient procedure
for the simulation of the temperature distribution in the slab where
the granularity \(h\) of the discretization in space and the step size \(\Delta t\) in
time are the input parameters.

**Exercise 21** (*Stability of extrapolation methods*)

Extrapolation techniques can also be used for stiff methods to construct
numerical integrators of higher order. However, the stability property
of the basis scheme are not automatically inherited by the resulting
method.

(i) Prove the A-stability of the implicit trapezoidal rule

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

(ii) It is well-known that the implicit trapezoidal rule admits an \(h^2\)-
expansion of the global discretization error. Extrapolate by means of
\(n_1 = 1\) and \(n_2 = 2\) (number of micro steps) and show that A-stability
gets lost for the resulting scheme.

(iii) Now, extrapolate with \(n_1 = 2, n_2 = 4\) and show that the applica-
tion to the scalar test equation \(y' = \lambda y, \lambda \in \mathbb{C}\),
gives rise to a stability function \(R_{22}\) of the resulting scheme with the property

\[ \lim_{z \to -\infty} |R_{22}(z)| = 1 , \quad z := \lambda h . \]

(iv) Prove that the scheme from part (iii) is not A-stable no matter
how \(n_1, n_2\) are chosen.
Exercise 22 (Mechanical two-body problem)

Differential-Algebraic Equations (DAE) naturally occur in the mathematical modeling and numerical simulation of mechanical systems. This will be illustrated using the example of a double pendulum:

The double pendulum consists of two point masses \( m_1 \) and \( m_2 \) joined by two strings of lengths \( \ell_1 \) and \( \ell_2 \) (cf. the figure below). Under an external excitement, the masses start to oscillate in a plane. As shown in the figure below, a Cartesian coordinate system will be used. The time-dependent locations of the masses will be described by the vector \( \mathbf{p}(t) = (x_1(t), y_1(t), x_2(t), y_2(t))^T \).

According to Hamilton’s principle, the dynamics of the system in the time interval \([0, T]\) results from the minimization of the functional

\[
\int_0^T L(\mathbf{p}(t), \dot{\mathbf{p}}(t)) \, dt
\]

under the geometrical constraints

\[
x_i^2 + y_i^2 = \ell_i^2, \quad 1 \leq i \leq 2.
\]

Here, the Lagrangian \( L \) is given by

\[
L(\mathbf{p}, \mathbf{q}) := E_{\text{kin}}(\mathbf{p}, \mathbf{q}) + E_{\text{pot}}(\mathbf{p}, \mathbf{q})
\]

as the sum of the kinetic energy

\[
E_{\text{kin}}(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \mathbf{q}^T M \mathbf{q}, \quad M := \text{diag}(m_1, m_2, m_1, m_2)
\]
and the potential energy
\[ E_{\text{pot}}(p, q) := -g(m_1 y_1 + m_2 y_2), \]
where \( g \) denotes the gravitational constant.

(i) Derive the saddle point problem which results by coupling the constraints using a Lagrangian multiplier \( \lambda \in \mathbb{R}^2 \).

(ii) Derive the necessary optimality conditions in case of a sufficiently smooth trajectory \( p \) (Euler-Lagrange equations).

(iii) Rewrite the Euler-Lagrange equations for the double pendulum and the geometrical constraints as a differential-algebraic system.

(iv) Compute the index of the DAE from part (iii).

**Exercise 23 (Laplace operator in polar coordinates)**
Consider the Laplace operator
\[ \Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} \]
and show that in polar coordinates \((r, \varphi)\) it is given by
\[ \Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r} + \frac{1}{r^2} \frac{\partial^2}{\partial \varphi^2}. \]

**Exercise 24 (Solution of an elliptic boundary value problem)**
Let \( \Omega \subset \mathbb{R}^2 \) be the domain given by
\[ \Omega := \{(r, \varphi) \mid 0 \leq r < 1, \ 0 < \varphi < \frac{7}{4}\pi\} \]
with boundary \( \Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \), where
\[ \Gamma_1 := \{(r, 0) \mid 0 < r < 1\}, \]
\[ \Gamma_2 := \{(r, \frac{7}{4}\pi) \mid 0 < r < 1\}, \]
\[ \Gamma_3 := \{(1, \varphi) \mid 0 < \varphi < \frac{7}{4}\pi\}. \]
Show that the solution of the elliptic boundary value problem
\[ -\Delta u = 0 \quad \text{in} \ \Omega, \]
\[ u = 0 \quad \text{on} \ \Gamma_1 \cup \Gamma_2, \]
\[ u = \sin\left(\frac{4}{7}\pi\right) \quad \text{on} \ \Gamma_3 \]
is given by \( u(r, \varphi) = r^{4/7} \sin\left(\frac{4}{7}\pi\right) \).
Exercise 25 (Fundamental solution)
Let $B_d^1$ be the unit ball in $\mathbb{R}^d$, $d \geq 2$, and let $|B_d^1|$ be its volume. Show that the function
\[
\Gamma(x, y) := \begin{cases} 
\frac{1}{2\pi} \ln(|x - y|), & d = 2 \\
\frac{1}{d(2-d)|B_d^1|} |x - y|^{2-d}, & d > 2
\end{cases}
\]
considered as a function of $x$, is a harmonic function in $\mathbb{R}^d \setminus \{y\}$.

Exercise 26 (Nine point approximation of the Laplace operator)
For a sufficiently smooth function $u$ show that the nine point formula
\[
\frac{1}{h^2} \left( \frac{8}{3} u(x, y) - \frac{1}{3} (u(x, y - h) + u(x - h, y) + u(x + h, y) + u(x, y + h)) \\
- \frac{1}{3} (u(x - h, y - h) + u(x + h, y - h) + u(x - h, y + h) + u(x + h, y + h)) \right)
\]
provides an approximation of $-\Delta u$ of order $O(h^2)$.

Exercise 27 (Checkerboard ordering of grid points)
Let $\Omega = (a, b)^2$, $a, b \in \mathbb{R}, a < b$, and let $\tilde{\Omega}_h$ be a uniform grid of step size $h = (b - a)/(N + 1)$, $N \in \mathbb{N}$. For the numerical solution of the Poisson equation with homogeneous Dirichlet boundary conditions
\[
-\Delta u = f \quad \text{in } \Omega, \\
u = 0 \quad \text{on } \Gamma
\]
use the five point approximation of the Laplace operator. Determine the block structure of the coefficient matrix $A_h$ of the resulting linear algebraic system
\[
A_h u_h = b_h
\]
for a checkerboard ordering of the grid points.

Exercise 28 (Eigenvalues and eigenfunctions of the discrete Laplace operator)
Let $\Omega = (a, b)^2$, $a, b \in \mathbb{R}, a < b$, and let $\tilde{\Omega}_h$ be a uniform grid with step size $h = (b - a)/(N + 1)$, $N \in \mathbb{N}$. For the numerical solution of the Poisson equation with homogeneous Dirichlet boundary conditions
\[
-\Delta u = f \quad \text{in } \Omega, \\
u = 0 \quad \text{on } \Gamma
\]
use the five point approximation of the Laplace operator. Show that the coefficient matrix $A_h$ of the resulting linear algebraic system
\[
A_h u_h = b_h
\]
has the eigenvalues
\[ \lambda_{ij} = 4h^2 - 2\sin^2\left(\frac{i\pi h}{2}\right) + 2\sin^2\left(\frac{j\pi h}{2}\right), \quad 1 \leq i, j \leq N, \]
and the associated eigenfunctions \( x^{(ij)} \in \mathbb{R}^{N^2} \) according to
\[ x^{(ij)}_{k\ell} = \frac{h}{2} \sin(ikh\pi) \sin(jh\ell\pi), \quad 1 \leq k, \ell \leq N. \]
Compute the spectral condition \( A_h \). How does the spectral condition behave \( N \to \infty \)?

**Exercise 29 (Non-smooth solution)**
Let \( \Omega \subset \mathbb{R}^2 \) be the domain given by
\[ \Omega := \{(r, \varphi) \mid 0 \leq r < 1, \; 0 < \varphi < \frac{7}{4}\pi\} \]
with boundary \( \Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \), where
\begin{align*}
\Gamma_1 &:= \{(r, 0) \mid 0 < r < 1\}, \\
\Gamma_2 &:= \{(r, \frac{7}{4}\pi) \mid 0 < r < 1\}, \\
\Gamma_3 &:= \{(1, \varphi) \mid 0 < \varphi < \frac{7}{4}\pi\}.
\end{align*}
The solution of the elliptic boundary value problem
\begin{align*}
-\Delta u &= 0 \quad \text{in} \; \Omega, \\
u &= 0 \quad \text{on} \; \Gamma_1 \cup \Gamma_2, \\
u &= \sin\left(\frac{4}{7}\pi\right) \quad \text{on} \; \Gamma_3
\end{align*}
is given by \( u(r, \varphi) = r^{4/7} \sin\left(\frac{4}{7}\pi\right) \).
Show that \( u \in H^1(\Omega) \), but \( u \notin H^2(\Omega) \).

**Exercise 30 (Poincaré-Friedrichs inequality)**
Let \( \Omega \subset \mathbb{R}^d \) be a bounded domain with boundary \( \Gamma = \partial \Omega \). For \( u \in H^1_0(\Omega) \) the Poincaré-Friedrichs inequality
\[ \int_\Omega |u|^2 \, dx \leq C_\Omega \int_\Omega |\nabla u|^2 \, dx \]
holds true with a constant \( C_\Omega > 0 \). Show that \( C_\Omega \geq 1/\lambda_1 \), where \( \lambda_1 \) is the smallest eigenvalue of the eigenvalue problem
\[ -\Delta u = \lambda u \quad \text{in} \; \Omega, \quad u = 0 \quad \text{on} \; \Gamma. \]
**Exercise 31 (Robin boundary condition)**

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with boundary $\Gamma = \partial \Omega$, $f \in L^2(\Omega)$, $g \in L^2(\Gamma)$, and $r \in L^\infty(\Gamma)$ with $r(x) \geq r_0 > 0$ for almost all $x \in \Gamma$. Consider the Poisson equation with Robin boundary conditions

\[ -\Delta u = f \quad \text{in } \Omega, \]
\[ \nu \cdot \nabla u + ru = g \quad \text{on } \Gamma, \]

where $\nu$ is the outward unit normal vector on $\Gamma$. Show that the weak solution of the boundary value problem satisfies the variational equation

\[ \int_{\Omega} \nabla u \cdot \nabla v \, dx + \int_{\Gamma} rv \, ds = \int_{\Omega} fv \, dx + \int_{\Gamma} gv \, ds, \quad v \in H^1(\Omega), \]

and prove existence and uniqueness of a weak solution.

**Exercise 32 (Representation of nodal basis functions)**

Let $T$ be a non-degenerate triangle in $\mathbb{R}^2$ with vertices $a_i$ and edges $E_i, 1 \leq i \leq 3$, where the edge $E_i$ is opposite to the vertex $a_i$. Further, let $m_{E_i}$ and $\nu_{E_i}$ be the midpoints and outward unit normal vectors of the edges. Show that the nodal basis functions $\varphi_i$ of the finite element space of simplicial Lagrangean finite elements of type (1) have the representation

\[ \varphi_i(x) = -\frac{|E_i|}{2|T|} \nu_{E_i} \cdot (x - m_{E_i}), \quad 1 \leq i \leq 3, \]

where $|E_i|$ is the length of $E_i$ and $|T|$ is the area of $T$.

**Exercise 33 (Representation of the local stiffness matrix)**

Let $T$ be a non-degenerate triangle in $\mathbb{R}^2$ of area $|T|$ with the vertices $a_i, 1 \leq i \leq 3$, and the edges $E_i$ of length $|E_i|$ opposite to $a_i$. Further, let $\alpha_{ij}, 1 \leq i < j \leq 3$, be the interior angles formed by the edges $E_i$ and $E_j$ and let $\varphi_i, 1 \leq i \leq 3$, be the nodal basis functions of the finite element space of simplicial Lagrangean finite elements of type (1). Show that the local stiffness matrix $S \in \mathbb{R}^{3 \times 3}$ with the elements $S_{ij} = \int_T \nabla \varphi_i \cdot \nabla \varphi_j \, dx, \quad 1 \leq i, j \leq 3$, is given by

\[ S = \frac{1}{4|T|} \begin{pmatrix}
|E_1|^2 & -\cot(\alpha_{12}) & -\cot(\alpha_{13}) \\
-\cot(\alpha_{12}) & |E_2|^2 & -\cot(\alpha_{23}) \\
-\cot(\alpha_{13}) & -\cot(\alpha_{23}) & |E_3|^2
\end{pmatrix}. \]
**Exercise 34** (Quadratic Lagrangean finite elements)

For the numerical solution of the elliptic boundary value problem

\[-\Delta u(x) = 0 \quad x = (x_1, x_2)^T \in \Omega := (0, 1)^2,\]

\[u(x) = x_1^2 + x_2^2 \quad \text{on } \Gamma := \partial \Omega\]

use a finite element discretization by simplicial Lagrangean finite elements of type (2) with respect to a uniform simplicial triangulation \(\mathcal{T}_h(\Omega)\) which is generated as follows: Use a partition of \(\Omega\) into a quadratic grid of step size \(h = 1/(N+1), N \in \mathbb{N}\), and subdivide each cell into two triangles by the diagonal from bottom left to top right. Use the canonical nodal basis functions with respect to a lexicographic ordering of the grid points.

(i) For the reference triangle \(\hat{T}\) with vertices \(a_1 = (0,0), a_2 = (1,0), \) and \(a_3 = (0,1)\) compute the nodal basis functions \(\varphi_{a_i}, 1 \leq i \leq 3\), and \(\varphi_{a_{ij}}, a_{ij} := (a_i + a_j)/2, 1 \leq i < j \leq 3\), as functions of \(x = (x_1, x_2)\) and compute their gradients.

(ii) For an actual triangle \(T \in \mathcal{T}_h(\Omega)\) of area \(|T|\) compute the local stiffness matrix using the quadrature formula

\[\int_T f(x) \, dx \approx \frac{1}{3} |T| \sum_{1 \leq i < j \leq 3} f(a_{ij}).\]

How big is the quadrature error?

(iii) For \(h = 1/3\) sketch the triangulation, mark the nodal points by \(\times\), and order them according to the lexicographic ordering.

(iv) For \(h = 1/3\) sketch the sparsity pattern of the global stiffness matrix.

**Exercise 35** (Method of alternating directions I)

For the numerical solution of the heat equation

\[\frac{\partial u}{\partial t} - \Delta u = f\]

in \(Q := \Omega \times (0,T), \Omega = (a,b)^2 \subset \mathbb{R}^2\), use uniform partitions \(\{t_m = m \tau, 0 \leq m \leq M\}\) of \([0,T]\) of step size \(\tau = T/M, M \in \mathbb{N}\), and \(\{(x_i, y_j) \mid x_i = a + ih, y_j = a + jh, 1 \leq i, j \leq N + 1\}\) of \(\Omega\) of step size \(h = (b-a)/(N+1), N \in \mathbb{N}\). Let \(U_{ij}^m, 0 \leq m \leq M, \) and \(U_{ij}^{m+1/2}, 0 \leq m \leq M - 1\), be approximations of \(u(x_i, y_j, t_m)\) resp. \(u(x_i, y_j, t_m + \tau/2)\). Show that for sufficiently smooth \(u\) the method
of alternating directions
\[
\frac{U_{ij}^{m+1/2} - U_{ij}^m}{\tau/2} - \frac{U_{i-1,j}^{m+1/2} - 2U_{ij}^{m+1/2} + U_{i+1,j}^{m+1/2}}{h^2} = f_{ij}^{m+1/2},
\]
\[
\frac{U_{ij}^{m+1} - U_{ij}^{m+1/2}}{\tau/2} - \frac{U_{i-1,j}^{m+1} - 2U_{ij}^{m+1} + U_{i+1,j}^{m+1}}{h^2} = f_{ij}^{m+1/2},
\]
has the consistency order 2 in space and time.

What is the advantage of the method with regard to the numerical solution of the resulting linear algebraic system?

**Exercise 36 (Method of alternating directions II)**

Consider the heat equation from Exercise 35 with the boundary condition
\[
u = u_D \quad \text{on } \Sigma := \partial \Omega \times (0, T]
\]
and the initial condition
\[
u(\cdot, 0) = u_0 \quad \text{in } \Omega.
\]
Apply the method of alternating directions from Exercise 35 and set
\[
U_j^{m+1/2} := (U_{1j}^{m+1/2}, U_{2j}^{m+1/2}, \ldots, U_{Nj}^{m+1/2})^T, \quad 1 \leq j \leq N.
\]
Show that for each \(j\) one has to solve a linear algebraic system
\[
AU_j^{m+1/2} = b
\]
with a symmetric tridiagonal matrix \(A \in \mathbb{R}^{N \times N}\) and a vector \(b \in \mathbb{R}^N\).
Compute \(A\) and \(b\) explicitly.