2 Elliptic Differential Equations

2.1 Classical solutions

As far as existence and uniqueness results for classical solutions are concerned, we restrict ourselves to linear elliptic second order elliptic differential equations in a bounded domain $\Omega \subset \mathbb{R}^d$

\begin{equation}
Lu := - \sum_{i,j=1}^{d} a_{ij} u_{x_i x_j} + \sum_{i=1}^{d} b_i u_{x_i} + cu = f .
\end{equation}

We assume that the data of the problem satisfies the following conditions:

(2.1.2a) There holds $a_{ij}, b_i, c \in C(\overline{\Omega}), 1 \leq i, j \leq d$, $f \in C(\overline{\Omega})$ and there exists a constant $C > 0$, such that for all $x \in \Omega$

$$
|a_{ij}(x)|, \ |b_i(x)|, \ |c(x)| \leq C, \ 1 \leq i, j \leq d ,
$$

(2.1.2b) The functions $a_{ij}$ are symmetric, i.e., for all $1 \leq i, j \leq d$ and $x \in \Omega$ we have

$$
a_{ij}(x) = a_{ji}(x) ,
$$

and the matrix-valued function $(a_{ij})_{i,j=1}^{d}$ is uniformly positive definite in $\Omega$, i.e., there exists a constant $\alpha > 0$, such that for all $x \in \Omega$ 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 (2.1.1) we need to specify boundary conditions on the boundary $\Gamma = \partial \Omega$.

(2.1.3) **Definition:** Under the assumptions (2.1.2a),(2.1.2b) let $L$ be the linear second order elliptic differential operator given by (2.1.1), and assume $f \in C(\Omega)$ and $u^D \in C(\Gamma)$. Then, the boundary value problem

(2.1.4a) $Lu(x) = f(x) , \ x \in \Omega ,$

(2.1.4b) $u(x) = u^D(x) , \ x \in \Gamma$
is called an inhomogeneous Dirichlet problem. If \( u^D \equiv 0 \), (2.1.4a),(2.1.4b) 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 \( u \) satisfies (2.1.4a),(2.1.4b) pointwise.

In case \( c(x) \geq 0 \), \( x \in \Omega \), in (2.1.1), the uniqueness of a classical solution of the Dirichlet problem (2.1.4a),(2.1.4b) follows from the Hopf maximum principle.

(2.1.5) **Theorem:** In addition to the conditions (2.1.2a),(2.1.2b) assume \( c(x) \geq 0 \), \( x \in \Omega \). Further, assume that the function \( u \in C^2(\Omega) \) satisfies

\[
(2.1.6) \quad Lu(x) \leq 0 \quad x \in \Omega .
\]

Then, if there exists \( x_0 \in \Omega \) such that

\[
(2.1.7) \quad u(x_0) = \sup_{x \in \Omega} u(x) \geq 0 ,
\]

there holds

\[
(2.1.8) \quad u(x) = \text{const.} , \quad x \in \Omega .
\]

**Proof.** We refer to Theorem 2.1.2 in Jost (2002). \( \square \)

**Corollary:** Under the assumptions of Theorem (2.1.5) the Dirichlet problem (2.1.4a),(2.1.4b) admits at most one classical solution.

**Proof.** The proof is left as an exercise. \( \square \)

For the prototype of a linear second order elliptic differential equation, the Poisson equation (1.1.14), we can prove the existence of a solution of the the Dirichlet problem (2.1.4a),(2.1.4b) using Green’s representation formula. We assume \( \Omega \) to be a \( C^1 \) domain (see Definition (1.2.2)).

Green’s first and second formula are as follows:

(2.1.9) **Theorem:** Let \( \Omega \subset \mathbb{R}^d \) be a \( C^1 \) domain and assume \( u,v \in C^2(\overline{\Omega}) \). Further, let \( \nu \) be the exterior unit normal vector on \( \Gamma \). Then, there hold the first Green’s formula

\[
(2.1.10) \quad \int_{\Omega} \Delta u \ v \ dx + \int_{\Omega} \nabla u \cdot \nabla v \ dx = \int_{\Gamma} \nu \cdot \nabla u \ v \ d\sigma
\]

and the second Green’s formula

\[
(2.1.11) \quad \int_{\Omega} \left( \Delta u \ v - u \ \Delta v \right) \ dx = \int_{\Gamma} \left( \nu \cdot \nabla u \ v - u \ \nu \cdot \nabla v \right) \ d\sigma .
\]
**Proof.** The proof of (2.1.10) follows from 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 the roles of $u$ and $v$ in (2.1.10) and subtracting the resulting equation from (2.1.10) implies (2.1.11).

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

$$
(2.1.13) \quad \Gamma(x,y) := \Gamma(|x-y|) := \begin{cases} \frac{1}{2\pi} \ln(|x-y|), & d = 2 \\ \frac{1}{d(2-d)|B_d^1|} |x-y|^{2-d}, & d > 2 \end{cases},
$$

where $x, y \in \mathbb{R}^d$, $x \neq y$, and $|B_d^1|$ denotes the volume of the unit ball in $\mathbb{R}^d$, is called the fundamental solution of the Laplace equation. As a function of $x$, the fundamental solution is a harmonic function in $\mathbb{R}^d \setminus \{y\}$.

The notion ‘fundamental solution’ is justified by Green’s representation formula:

(2.1.14) **Theorem:** Let $\Omega \subset \mathbb{R}^d$ be a $C^1$ domain and $\Gamma(\cdot, \cdot)$ the fundamental solution as given by (2.1.13). If $u \in C^2(\overline{\Omega})$, in $y \in \Omega$ there holds

$$
(2.1.15) \quad u(y) = \int_{\Omega} \Gamma(x,y) \Delta u(x) \, dx + \int_{\partial \Omega} \left( u(x) \nu \cdot \nabla_x \Gamma(x,y) - \Gamma(x,y) \nu \cdot \nabla u(x) \right) \, d\sigma(x),
$$

where $\nabla_x$ denotes the gradient with respect to the differentiation in the variable $x$.

**Proof.** The proof follows from Green’s second formula (2.1.11). For details we refer to Theorem 1.1.1 in Jost (2002). □

The application of (2.1.15) to a test function $\phi \in C_0^\infty(\Omega)$ results in

$$
\phi(y) = \int_{\Omega} \Gamma(x,y) \Delta \phi(x) \, dx, \quad y \in \Omega.
$$

If we denote by $\Delta_x$ the Laplace operator with respect to the variable $x$ and by $\delta_y$ Dirac’s delta function as a distribution with $\delta_y(\phi) = \phi(y)$, we can interpret $\Delta_x \Gamma(\cdot, y)$ according to

$$
\Delta_x \Gamma(\cdot, y) = \delta_y.
$$
Definition: A function $G(x, y), x, y \in \Omega, x \neq y,$ is called Green's function for $\Omega,$ if the following conditions are satisfied

$$G(x, y) = 0, \quad x \in \partial \Omega,$$

$$\text{The function } G(x, y) - \Gamma(x, y) \text{ is harmonic in } x \in \Omega.$$}

For the domains $\Omega \subset \mathbb{R}^d$ considered here, the existence of a Green's function is guaranteed. Under certain assumptions, it enables the explicit representation of the solution of the inhomogeneous Dirichlet problem for Poisson's equation (1.1.14).

Theorem: Under the assumptions of Theorem (2.1.14) let $G(x, y)$ be a Green's function for $\Omega.$ Then, the following holds true:

(i) If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is a classical solution of the inhomogeneous Dirichlet problem (2.1.4a), (2.1.4b) for Poisson's equation (1.1.14), for $y \in \Omega$ we have the representation

$$u(y) = \int_{\Omega} G(x, y) f(x) \, dx + \int_{\partial \Omega} u_D(x) \nu \cdot \nabla_x G(x, y) \, d\sigma(x).$$

(ii) If $f$ is a Hölder continuous function in $\Omega$ and $u^D \in C(\partial \Omega),$ then the function $u$ given by (2.1.19) is a classical solution of the inhomogeneous Dirichlet problem (2.1.4a), (2.1.4b) for Poisson's equation (1.1.14).

Proof. Assertion (i) follows readily from the application of Green's second formula (2.1.11) to the function $v(x) = G(x, y) - \Gamma(x, y).$ The proof of (ii) requires additional results concerning the regularity of solutions of elliptic differential equations. We refer to chapters 9.1 and 10.1 in Jost (2002).

2.2 Finite difference methods

We consider the boundary value problem (2.1.4a), (2.1.4b) under the assumptions (2.1.2a), (2.1.2b) and $c(x) > 0, x \in \Omega.$ For the numerical solution we use finite difference methods on the basis of an approximation of the partial derivatives in (2.1.4a) by difference quotients (cf. Chapter 3.5 in Stoer, Bulirsch (2002)). We begin by assuming that the domain $\Omega$ is a $d$-dimensional cube.

Definition: Let $\Omega := (a, b)^d, a, b \in \mathbb{R}, a < b, h := (b - a)/(N + 1), N \in \mathbb{N}.$ Then, the set

$$\Omega_h := \{(x_{i_1}, \cdots, x_{i_d})^T | x_{i_j} = a + i_j h, \ 0 \leq i_j \leq N + 1, \ 1 \leq j \leq d\}$$

is called a grid-point set with step size $h.$ It represents a uniform or equidistant grid. A grid is called uniform or equidistant, if the grid-points have the same distance from each other. We further refer to $\Omega_h := \overline{\Omega}_h \cap \Omega$ as the set of interior grid-points and to $\Gamma_h := \overline{\Omega}_h \setminus \Omega_h$ as the set of boundary grid-points.
For the approximation of the first and second partial derivatives $u_{x_i}, u_{x_ix_i}$ we consider the following difference quotients:

\begin{equation}
\text{(2.2.2) Definition: For } u : \Omega \to \mathbb{R} \text{ and } i \in \{1, \cdots, d\}, \text{ the difference quotients}
\end{equation}

\begin{align}
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)) ,
\end{align}

where $e_i$ is the $i$-th unit vector in $\mathbb{R}^d$, are called the forward, backward and central difference quotients for the first partial derivatives $u_{x_i}, 1 \leq i \leq d$. The difference quotients given by

\begin{equation}
D_{h,i}^2 u(x) := h^{-2}(u(x + he_i) - 2u(x) + u(x - he_i))
\end{equation}

are called central difference quotients for the second derivatives $u_{x_ix_i}$.

For $u \in C^2(\Omega)$, Taylor expansion shows that the difference quotients $D_{h,i}^+ u$ provide an approximation of $u_{x_i}$ of order $O(h)$, whereas for $u \in C^4(\Omega)$ the central difference quotients $D_{h,i}^1 u$ and $D_{h,i}^2 u$ result in an approximation of $u_{x_i}$ and $u_{x_ix_i}$ of order $O(h^2)$.

As far as the approximation of the mixed second partial derivatives $u_{x_ix_j}, 1 \leq i \neq j \leq d$, in $x \in \Omega_h$ is concerned, it is obvious to use a weighted sum of functions values of $u$ in points of the $(x_i, x_j)$-plane which are neighbors of $x$. This leads to

\begin{equation}
D_{h,i,j}^2 u(x) = \sum_{|\alpha|,|\beta| \leq 1} \kappa_{\alpha,\beta} u(x + \alpha he_i + \beta he_j),
\end{equation}

where $\kappa_{\alpha,\beta} \in \mathbb{R}, \alpha, \beta \in \mathbb{Z}$. Assuming sufficient smoothness of $u$ and requiring that these difference quotients approximate the mixed second partial derivatives of order $O(h^2)$, by Taylor expansion we obtain the conditions

\begin{align}
\kappa_{0,0} &= -(h^2)^{-1} \gamma_1 , \\
\kappa_{\alpha,\beta} &= (2h^2)^{-1}(\gamma_1 - \text{sgn}(\alpha) \gamma_2 - \text{sgn}(\beta) \gamma_3), \quad |\alpha| + |\beta| = 1 , \\
\kappa_{\alpha,\beta} &= (4h^2)^{-1}(\text{sgn}(\alpha) \beta - \gamma_1 + \text{sgn}(\alpha) \gamma_2 + \text{sgn}(\beta) \gamma_3), \quad |\alpha| = |\beta| = 1 ,
\end{align}

where $\gamma_i \in \mathbb{R}, 1 \leq i \leq 3$, are free parameters. Hence, there exist several possibilities for the approximation of the mixed second partial derivatives of order $O(h^2)$. We will discuss later which is the most suited one with regard to the finite difference approximation of (2.1.4a)-(2.1.4b).

An appropriate tool to characterize difference quotients are difference stars. (see Fig. 1). Here, the points displayed by $\bullet$ mark those points used in the difference quotient and the number next to a point refers to the corresponding weight.
We also use the following characterization

\[ \begin{bmatrix} \kappa_{-1,1} & \kappa_{0,1} & \kappa_{1,1} \\ \kappa_{-1,0} & \kappa_{0,0} & \kappa_{1,0} \\ \kappa_{-1,-1} & \kappa_{0,-1} & \kappa_{1,-1} \end{bmatrix} \]

The difference quotient given by (2.2.5) is also an example for a so-called compact difference approximation. A difference approximation

\[ \sum_{|\alpha| \leq m} \kappa_{\alpha} u(x + \alpha h), \quad \alpha \in \mathbb{Z}^d, \quad m \in \mathbb{N}, \]

and the associated difference star are called compact, if besides the grid point \( x \) only its next neighbors are used in the difference approximation, i.e., if \( m = 1 \).

An important special case of (2.1.4a) is Poisson's equation (cf. (1.1.14)), i.e., \( A = -\Delta \) where \( \Delta \) is the Laplace operator \( \Delta = \sum_{i=1}^{d} u_{x_i x_i} \). In view of the definition of the central difference quotients for the second partial derivatives \( u_{x_i x_i} \), we use the following difference approximation of the Laplace operator:

(2.2.8) Definition: Let \( D^2_{h, i, i} \) be given as in Definition (??). Then, the difference operator

\[ \Delta_h := \sum_{i=1}^{d} D^2_{h, i, i} \]

is called the discrete Laplace operator. In case \( u \in C^4(\Omega) \), \( \Delta_h u \) provides an approximation of \( \Delta u \) of order \( O(h^2) \).

For \( d = 2 \), the difference star associated with the difference operator \( -\Delta_h \) is displayed in Fig. 2. Since exactly five grid points are involved in the difference approximation, the difference star is referred to as the five-point difference star.

In order to derive compact difference approximations of \( -\Delta u \) in two space dimensions, as in (2.2.5) we consider
Fig. 2. Five-point and nine-point difference stars for the approximation of the differential operator $-\Delta$ in two space dimensions

$$\sum_{|\alpha|,|\beta|\leq 1} \kappa_{\alpha,\beta} u(x + \alpha h e_i + \beta h e_j) , \ \alpha, \beta \in \mathbb{Z}$$

and try to determine the weights $\kappa_{\alpha,\beta}$ such that we obtain a difference approximation of order $O(h^2)$ for sufficiently smooth $u$. This leads to the conditions

$$\begin{align*}
k_{1,1} &= k_{1,-1} = k_{-1,1} = k_{-1,-1} , \\
k_{1,0} &= k_{0,1} = k_{0,-1} = 4k_{-1,0} , \\
k_{1,1} + 4k_{1,0} + k_{0,0} &= 0 , \\
2k_{1,1} + k_{1,0} &= -1 .
\end{align*}$$

For $k_{1,1} = 0$ and $k_{1,0} = -1/h^2$ we get the five-point difference star. For $k_{1,1} = k_{1,0} = -1/(3h^2)$ we obtain the nine-point difference star displayed in Fig. 2 (right). However, it is not possible to obtain the order $O(h^3)$. On the other hand, using non-compact difference approximations one can realize approximations of higher order than 2.

In case of a more general bounded domain $\Omega \subset \mathbb{R}^d$ with boundary $\Gamma$, the definition of the set of interior grid points has to be renewed.

(2.2.9) **Definition:** Let $\Omega \subset \mathbb{R}^d$ be a bounded, simply connected domain with boundary $\Gamma$ and $\mathbb{R}^d_h, h > 0$ the grid-point set given by

$$\mathbb{R}^d_h := \{x = (x_1, \cdots, x_d)^T \mid x_i = ih , \ i \in \mathbb{Z}\} .$$

Then, the set $\Omega_h := \mathbb{R}^d_h \cap \Omega$ is called the set of interior grid points. A grid point $x \in \Gamma$ is called a boundary grid point, if there exists an interior grid point $x^* \in \Omega_h$ such that $x = x^* + \alpha he_i$ for some $i \in \{1, \cdots, d\}$ and $\alpha \in \mathbb{R}, |\alpha| < 1$. In this case, $x^*$ is said to be next to the boundary. All interior grid points that are not next to the boundary are called far from the boundary. We denote by $\Gamma_h, \Omega_h^\Gamma$ and $\Omega_h^\Omega$ the set of boundary grid points, the set of grid points next to the boundary, and the set of grid points far from the boundary, respectively (see Fig. 3). We set $\overline{\Omega}_h := \Omega_h \cup \Gamma_h$. 
The line between two neighboring grid points along a grid-line is called an edge, and the two grid points are called the vertices of the edge. We call a grid-point set discretely connected, if any two grid points in \( \Omega_h \) can be connected by a path consisting of a finite number of edges with vertices in \( \Omega_h \).

**Remark.** Observe that an interior grid point \( x \) can be next to the boundary although \( x \pm h e_i \in \Omega_h, 1 \leq i \leq d \).

The definition of difference quotients and of the discrete Laplace operator also has to be modified. This leads to the Shortley-Weller approximation.

**Definition:** Let \( \overline{\Omega}_h \subset \mathbb{R}^d \) be a grid-point set as in Definition (2.2.9) and \( x \in \Omega_h, x \pm h_i^\pm e_i \in \overline{\Omega}_h, h_i^\pm = \alpha_i^\pm h, |\alpha_i^\pm| \leq 1, 1 \leq i \leq d \). For the approximation of \( u_{x_i} \) in \( x \), the forward and backward difference quotients are given by

\[
D_{h,i}^+ u(x) := \left( u(x + h_i^+ e_i) - u(x) \right) / h_i^+ ,
\]
\[
D_{h,i}^- u(x) := \left( u(x) - u(x - h_i^- e_i) \right) / h_i^- .
\]

The central difference quotient for the second derivatives \( u_{x_i,x_i} \) is given by

\[
D_{h,i}^2 u(x) := \frac{2}{h_i^+ + h_i^-} \left( \frac{u(x + h_i^+ e_i) - u(x)}{h_i^+} + \frac{u(x - h_i^- e_i) - u(x)}{h_i^-} \right).
\]

For the mixed second derivatives \( u_{x_i,x_j} \), \( 1 \leq i \neq j \leq d \), we may proceed as in (2.2.5).

We define the discrete Laplace operator \( \Delta_h \) as in Definition (2.2.8). For \( d = 2 \) and \( x = (x_1, x_2) \in \Omega_h \) such that \( (x_1 \pm \alpha_1^\pm h, x_2 \pm \alpha_2^\pm h) \in \overline{\Omega}_h \), we obtain the explicit representation

**Fig. 3.** Grid point next to the boundary \( \bullet \) and boundary grid points \( \circ \) in a domain with curved boundary (left) and approximation of the normal derivative in a boundary grid point (right).
\[ \Delta_h u(x) := \frac{2}{h^2} \left( \frac{1}{\alpha_1^+ (\alpha_1^+ + \alpha_1^-)} u(x_1 + \alpha_1^+ h, x_2) + \frac{1}{\alpha_1^- (\alpha_1^- + \alpha_1^+)} u(x_1 - \alpha_1^- h, x_2) + \frac{1}{\alpha_2^+ (\alpha_2^+ + \alpha_2^-)} u(x_1, x_2 + \alpha_2^+ h) + \frac{1}{\alpha_2^- (\alpha_2^- + \alpha_2^+)} u(x_1, x_2 - \alpha_2^- h) - \frac{1}{\alpha_1^+ \alpha_1^- + \alpha_2^+ \alpha_2^-} u(x_1, x_2) \right). \]

**Remark.**

(i) In grid points next to the boundary, the central difference quotient for the second derivatives and the discrete Laplace operator in general only provide an approximation of \( u_{x_i x_i} \), and \( \Delta u \) of order \( O(h) \).

(ii) The definition of the difference quotients and the discrete Laplace operator in (2.2.10) can be easily generalized to non-equidistant grids.

For the discretization of the normal derivative \( u_\nu = \nu \cdot \nabla u \) in \( x_\Gamma \in \Gamma_h \), where \( \nu \) denotes the exterior normal unit vector in \( x_\nu \), let \( x_\Gamma^* \in \Omega_h \) be the first intersection in \( \Omega_h \) of the normal through \( x_\Gamma \) and a grid line (cf. Fig. 3 (right)). Let \( x_i, 1 \leq i \leq 2 \), be those grid points on that grid line next to \( x_\Gamma^* \), assuming that the grid is sufficiently fine such that \( x_i \in \Omega_h, 1 \leq i \leq 2 \). We refer to \( \tilde{u}(x_\Gamma^*) \) as the value obtained by linear interpolation with respect to \( u(x_1) \) and \( u(x_2) \). Then, an approximation of the normal derivative \( \nu \cdot \nabla u \) in \( x_\Gamma \) is given by

\[ \partial^h_\nu u(x_\Gamma) := \frac{u(x_\Gamma) - \tilde{u}(x_\Gamma^*)}{|x_\Gamma - x_\Gamma^*|}. \]

This approach can also be used for directional derivatives \( u_\mu = \mu \cdot \nabla u \) with respect to a non-tangential vector \( \mu \) in \( x_\Gamma \).

We compute approximations within the linear space of functions defined on a grid-point set.

**Definition:** Let \( \overline{\Omega}_h \) be a grid-point set as in (2.2.10). A function \( u_h : \overline{\Omega}_h \to \mathbb{R} \) is called a grid function. We refer to \( C(\overline{\Omega}_h) \) as the linear space of grid functions. The space \( C(\overline{\Omega}_h) \) will be equipped with the norm

\[ \| u_h \|_h := \| u_h \|_{\ell^\infty(\overline{\Omega}_h)} = \max_{x \in \overline{\Omega}_h} |u_h(x)|. \]

We first consider a finite difference method for the approximation of Poisson’s equation with inhomogeneous Dirichlet boundary conditions by means of the discrete Laplace operator. Given grid functions \( f_h \in C(\Omega_h) \) and \( u^D_h \in C(\Gamma_h) \), we are looking for \( u_h \in C(\overline{\Omega}_h) \) such that

\[ -\Delta_h u_h = f_h \quad \text{in} \ \Omega_h, \]

\[ u_h = u^D_h \quad \text{auf} \ \Gamma_h. \]

The existence and uniqueness of a solution of (2.2.14a),(2.2.14b) can be shown either using a discrete maximum principle, a discrete Green’s function (cf.,
e.g., Jost (2002) or by means of the investigation of the linear algebraic system (2.2.14a), (2.2.14b). Here, we will follow the latter approach. For simplicity, we will restrict ourselves to the case $d = 2$ and $\Omega = (a, b)^2$ and assume a uniform grid $\Omega_h$. The unknown values of the grid function $u_h$ in the interior grid points $x \in \Omega_h$ can be interpreted as the components of a vector. The structure of the resulting linear algebraic system depends on the ordering of the grid points.

(2.2.15) **Definition:** Assume $\Omega = (a, b)^2, a, b \in \mathbb{R}, a < b$, and let $\Omega_h$ be a uniform grid of step size $h := (b - a)/N, N \in \mathbb{N}$, with interior grid points $x = (x_i, x_j)^T, 1 \leq i, j \leq N$, where $x_i = a + ih$. An ordering of the 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 a lexicographic ordering. If we first count the grid points with even $i + j$ and then those with odd $i + j$, the ordering is said to be a checkerboard ordering.

If we order the interior grid points according to Definition (2.2.15), the values of the grid function in the interior grid points can be assigned a vector with $n_h := N^2$ components. For simplifying the notation, this vector will be denoted in the same way as the grid function, i.e., we define $u_h \in \mathbb{R}^{n_h}$ according to $u_h,i := u_h(x_i), 1 \leq i \leq n_h$. Then, (2.2.14a), (2.2.14b) corresponds to a linear algebraic system

(2.2.16) \[ A_h u_h = b_h \]

with the 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 function values of $f_h$ and $u_h^D$.

In case of a lexicographic ordering of the grid points, $A_h$ is the block tridiagonal matrix

(2.2.17) \[ A_h = h^{-2} \text{tridiag}(-I, T, -I), \]

where the $n_h$ diagonal blocks $T \in \mathbb{R}^{N \times N}$ are tridiagonal matrices of the form

(2.2.18) \[ T = \text{tridiag}(-1, 4, -1) \]

and the off-diagonal blocks represent the negative $N \times N$ unit matrix.

In view of (2.2.17) and (2.2.18) it is easily seen that $A_h$ is a symmetric positive definite matrix. with the eigenvalues

(2.2.19) \[ \lambda_{ij}(A_h) = 4h^{-2}\left(\sin^2(i\pi h/2) + \sin^2(j\pi h/2)\right), \quad 1 \leq i, j \leq N. \]

The components of the associated orthonormal eigenvectors $x^{(ij)} \in \mathbb{R}^{n_h}, 1 \leq i, j \leq N$, are given by
The structure of the linear algebraic system (2.2.17) in case of a checkerboard ordering of the grid points is left as an exercise.

The properties of $A_h$ imply the unique solvability of the linear system (2.2.16) and hence of the finite difference method (2.2.14a),(2.2.14b). The properties of $A_h$ further imply the convergence of $u_h$ to the solution $u$ of Poisson’s equation provided $f_h \to f$ and $u_h^D \to u^D$ for $h \to 0$. We want to address the issue of the convergence of finite difference methods in a more general framework. For that purpose we consider such a method for the approximate solution of the boundary value problem (2.1.4a),(2.1.4b) in $\Omega = (a,b)^2$ with respect to a uniform grid $\Omega_h$ with step size $h > 0$. For the approximation of the second partial derivatives $u_{x_i x_j}, i \neq j$, and the first partial derivatives $u_{x_i}$ we use central difference quotients. The mixed second partial derivatives $u_{x_i x_j}, i \neq j$, are approximated by difference quotients of the form (2.2.5),(2.2.6). This leads to the finite difference method

\begin{align}
(2.2.20a) & 
L_h u_h = f_h \quad \text{in} \quad \Omega_h , \\
(2.2.20b) & 
u_h = u_h^D \quad \text{on} \quad \Gamma_h .
\end{align}

Here, $L_h$ stands for the difference operator

\[
L_h u_h(x) := - \sum_{i,j=1}^{2} a_{ij}(x) D_{h,i,j}^2 u_h(x) + \sum_{i=1}^{2} b_i(x) D_{h,i}^1 u_h(x) + c(x) u_h(x)
\]

and $f_h \in C(\Omega_h), u_h^D \in C(\Gamma_h)$ are given grid functions.

Assuming the existence and uniqueness of a solution $u_h$, we define the convergence and the order of convergence of the finite difference method as follows:

\textbf{(2.2.21) Definition:} Given $\Omega := (a,b)^2, a, b \in \mathbb{R}, a < b$, and a uniform grid $\Omega_h$ with step size $h > 0$, let $u : \Omega \to \mathbb{R}$ be the classical solution of the boundary value problem (2.1.4a),(2.1.4b) and let $u_h \in C(\Omega_h)$ be the solution of the finite difference method (2.2.20a),(2.2.20b). Then, the finite difference method is said to be convergent, if

\[
\| u - u_h \|_h = \max_{x \in \Omega_h} | u(x) - u_h(x) | \to 0 \quad \text{for} \quad h \to 0 .
\]

It is said to be convergent of order $p$, if for sufficiently small $h$ there holds

\[
\| u - u_h \|_h = \max_{x \in \Omega_h} | u(x) - u_h(x) | \leq C \, h^p ,
\]

where $C > 0$ is a constant independent of $h$.

Sufficient conditions for the convergence can be stated in terms of the consistency with the boundary value problem (2.1.4a),(2.1.4b) and the stability
of the finite difference method. The consistency is measured by means of the local discretization error which is obtained by evaluating the finite difference method for the solution of the boundary value problem.

(2.2.22) Definition: Let \( u \) be the classical solution of the boundary value problem (2.1.4a), (2.1.4b) and let \( L_h \) and \( f_h, u^D_h \) be the difference operator and the grid functions from (2.2.20a), (2.2.20b). Then, the grid function \( \tau_h \in C(\Omega_h) \) given by

\[
\tau_h(x) := \begin{cases} 
L_h u(x) - f_h(x) & , x \in \Omega_h \\
 u(x) - u^D_h(x) & , x \in \Gamma_h 
\end{cases}
\]

is called the local discretization error. The finite difference method (2.2.20a), (2.2.20b) is said to be consistent with the boundary value problem (2.1.4a), (2.1.4b), if

\[
\max_{x \in \Omega_h} |\tau_h(x)| \to 0 \quad \text{as} \ h \to 0.
\]

It is called consistent of order \( p \), if there exists \( C > 0 \), independent of \( h \) such that

\[
\max_{x \in \Omega_h} |\tau_h(x)| \leq C h^p.
\]

(2.2.23) Lemma: Assume that the boundary value problem (2.1.4a), (2.1.4b) has a classical solution \( u \) such that \( u \in C^4(\Omega) \), Further, suppose that there holds

\[
\max_{x \in \Omega_h} |f(x) - f_h(x)| = O(h^2) \quad \text{and} \quad \max_{x \in \Gamma_h} |u^D(x) - u^D_h(x)| = O(h^2).
\]

Then, the finite difference method (2.2.20a), (2.2.20b) is consistent with the boundary value problem (2.1.4a), (2.1.4b) of order \( p = 2 \).

Proof. Under the assumptions on \( f_h \) and \( u^D_h \) the proof follows easily taking into account the properties of the difference quotients used in the definition of the difference operator \( L_h \). \( \square \)

Remark. The result of Lemma (2.2.23) can be generalized without difficulties to higher dimensions. In case of more general bounded domains and the use of the Shortley-Weller approximation, we can only expect the order of consistency \( p = 1 \).

For linear finite difference methods the stability corresponds to the asymptotically uniform boundedness of the inverse of the difference operator.

(2.2.24) Definition: The finite difference method (2.2.20a), (2.2.20b) is called stable, if there exists \( h_{\text{max}} > 0 \) such that for \( h \leq h_{\text{max}} \) the operator \( L_h \) is invertible and there is a constant \( C > 0 \) independent of \( h \) such that

\[
\|L_h^{-1}\| \leq C.
\]
(2.2.25) **Theorem:** Assume that (2.2.20a),(2.2.20b) is stable and consistent with the boundary value problem (2.1.4a),(2.1.4b). Then, the finite difference method is convergent and the order of convergence is the same as the order of consistency.

**Proof.** The proof is an immediate consequence of
\[
L_h(u(x) - u_h(x)) = \tau_h(x), \quad x \in \Omega_h,
\]
and \(u(x) - u_h(x) = \tau_h(x), x \in \Gamma_h.\) □

If the finite difference method is consistent with the boundary value problem, a sufficient condition for stability is the \(L_0\)-matrix property of the matrix associated with the finite difference method.

(2.2.26) **Definition:** A matrix \(A = (a_{ij})_{i,j=1}^n\) is called an \(L_0\)-matrix, if \(a_{ij} \leq 0, 1 \leq i \neq j \leq n.\) An \(L_0\)-matrix is said to be an \(L\)-matrix, if \(a_{ii} > 0, 1 \leq i \leq n.\) An \(L_0\)-matrix is called an \(M\)-matrix, if it is regular and \(A^{-1} \geq 0.\)

We have the following characterization of \(M\)-matrices.

(2.2.27) **Theorem:** An \(L_0\)-matrix \(A \in \mathbb{R}^{n \times n}\) is an \(M\)-matrix if and only if there exists a vector \(z \in \mathbb{R}^n\) such that \(z > 0\) and \(Az > 0.\) Moreover, there holds
\[
\|A^{-1}\| \leq \frac{\|z\|_\infty}{\min_{1 \leq i \leq n} (Az)_i},
\]
where \(\| \cdot \|\) refers to the row-sum norm.

**Proof.** Let \(A\) be an \(M\)-matrix. Then \(A^{-1} \geq 0\) and we may choose \(z = A^{-1}e,\) where \(e = (1, \cdots, 1)^T.\)

Conversely, assume \(z > 0\) and \(Az > 0,\) i.e., \(\sum_{j=1}^n a_{ij}z_j > 0, 1 \leq i \leq n.\) Since \(a_{ij}z_j \leq 0\) for \(i \neq j,\) we must have \(a_{ii} > 0, 1 \leq i \leq n.\) Hence, the matrix \(D_A := \text{diag}(A)\) is invertible. We set \(P := D_A^{-1}(D_A - A)\) and hence, we have \(A = D_A(I - P).\) Obviously, \(P \geq 0\) and \((I - P)z = D_A^{-1}Az > 0,\) whence \(Pz < z.\) For the special norm
\[
\|x\|_z := \max_{1 \leq i \leq n} \frac{|x_i|}{z_i},
\]
we thus obtain
\[
\|P\|_z = \max_{\|x\|_z = 1} \|Px\|_z = \|Pz\|_z = \max_{1 \leq i \leq n} \frac{(Pz)_i}{z_i} < 1.
\]
Consequently, \(I - P\) is invertible with \((I - P)^{-1} = \sum_{i=0}^\infty P^i \geq 0.\) We then conclude \(A^{-1} = (I - P)^{-1}D_A^{-1} \geq 0.\)
In order to prove (2.2.28), for \( b \in \mathbb{R}^n \) let \( x \in \mathbb{R}^n \) be the solution of \( Ax = b \). Then, there holds

\[
\pm x = \pm A^{-1}b \leq \|b\|_{\infty} A^{-1}e.
\]

Moreover, \( Az \geq \min_{1 \leq i \leq n} (Az)_i e \), whence

\[
A^{-1}e \leq \frac{z}{\min_{1 \leq i \leq n} (Az)_i}.
\]

From (2.2.29) and (2.2.30) we conclude

\[
\|A^{-1}b\|_{\infty} = \|x\|_{\infty} \leq \|b\|_{\infty} \frac{\|z\|_{\infty}}{\min_{1 \leq i \leq n} (Az)_i},
\]

which gives the assertion.

We apply the previous result to the finite difference approximation under consideration.

\[\text{(2.2.31) Lemma: Assume that (2.2.20a),(2.2.20b) is consistent with the boundary value problem (2.1.4a),(2.1.4b), and let } A_h \in \mathbb{R}^{n_h \times n_h} \text{ be the coefficient matrix of the linear algebraic system associated with the finite difference method. If } A_h \text{ is an } L_0-\text{matrix, then the finite difference method is stable.}\]

\[\text{Proof. In view of Theorem (2.2.27) it suffices to find } z_h \in \mathbb{R}^{n_h} \text{ such that } z_h > 0, A_h z_h > 0 \text{ and the upper bound in}
\]

\[\|A_h^{-1}\| \leq \frac{\|z_h\|_{\infty}}{\min_{1 \leq i \leq n_h} (A_h z_h)_i}
\]

is asymptotically uniformly bounded. Without restriction of generality we may assume that \( u_D = 0 \) in (2.1.4b). Let \( u^* \) be the solution of (2.1.4a),(2.1.4b) for \( f = 1 \) and \( u_D = 0 \). According to the maximum principle due to Hopf (2.1.5) we have \( u^*(x) > 0, x \in \Omega \). Moreover, \( Lu^*(x) \geq 1, x \in \Omega \). We define the vector \( z_h \) according to \( z_h := (u^*(x_1), \ldots, u^*(x_{n_h}))^T \). In view of the consistency, for sufficiently small \( h \) it follows that \( A_h z_h \geq 1/2 \). Hence, (2.2.32) implies the asymptotically uniform boundedness of \( A_h^{-1} \) and thus the stability of the finite difference method.

According to Lemma (2.2.31) the stability of the finite difference method follows from the \( L_0 \)-matrix property of \( A_h \in \mathbb{R}^{n_h \times n_h} \). It is easily shown that for \( a_{12}(x) = 0, x \in \Omega \), the difference operator \( L_h \) gives rise to the difference star

\[
\frac{1}{h^2} \begin{bmatrix}
0 & \frac{h}{2} b_2 - a_{22} & 0 \\
-\frac{h}{2} b_1 - a_{11} & 2(a_{11} + a_{22}) & \frac{h}{2} b_1 - a_{11} \\
0 & -\frac{h}{2} b_2 - a_{22} & 0
\end{bmatrix}.
\]
Assuming \( h|b_i(x)| < 2a_{ii}(x), 1 \leq i \leq 2 \), \( A_h \) turns out to be an M-matrix. However, for non-vanishing \( a_{12} \) we have to choose the difference quotients \( D_{h,i,j}^2 u_h(x), 1 \leq i \neq j \leq 2 \), from (2.2.5) with (2.2.6) depending on the sign of \( a_{12}(x) \). For \( \gamma_1 = \pm 1, \gamma_2 = \gamma_3 = 0 \) in (2.2.6) we obtain the difference stars

\[
\frac{1}{2h^2} \begin{bmatrix}
-1 & 1 & 0 \\
1 & -2 & 1 \\
0 & 1 & -1
\end{bmatrix}, \quad \frac{1}{2h^2} \begin{bmatrix}
0 & -1 & 1 \\
-1 & 2 & -1 \\
1 & -1 & 0
\end{bmatrix}.
\]

We denote by \( D_{h,i,j}^{2,-} \) and by \( D_{h,i,j}^{2,+} \) the difference operators associated with the left and right difference star. Then, if \( \pm a_{12}(x) > 0 \) and we use the difference quotient \( D_{h,i,j}^{2,\pm} u_h(x) \) in the definition of \( A_h \), it is easily shown that the condition

\[
(2.2.33) \quad h |b_i(x)| < 2 \left( a_{ii}(x) - |a_{12}(x)| \right), \quad 1 \leq i \leq 2, \ x \in \Omega
\]

implies the \( L_0 \)-matrix property of \( A_h \).

(2.2.34) **Lemma:** Let \( A_h \) be the difference operator from (2.2.20a) such that \( D_{h,i,j}^2 u_h(x) = D_{h,i,j}^{2,\pm} u_h(x), 1 \leq i \neq j \leq 2 \), if \( \pm a_{12}(x) > 0, x \in \Omega \), and suppose that (2.2.33) holds true. Then, the finite difference method (2.2.20a),(2.2.20b) is stable.

**Proof.** Condition (2.2.33) implies the M-matrix property of the matrix associated with the difference operator and hence, the assertion follows from Lemma (2.2.31). \( \square \)

In summary we obtain the following result:

(2.2.35) **Theorem:** Under the assumptions of Lemma (2.2.23) and Lemma (2.2.34) the finite difference method (2.2.20a),(2.2.20b) is convergent of order \( p = 2 \).

**Proof.** Lemma (2.2.23) proves the consistency of order \( p = 2 \) and Lemma (2.2.34) guarantees the stability of the finite difference method. Theorem (2.2.25) allows to conclude. \( \square \)

**Remark.** (i) For general bounded domains and the Shortley-Weller approximation the stability of finite difference methods can be shown under appropriate conditions on the coefficient functions of the difference operator. According to Remark (10.3.2.26) the order of convergence reduces to \( p = 1 \).

(ii) If the coefficient function \( b_i \) significantly dominates the coefficient function \( a_{ii} \) in absolute value, the step size \( h \) has to be chosen unrealistically small in order to guarantee condition (2.2.33). Such a problem is called convection-dominated, since for diffusion-convection-reaction problems as described by (2.1.4a) the term \( b \cdot \nabla u \) represents the convective part of the process. For such problems, the discretization of the first partial derivatives is done by
the forward or the backward difference quotient depending on the sign of $b_i$ (upwind discretization).

The convergence analysis presented in this section relies on smoothness assumptions with respect to the classical solution of the boundary value problem. For many problems, such assumptions do not hold true, e.g., in case of singularities due to inconsistent boundary conditions or due to the geometry of the domain. As far as the development, analysis and implementation of finite difference methods for such problems is concerned, we refer to Samarskii, Lazarov and Makarov (1987).

### 2.3 Weak solutions

We consider a boundary value problem for a linear second order elliptic differential operator in selfadjoint form with homogeneous Dirichlet boundary conditions

\begin{align}
(2.3.1a) & \quad Lu = f \quad \text{in } \Omega, \\
(2.3.1b) & \quad u = 0 \quad \text{on } \Gamma,
\end{align}

where

$$Lu := -\nabla\left(a\nabla u\right) + cu.$$  

We assume $\Omega \subset \mathbb{R}^d$ to be a bounded Lipschitz domain with boundary $\Gamma = \partial \Omega$. We further suppose that $a = (a_{ij})_{i,j=1}^d$ with $a_{ij} \in L^\infty(\Omega)$, $1 \leq i, j \leq d$, is a uniformly positive definite matrix-valued function, i.e., there exists a constant $\alpha > 0$ such that for almost all $x \in \Omega$

\begin{equation}
\sum_{i,j=1}^d a_{ij}(x)\xi_i\xi_j \geq \alpha|\xi|^2, \quad \xi = (\xi_1, \cdots, \xi_d)^T \in \mathbb{R}^d.
\end{equation}

Moreover, we assume $c \in L^\infty(\Omega)$ such that $c(x) \geq 0$ for almost all $x \in \Omega$ and $f \in L^2(\Omega)$.

Multiplication of (2.3.1a) by a function $v \in C^\infty_0(\Omega)$, integration over $\Omega$ and the application of Green’s formula yields

\begin{equation}
\int_{\Omega} \left( a\nabla u \cdot \nabla v + cuv \right) \, dx = \int_{\Omega} fv \, dx.
\end{equation}

Obviously, the integrals in (2.3.3) are well defined for functions $u$ and $v$ in the Sobolev space $H^1_0(\Omega)$ (see section (1.2)). Therefore, we introduce the notion of weak (generalized) solution.

\begin{equation}
\textbf{Definition:} A function } u \in H^1_0(\Omega) \text{ is called a weak or generalized solution of } (2.3.1a), (2.3.1b), \text{ if (2.3.3) is satisfied for all } v \in H^1_0(\Omega).
\end{equation}
2.3 Weak solutions

In case \( f \in C(\Omega) \) it is obvious that a classical solution of (2.3.1) in the sense of Definition 2.1.3 also represents a solution in the weak sense. On the other hand, if \( u \in H^0_0(\Omega) \) is a weak solution such that \( u \in H^2(\Omega) \), the application of Green’s formula to (2.3.3) yields

\[
\int_{\Omega} (Lu - f)v \, dx = 0, \quad v \in H^1_0(\Omega),
\]

and hence, \( Lu(x) = f(x) \) for almost all \( x \in \Omega \).

(2.3.5) Definition: A function \( u \) such that \( Lu \in L^2(\Omega) \), which satisfies (2.3.1a) for almost all \( x \in \Omega \), is called a strong solution.

If \( f \in C(\Omega) \) and if a weak solution satisfies \( u \in H^m(\Omega) \) for sufficiently large \( m \) such that \( H^m(\Omega) \hookrightarrow C^2(\Omega) \cap C_0(\overline{\Omega}) \), it follows that \( Lu(x) = f(x), x \in \Omega, \) and \( u(x) = 0, x \in \Gamma, \) i.e., \( u \) is a solution in the classical sense.

We observe that a weak solution is a solution in the strong or in the classical sense only if it has additional smoothness properties. The study of the regularity of weak solutions is a central issue in the theory of partial differential equations (cf. e.g., Evans (2000)).

The left-hand side in (2.3.3) is a bilinear form on \( H^1_0(\Omega) \times H^1_0(\Omega) \), whereas the right-hand side represents a bounded linear functional on \( H^1_0(\Omega) \). Therefore, we study the existence and uniqueness of weak solutions within the framework of variational equations in Hilbert spaces:

(2.3.6) Definition: Let \( V \) be a real Hilbert space with inner product \((\cdot, \cdot)_V\) and associated norm \( \| \cdot \|_V \), let \( a(\cdot, \cdot) : V \times V \to \mathbb{R} \) be a bilinear form and \( \ell \in V^* \). The problem to find a function \( u \in V \) such that

\[
a(u, v) = \ell(v), \quad v \in V,
\]

is called a variational equation.

As we shall see now, the existence and uniqueness of a solution of the variational equation (2.3.7) essentially requires the boundedness and \( V \)-ellipticity of the bilinear form.

(2.3.8) Definition: 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

\[
|a(u, v)| \leq C \| u \|_V \| v \|_V, \quad u, v \in V.
\]

(2.3.10) Definition: A bilinear form \( a(\cdot, \cdot) : V \times V \to \mathbb{R} \) is called \( V \)-elliptic, if there exists a constant \( \alpha > 0 \) such that

\[
|a(u, u)| \geq \alpha \| u \|_V^2, \quad u \in V.
\]
The following celebrated *Lemma of Lax-Milgram* provides an existence and uniqueness result for variational equations of the form (2.3.7).

(2.3.12) **Theorem:** Let $V$ be a Hilbert space with dual $V^*$ and suppose that $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is a bounded, $V$-elliptic bilinear form and $\ell \in V^*$. Then, the variational equation (2.3.7) has a unique solution $u \in V$.

**Proof.** We denote by $A : V \to V^*$ the operator associated with the bilinear form according to

$$
(2.3.13) \quad Au(v) = a(u,v) , \quad u, v \in V.
$$

The operator $A$ is a bounded linear operator, since in view of (2.3.9) there holds

$$
\|Au\|_{V^*} = \sup_{v \neq 0} \frac{|Au(v)|}{\|v\|_V} \leq C \|u\|_V,
$$

and hence,

$$
\|A\| = \sup_{u \neq 0} \frac{\|Au\|_{V^*}}{\|u\|_V} \leq C.
$$

We will reformulate the variational equation (2.3.7) as an operator equation in $V$ by means of the Riesz mapping $\tau : V^* \to V$, $\ell \mapsto \tau\ell$ which is given by

$$
(2.3.14) \quad \ell(v) = (\tau\ell, v)_V , \quad v \in V.
$$

It can be easily verified that the Riesz mapping is an isometry, i.e.,

$$
(2.3.15) \quad \|\tau\ell\|_V = \|\ell\|_{V^*} , \quad \ell \in V^*.
$$

Indeed, using the Riesz mapping, the variational equation (2.3.7) can be equivalently written as the operator equation

$$
(2.3.16) \quad \tau(Au - \ell) = 0 .
$$

We prove the existence and uniqueness of a solution of (2.3.16) by an application of Banach’s fixed point theorem. For that purpose we consider the mapping $T : V \to V$ defined by means of

$$
(2.3.17) \quad Tv := v - \rho (\tau(Av - \ell)) , \quad \rho > 0.
$$

Obviously, $u \in V$ solves (2.3.16) if and only if $u \in V$ is a fixed point of the operator $T$. For the proof of the existence and uniqueness of a fixed point in $V$ it suffices to show that the operator $T$ is a contraction on $V$, i.e., there exists a constant $0 \leq q < 1$ such that
\[ \|Tv_1 - Tv_2\|_V \leq q \|v_1 - v_2\|_V, \quad v_1, v_2 \in V. \]

Setting \( w := v_1 - v_2 \) and taking (2.3.13) as well as (2.3.15) into account it follows that

\[ \|Tw\|_V^2 = \|w - \rho \tau Aw\|_V^2 = \|w\|_V^2 - 2\rho (\tau Aw, w)_V + \rho^2 \|\tau Aw\|_V^2 \leq \left( 1 - 2\rho \alpha + \rho^2 C^2 \right) \|w\|_V^2. \]

We thus obtain

\[ \left( 1 - 2\rho \alpha + \rho^2 C^2 \right) < 1 \iff \rho < \frac{2\alpha}{C^2}, \]

which proves the assertion.

**Example.** We consider the elliptic boundary value problem (2.3.1). The associated bilinear form \( a(\cdot, \cdot) : H^1_0(\Omega) \times H^1_0(\Omega) \rightarrow \mathbb{R} \) is given by

\[ a(u, v) := \int_{\Omega} \left( a \nabla u \cdot \nabla v + cuv \right) \, dx, \quad u, v \in H^1_0(\Omega). \]

in view of \( a_{ij} \in L^\infty(\Omega), 1 \leq i, j \leq d, \) and \( c \in L^\infty(\Omega) \) we easily deduce the boundedness. On the other hand, (2.3.2) and \( c(x) \geq 0 \) f.a.a. \( x \in \Omega \) readily imply

\[ a(u, u) \geq \alpha |u|_{1,\Omega}^2, \quad u \in H^1_0(\Omega). \]

Since on \( H^1_0(\Omega) \) the \( |\cdot|_{1,\Omega}\)-norm is equivalent to the \( \|\cdot\|_{1,\Omega}\)-norm, the previous inequality proves the \( H^1_0(\Omega)\)-ellipticity. Finally, for \( f \in L^2(\Omega) \) the linear functional \( \ell(v) := \int_{\Omega} f v \, dx \) is bounded on \( H^1_0(\Omega) \). The Lax-Milgram lemma (Theorem 2.3.12) implies the existence and uniqueness of a weak solution of (2.3.1).

### 2.4 Conforming finite element methods

We assume that \( V \subseteq H^1(\Omega) \) is 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-Milgram (2.3.12) implies that the variational equation

\[ (2.4.1) \quad a(u, v) = \ell(v), \quad v \in V \]

admits a unique solution \( u \in V \). The *Galerkin method* consists in the approximation of \( u \in V \) by an approximate solution \( u_h \in V_h \) in a finite dimensional subspace \( V_h \subset V, \dim V_h = m_h \), which is the solution of the variational equation (2.4.1) restricted to \( V_h \), i.e.,
Another application of the Lemma of Lax-Milgram reveals the existence and uniqueness of $u_h \in V_h$.

**Remark.** If the bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ is symmetric, the Galerkin method is also referred to as the Ritz-Galerkin method.

Specifying a basis $(\varphi_h^{(i)})_{i=1}^{n_h}$ of $V_h$, it is easily seen that the Galerkin method leads to a linear algebraic system. Since the solution $u_h \in V_h$ of (2.4.2) can be written as a linear combination of the basis functions according to

$$u_h = \sum_{j=1}^{n_h} \alpha_j \varphi_h^{(j)}$$

$u_h$ satisfies the variational equation (2.4.2) if and only if

$$\sum_{j=1}^{n_h} a(\varphi_h^{(j)}, \varphi_h^{(i)}) \alpha_j = \ell(\varphi_h^{(i)}) , \quad 1 \leq i \leq n_h .$$

Obviously, (2.4.4) represents a linear algebraic system

$$A_h \alpha_h = b_h$$

in the unknown vector $\alpha_h = (\alpha_1, \ldots, \alpha_{n_h})^T$. The *stiffness matrix* $A_h \in \mathbb{R}^{n_h \times n_h}$ and the *load vector* $b_h = (b_{h,1}, \ldots, b_{h,n_h})^T \in \mathbb{R}^{n_h}$ are given by

$$A_h := \begin{pmatrix} a(\varphi_h^{(1)}, \varphi_h^{(1)}) & \cdots & a(\varphi_h^{(n_h)}, \varphi_h^{(1)}) \\ \vdots & \ddots & \vdots \\ a(\varphi_h^{(1)}, \varphi_h^{(n_h)}) & \cdots & a(\varphi_h^{(n_h)}, \varphi_h^{(n_h)}) \end{pmatrix}$$

and

$$b_{h,i} := \ell(\varphi_h^{(i)}) , \quad 1 \leq i \leq n_h .$$

**Remark.** We remark that the notions 'stiffness matrix' and 'load vector' are stemming from mechanical applications.

With regard to the appropriate choice of finite dimensional subspaces, there are essentially two aspects, namely the

- efficient numerical solution of the linear algebraic system (2.4.5) and the
- accuracy of the approximation of the solution $u \in V$ of (2.4.1) by the solution $u_h \in V_h$ of (2.4.2).
The latter aspect will be addressed in Céa’s Lemma. Céa’s Lemma states that under the assumptions of the Lax-Milgram Lemma the accuracy of the approximate solution corresponds to the best approximation of the solution \( u \in V \) by a function in \( V_h \). Hence, the issue is reduced to a problem in approximation theory. Céa’s Lemma is based on the observation that the error \( u - u_h \) is \( a \)-orthogonal with respect to \( V_h \), i.e.,

\[
a(u - u_h, v_h) = 0, \quad v_h \in V_h.
\]

This property is called the Galerkin orthogonality.

(2.4.9) **Definition:** If the bilinear form \( a(\cdot, \cdot) \) is symmetric 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. Galerkin orthogonality (2.4.8) means that the solution \( u_h \in V_h \) of (2.4.2) is the projection of the solution \( u \in V \) of (2.4.1) onto \( V_h \). Therefore, the approximate solution is referred to as the elliptic projection.

(2.4.10) **Theorem:** Under the assumptions of the Lemma of Lax-Milgram (2.3.12) let \( u \in V \) and \( u_h \in V_h \) be the unique solutions of (2.4.1) and (2.4.2). Then, there holds

\[
\| u - u_h \|_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} \| u - v_h \|_V.
\]

**Proof.** Using the \( V \)-ellipticity and the boundedness of \( a(\cdot, \cdot) \) as well as the Galerkin orthogonality (2.4.8), for an arbitrary \( v_h \in V_h \) we have

\[
\alpha \| u - u_h \|_V^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \leq 0
\]

\[
\leq C \| u - u_h \|_V \| u - v_h \|_V,
\]

which implies (2.4.11).

The efficient numerical solution of the linear algebraic system (2.4.5) resulting from the Galerkin method depends on the structure of the stiffness matrix \( A_h \). In particular, we are interested in sparsely populated matrices with a specific sparsity pattern which can be realized by an appropriate choice of the basis functions of \( V_h \subset V \) with minimal support. In case of conforming finite element methods, \( V_h \) is constructed on the basis of triangulations of the computational domain \( \Omega \subset \mathbb{R}^d \).

(2.4.12) **Definition:** Let \( \Omega \subset \mathbb{R}^d \) be a bounded Lipschitz domain. A triangulation \( T_h(\Omega) \) of \( \Omega \) is a partition of \( \Omega \) into a finite number of subsets \( T \), called finite elements such that
We distinguish between two types of finite elements: simplicial elements and quadrilateral elements.

**Definition:** 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 $T$ is called non-degenerate, if every point $x \in \mathbb{R}^d$ can be uniquely represented according to

$$x = \sum_{j=1}^{d+1} \lambda_j a_j, \lambda_j \in \mathbb{R}, \sum_{j=1}^{d+1} \lambda_j = 1.$$  

In $\mathbb{R}^2$, a simplex is a triangle, in $\mathbb{R}^3$ a tetrahedron.

**Remark.** The non-degeneracy of a simplex is related to the unique solvability of the linear algebraic system

$$\begin{pmatrix}
  a_{11} & \cdots & a_{1,d+1} \\
  \vdots & \ddots & \vdots \\
  a_{d1} & \cdots & a_{d,d+1} \\
  1 & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
  \lambda_1 \\
  \vdots \\
  \vdots \\
  \lambda_d
\end{pmatrix}
=:
\begin{pmatrix}
  x_1 \\
  \vdots \\
  \vdots \\
  x_d
\end{pmatrix}.$$  

Obviously, $T$ is non-degenerate if and only if the matrix $A$ is non-singular.

**Definition:** The points $a_j, 1 \leq j \leq d + 1$, of a simplex $T$ are called vertices. An $m$-dimensional face of a simplex $T$, $0 \leq m \leq d$, is a simplex in $\mathbb{R}^m$, whose vertices are vertices of $T$ in $D$. A one-dimensional face is said to be 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 $\mathcal{T}_h(\Omega)$. The simplex $\hat{T}$ with the vertices $\hat{a}_1 = (0, \ldots, 0)^T$ and $\hat{a}_{i+1} = e_i, 1 \leq i \leq d$, is called simplicial reference element.

**Definition:** 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 (2.4.15).
The center of gravity \(x_S\) of a non-degenerate simplex \(T\) is the point satisfying
\[
\lambda_j(x_S) = \frac{1}{d+1}, \quad 1 \leq j \leq d + 1.
\]

(2.4.18) **Lemma:** Any non-degenerate simplex \(T \subset \mathbb{R}^d\) is the image of the simplicial reference element \(\hat{T}\) under an affine transformation
\[
F_T : \mathbb{R}^d \to \mathbb{R}^d
\]
\[
\hat{x} \mapsto F_T(\hat{x}) = B_T \hat{x} + b_T
\]
with a non-singular matrix \(B_T \in \mathbb{R}^{d \times d}\) and a vector \(b_T \in \mathbb{R}^d\).

**Proof.** The proof is left as an exercise. \(\square\)

(2.4.19) **Definition:** A triangulation \(T_h(\Omega)\) of a polyhedral domain \(\Omega \subset \mathbb{R}^d\) is called a simplicial triangulation, if any of its elements is a non-degenerate simplex in \(\mathbb{R}^d\).

An alternative to simplicial triangulations is to use quadrilateral elements.

(2.4.20) **Definition:** 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.,
\[
T = \prod_{i=1}^d [c_i, d_i] = \{ x = (x_1, ..., x_d)^T \mid c_i \leq x_i \leq d_i, \ 1 \leq i \leq d \}.
\]

A quadrilateral element in \(\mathbb{R}^2\) is a rectangle, in \(\mathbb{R}^3\) cube. A quadrilateral element \(T\) is said to be non-degenerate, if \(c_i < d_i, 1 \leq i \leq d\).

The quadrilateral element \(\hat{T} := [0, 1]^d\) (unit cube) is called the quadrilateral reference element in \(\mathbb{R}^d\). The points
\[
a_j = (a_{j1}, ..., a_{jd})^T, \quad a_{ji} = c_i \text{ or } a_{ji} = d_i, \ 1 \leq i \leq d,
\]
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 vertices of \(T\). A one-dimensional face is said to be an edge. For \(D \subseteq \Omega\) we refer to \(F_h(D)\) and \(E_h(D)\) as the sets of \((d - 1)\)-dimensional faces and one-dimensional edges of \(T_h(\Omega)\) in \(D\).

(2.4.21) **Lemma:** Any non-degenerate quadrilateral element \(T \subset \mathbb{R}^d\) is the image of the quadrilateral reference element \(\hat{T}\) under a diagonal-affine transformation
\[
F_T : \mathbb{R}^d \to \mathbb{R}^d
\]
\[
\hat{x} \mapsto F_T(\hat{x}) = B_T \hat{x} + b_T
\]
with a non-singular diagonal matrix \(B_T = (b_{ii})_{i=1}^d\) and a vector \(b_T \in \mathbb{R}^d\).
**Proof.** The proof is left as an exercise. □

(2.4.22) **Definition:** A triangulation $T_h(\Omega)$ of a quadrilateral domain $\Omega \subset \mathbb{R}^d$ is called a quadrilateral triangulation, if any of its elements $T$ is a quadrilateral element.

Given a triangulation $T_h(\Omega)$, conforming finite element functions will be specified locally for the elements $T \in T_h(\Omega)$ and then composed such that the resulting global function belongs to the underlying function space $V$.

(2.4.23) **Definition:** Let $T_h(\Omega)$ be a triangulation of $\Omega \subset \mathbb{R}^d$, and $P_T$ be a linear space of functions $p : T \to \mathbb{R}$ such that $\dim P_T = n_T$. The elements of $P_T$ are called local trial functions. For linear bounded functionals $\ell_i : P_T \to \mathbb{R}, 1 \leq i \leq n_T$, we define

(2.4.24) $\Sigma_T := \{ \ell_i(p) \mid p \in P_T \ , \ 1 \leq i \leq n_T \}.$

The elements of the set $\Sigma_T$ are called degrees of freedom.

(2.4.25) **Definition:** Let $T_h(\Omega)$ be a triangulation of $\Omega \subset \mathbb{R}^d$, $T \in T_h(\Omega)$, and $P_T, \Sigma_T$ as in Definition (2.4.23). Then, the triple $(T, P_T, \Sigma_T)$ is called a finite element. A finite element $(T, P_T, \Sigma_T)$ is said to be unisolvent, if any $p \in P_T$ is uniquely determined by the degrees of freedom in $\Sigma_T$.

**Remark.** We remark that the notion 'finite element' is used for $T \in T_h(\Omega)$ as well as for the triple $(T, P_T, \Sigma_T)$.

The numerical computation of the elements of the stiffness matrix and the components of the load vector is greatly facilitated in case of affine equivalent finite elements.

(2.4.26) **Definition:** Let $T_h(\Omega)$ be a triangulation of $\Omega \subset \mathbb{R}^d$, $T \in T_h(\Omega)$, and $P_T, \Sigma_T$ as in Definition (2.4.23). Moreover, let $(\hat{T}, \hat{P}_T, \hat{\Sigma}_T)$ be a reference element. Then, the finite elements $(T, P_T, \Sigma_T), T \in T_h(\Omega)$ are called affine equivalent, if there exists an invertible affine mapping $F_T : \mathbb{R}^d \to \mathbb{R}^d$ such that for all $T \in T_h(\Omega)$

(2.4.27a) $T = F_T(\hat{T})$,  
(2.4.27b) $P_K = \{ p : T \to \mathbb{R} \mid p = \hat{p} \circ F_T^{-1}, \hat{p} \in \hat{P}_T \}$,  
(2.4.27c) $\Sigma_T = \{ \ell_i : P_T \to \mathbb{R} \mid \ell_i = \hat{\ell}_i \circ F_T^{-1}, \hat{\ell}_i \in \hat{\Sigma}_T \ , \ 1 \leq i \leq n_T \}.$

(2.4.28) **Definition:** Let $T_h(\Omega)$ be a triangulation of $\Omega \subset \mathbb{R}^d$, $T \in T_h(\Omega)$, and $P_T, \Sigma_T$ as in Definition (refdef:10.3.4.25). Then,

(2.4.29) $V_h := \{ v_h : \overline{\Omega} \to \mathbb{R} \mid v_h|_T \in P_T \ , \ T \in T_h(\Omega) \}$

is said to be a finite element space.
The following result provides sufficient conditions for a finite element space $V_h$ to be a subspace of $H^1(\Omega)$.

(2.4.30) **Theorem:** Let $V_h$ be a finite element space and assume that there holds

\begin{align}
P_T &\subset H^1(T) \quad , \quad T \in T_h(\Omega) , \\
V_h &\subset C(\Omega) .
\end{align}

Then, we have

(2.4.32) \quad $V_h \subset H^1(\Omega)$.

**Proof.** Let $v_h \in V_h$. Obviously, we have $v_h \in L^2(\Omega)$. It remains to be shown that $v_h$ has weak first derivatives $w_\alpha^h \in L^2(\Omega)$, $|\alpha| = 1$, i.e.,

(2.4.33) \quad \int_\Omega v_h D^\alpha z \, dx = (-1)^{|\alpha|} \int_\Omega w_\alpha^h \, z \, dx , \quad z \in C_0^\infty(\Omega) .

Since $v_h |_{T} \in H^1(T)$, we apply Green’s formula elementwise and obtain

(2.4.34) \quad \int_\Omega v_h D^\alpha z \, dx = \sum_{T \in T_h(\Omega)} \int_T v_h D^\alpha z \, dx =

\quad = - \sum_{T \in T_h(\Omega)} \int_T D^\alpha v_h \, z \, dx + \sum_{T \in T_h(\Omega)} \int_{\partial T} v_h \nu_\alpha D^\alpha z \, d\sigma =

\quad = - \sum_{T \in T_h(\Omega)} \int_T D^\alpha v_h \, z \, dx + \sum_{F \in 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 $F = T_1 \cap T_2, T_i \in T_h(\Omega), 1 \leq i \leq 2$. But $v_h \in C(\Omega)$ and hence, $[v_h] = 0$ in (2.4.34) which shows (2.4.33) with $w_\alpha^h |_{T} := D^\alpha v_h |_{T}, T \in T_h(\Omega)$.

\[\square\]

**Corollary:** Let $V_h$ be a finite element space and assume that

\begin{align}
P_T &\subset H^1(T) \quad , \quad T \in T_h(\Omega) , \\
V_h &\subset C_0(\Omega) .
\end{align}

Then, there holds

(2.4.36) \quad $V_h \subset H^1_0(\Omega)$.

**Proof.** The proof is an immediate consequence of Theorem (2.4.30).

In the following we will be concerned with the construction of simplicial Lagrangian finite elements. The local trial functions are chosen as polynomials.
of degree \( k \): Let \( T \) be a simplex in \( \mathbb{R}^d \). For \( k \in \mathbb{N}_0 \) we define \( P_k(T) \) as den
the linear space of polynomials of degree \( \leq k \) in \( T \), i.e., \( p \in P_k(T) \), if
\[
p(x) = \sum_{|\alpha| \leq k} a_\alpha \, x^\alpha \quad , \quad a_\alpha \in \mathbb{R} , \quad |\alpha| \leq k ,
\]
where
\[
x^\alpha := \prod_{i=1}^d x_i^{\alpha_i} \quad , \quad \alpha := (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d \quad , \quad |\alpha| := \sum_{i=1}^d \alpha_i .
\]
We note that
\[
(2.4.37) \quad \dim P_k(T) = \binom{k + d}{d} = \frac{(k + d)!}{d! \, k!} .
\]

(2.4.38) **Definition:** Let \( T \) be a simplex in \( \mathbb{R}^d \). Denoting by \( a_i, 1 \leq i \leq d + 1 \), the vertices of \( T \), the set
\[
(2.4.39) \quad L_k(T) := \{ x = \sum_{i=1}^{d+1} \lambda_i \, a_i \mid \lambda_i \in \{0, \frac{1}{k}, \ldots, \frac{k-1}{k}, 1\} , \sum_{i=1}^{d+1} \lambda_i = 1 \}
\]
is called the principal lattice of order \( k \) of \( T \). We have
\[
\text{card } L_k(T) = \binom{k + d}{d} .
\]
The elements of the principal lattice of order \( k \) are called nodal points. For \( D \subseteq \Omega \) we denote by \( \mathcal{N}_h(D) \) the set of nodal points in \( D \).

Defining \( \Sigma_T \) as the set of degrees of freedom as given by the values of \( p \in P_k(T) \) in the nodal points, we obtain a simplicial Lagrangian finite element.

(2.4.40) **Definition:** Let \( T_h(\Omega) \) be a simplicial triangulation of \( \Omega \subset \mathbb{R}^d \). The triple \((T, P_k(T), L_k(T))\), \( T \in T_h(\Omega) \), is called a simplicial Lagrangian finite element of type \((k)\).

(2.4.41) **Lemma:** A simplicial Lagrangian finite element of type \((k)\) is unisolvant.

**Proof.** The proof is left as an exercise. \( \square \)

(2.4.42) **Lemma:** Simplicial finite elements of type \((k)\) are affine-equivalent to the simplicial reference element \((\hat{T}, \hat{P}_T, \hat{\Sigma}_T)\).

**Proof.** Let \( F_T : \mathbb{R}^d \to \mathbb{R}^d \) be the invertible affine mapping such that \( T = F_T(\hat{T}) \). Then, there holds \( P_k(T) = \{ p = \hat{p} \circ F_T^{-1} \mid \hat{p} \in \hat{P}_k(\hat{T}) \} \). Since \( x \in L_k(T) \) iff \( x = F_T(\hat{x}), \hat{x} \in L_k(\hat{T}) \), we also have \((2.4.27c)\). \( \square \)

**Remark.** In case \( d = 2 \), a simplicial Lagrangian finite element of type \((1)\) is called Courant’s triangle.
(2.4.43) **Definition:** The finite element space $V_h$ consisting of simplicial Lagrangian finite elements of type $(k)$ is called simplicial Lagrangian finite element space and will be denoted by $S_k(\Omega, T_h(\Omega))$.

In order to guarantee the conformity of $S_k(\Omega, T_h(\Omega))$ we have to assume that the simplicial triangulation $T_h(\Omega)$ is geometrically conforming.

(2.4.44) **Definition:** A simplicial triangulation $T_h(\Omega)$ is said to be geometrically conforming, if the intersection of two distinct elements of the triangulation is either empty or consists of a common face or a common edge or a common vertex.

(2.4.45) **Lemma:** Let $T_h(\Omega)$ be a geometrically conforming triangulation. Then, there holds

\[(2.4.46) \quad S_k(\Omega, T_h(\Omega)) = \{v_h \in C(\overline{\Omega}) \mid v_h|_T \in P_k(T), T \in T_h(\Omega)\} \subset H^1(\Omega).\]

**Proof.** Since $P_k(T) \subset H^1(T), T \in T_h(\Omega)$, and in view of Theorem (2.4.30) we have to show that $S_k(\Omega, T_h(\Omega)) \subset C(\overline{\Omega})$. The proof will be given in case $d = 2$ and $k = 2$: Let $T_1 \in T_h(\Omega), 1 \leq i \leq 2$, be two adjacent elements such that $E = T_1 \cap T_2 \in \mathcal{E}_h(\Omega)$ with nodal points $a_j \in N_h(E), 1 \leq j \leq 3$. Moreover, assume $p_i \in P_2(K_i), 1 \leq i \leq 2$. Then $p_i|_E \in P_2(E)$. Since $p_1(a_j) = p_2(a_j), 1 \leq j \leq 3$, we conclude that $p_1|_E \equiv p_2|_E$. □

**Corollary:** We define the subspace

\[(2.4.47) \quad S_{k,0}(\Omega, T_h(\Omega)) := \{v_h \in S_k(\Omega, T_h(\Omega)) \mid v_h|_{\partial T \cap \partial \Omega} = 0, T \in T_h(\Omega), T \cap \partial \Omega \neq \emptyset\}.\]

Under the same assumptions as in Lemma (2.4.45) there holds

\[(2.4.48) \quad S_{k,0}(\Omega, T_h(\Omega)) \subset H^1_0(\Omega).\]

It remains to specify the nodal basis functions:

(2.4.49) **Definition:** Let $N_h(\overline{\Omega}) = \{x_j \mid 1 \leq j \leq n_h\}$. The function $\varphi_i \in S_k(\Omega, T_h(\Omega))$ as given by

\[(2.4.50) \quad \varphi_i(x_j) = \delta_{ij}, \quad 1 \leq i, j \leq n_h,\]

is called the nodal basis function associated with the nodal point $x_i$. The set of these functions is a basis of the Lagrangian finite element space $S_k(\Omega, T_h(\Omega))$.

The representation of nodal basis functions by means of barycentric coordinates as well as a geometric characterization are left as exercises.

We will now deal with the construction of quadrilateral Lagrangian 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 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

$$p(x) = \sum_{\alpha_1, \ldots, \alpha_d \leq k} a_{\alpha_1 \ldots \alpha_d} x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_d^{\alpha_d},$$

whence

$$\dim Q_k(T) = (k + 1)^d.$$
(2.4.59) **Lemma:** A Lagrangian finite element of type $[k]$ is unisolvent.

**Proof.** The proof is an immediate consequence of Lemma 2.4.56. □

(2.4.60) **Lemma:** Let $\hat{T} := [0, 1]^d$ be the quadrilateral reference element and

(2.4.61a) $\hat{P}_k := Q_k(\hat{T}), \ k \in \mathbb{N}$,
(2.4.61b) $\hat{\Sigma}_k := \{ \hat{p}(\hat{x}) | \hat{x} \in L_k(\hat{T}), \hat{p} \in \hat{Q}_k \}$.

The Lagrangian finite elements of type $[k]$ are affine equivalent to the reference element $(\hat{T}, \hat{Q}_k, \hat{\Sigma}_k)$.

**Proof.** The proof is left as an exercise. □

(2.4.62) **Definition:** A quadrilateral triangulation $\mathcal{T}_h(\Omega)$ is called geometrically conforming, if the intersection of two distinct elements is either empty or consists of a common face or a common edge or a common vertex.

(2.4.63) **Definition:** Let $\Omega \subset \mathbb{R}^d$ be the union of a finite number of quadrilaterals in $\mathbb{R}^d$ and let $\mathcal{T}_h(\Omega)$ be a quadrilateral triangulation. The finite element space $V_h$ consisting of quadrilateral finite elements of type $[k]$ is called a quadrilateral Lagrangian finite element space and will be denoted by $S_{[k]}(\Omega, \mathcal{T}_h(\Omega))$, i.e., we have

$$S_{[k]}(\Omega, \mathcal{T}_h(\Omega)) := \{ v_h : \overline{\Omega} \to \mathbb{R} | v_h|_T \in Q_k(T), T \in \mathcal{T}_h(\Omega) \}.$$

We further define the subspace

$$S_{[k],0}(\Omega, \mathcal{T}_h(\Omega)) := \{ v_h \in S_{[k]}(\Omega, \mathcal{T}_h(\Omega)) | v_h|_{\partial T \cap \partial \Omega} = 0, T \in \mathcal{T}_h(\Omega), T \cap \partial \Omega \neq \emptyset \}.$$

(2.4.64) **Lemma:** Let $\mathcal{T}_h(\Omega)$ be a geometrically conforming triangulation of the computational domain $\Omega$. Then, there holds

(2.4.65a) $S_{[k]}(\Omega, \mathcal{T}_h(\Omega)) \subset H^1(\Omega)$,
(2.4.65b) $S_{[k],0}(\Omega, \mathcal{T}_h(\Omega)) \subset H^1_0(\Omega)$.

**Proof.** The proof is left as an exercise. □

(2.4.66) **Definition:** Let $\mathcal{N}_h(\overline{\Omega}) = \{ x \in L_{[k]}(T) | T \in \mathcal{T}_h(\Omega) \}, \ \text{card}(\mathcal{N}_h(\overline{\Omega})) = n_h$. The function $\varphi_i \in S_{[k]}(\Omega, \mathcal{T}_h(\Omega))$ as given by

$$\varphi_i(x_j) = \delta_{ij}, \ 1 \leq i, j \leq n_h,$$

is called the nodal basis function associated with the nodal point $x_i \in \mathcal{N}_h(\overline{\Omega})$. The set of these functions is a basis of $S_{[k]}(\Omega, \mathcal{T}_h(\Omega))$. 
The notion ‘Lagrangian finite element’ stems from the underlying polynomial interpolation of Lagrangian type. Another type of polynomial interpolation is given by \textit{Hermite interpolation} which is based on the interpolation of point values and derivatives of a function (cf. Section 2.1.5 in Stöer and Bulirsch (2002)). As far as the construction of \textit{Hermitian finite elements} and associated \textit{Hermitian finite element spaces} is concerned, we refer to chapter 2.3 in Ciarlet (2002).

The accuracy of finite element approximations \( u_h \in V_h \) of the solution \( u \in V \) of (2.4.1) can be determined by \textit{a priori error estimates} of the global discretization error \( u - u_h \) within the framework of \textit{interpolation in Sobolev spaces}.

\textbf{(2.4.68) Definition:} Let \( \mathcal{T}_h(\Omega) \) be a simplicial triangulation of \( \Omega \subset \mathbb{R}^d \), and let \((T, P_T, \Sigma_T), T \in \mathcal{T}_h(\Omega)\), where \( P_T = P_k(T) \) such that \( \dim(P_T) = n_k \) and \( \Sigma_T = \{ p(x_i \mid x_i \in L_k(T), 1 \leq i \leq n_k \} \) be a simplicial Lagrangian finite element of type \((k)\). Moreover, let \( \varphi_{T}^{(i)}, 1 \leq i \leq n_k \), be the basis functions associated with the nodal points \( x_i \). Then, \( I_T : V \cap C(T) \to P_T \) as given by

\[
I_T v := \sum_{i=1}^{n_k} v(x_i) \varphi_{T}^{(i)} , \quad v \in V \cap C(T)
\]

is called the \textit{local interpolation operator}. If \( V_h \) is the associated simplicial Lagrangian finite element space, the operator \( I_h : V \cap C(\Omega) \to V_h \) given by

\[
I_h v|_T := I_T v|_T , \quad T \in \mathcal{T}_h(\Omega)
\]

is called the \textit{global interpolation operator}.

Now, if \( u \in V \) is the solution of (2.4.1) such that \( u \in V \cap H^{k+1}(\Omega) \) for sufficiently large \( k \in \mathbb{N} \) such that \( H^{k+1}(\Omega) \subset C(\Omega) \), then \( I_h u \) is well defined and Céa’s Lemma (cf. Theorem 2.4.10) implies

\[
\| u - u_h \|_{1, \Omega} \leq \frac{C}{\alpha} \| u - I_h u \|_{1, \Omega} .
\]

In order to derive an upper bound for the interpolation error \( \| u - I_h u \|_{1, \Omega} \), we use that in view of the definition of \( I_h \) and due to \( P_T = P_k(T) \subset H^1(T), T \in \mathcal{T}_h(\Omega) \), there holds

\[
\| u - I_h u \|_{1, \Omega} = \left( \sum_{T \in \mathcal{T}_h(\Omega)} \| u - I_T u \|^2_{1, T} \right)^{1/2},
\]

i.e., we obtain an estimate of the global discretization error via an estimate of the global interpolation error by means of estimates of the local interpolation error \( \| u - I_T u \|_{1, T} \), \( T \in \mathcal{T}_h(\Omega) \). Due to the affine equivalence of simplicial Lagrangian finite elements (cf. Lemma (2.4.42)), it suffices to provide such an estimate for the reference element \( \hat{T} \). Moreover, we need estimates of affine transformed functions in Sobolev spaces.
(2.4.71) **Lemma:** Let $\hat{T}$ be the simplicial reference element and $T = F_T(\hat{T})$ such that $F_T(\hat{x}) = B_T \hat{x} + b_T$, $B_T \in \mathbb{R}^{d \times d}$, $b_T \in \mathbb{R}^d$, and let $v \in H^m(T)$, $m \in \{0,1\}$. For $\hat{v} := v \circ F_T$ there holds $\hat{v} \in H^m(\hat{T})$, and there exists a constant $C > 0$ only depending on $d$ and $m$ such that

\[
|\hat{v}|_{m,\hat{T}} \leq C \|B_T\|^m |\det(B_T)|^{-1/2} |v|_{m,T},
\]

(2.4.72a)

\[
|v|_{m,T} \leq C \|B_T^{-1}\|^m |\det(B_T)|^{1/2} |\hat{v}|_{m,\hat{T}}.
\]

(2.4.72b)

Further, denoting by $h_T$ the diameter of $T$ and by $\rho_T$ the diameter of the largest ball contained in $T$, i.e.,

\[
\rho_T := \sup\{\text{diam}(K) \mid K \text{ is a ball such that } K \subset T\},
\]

there holds

\[
\|B_T\| \leq \frac{h_T}{\rho_T}, \quad \|B_T^{-1}\| \leq \frac{h_T}{\rho_T},
\]

(2.4.73a)

\[
|\det(B_T)| \leq \frac{|T|}{|\hat{T}|}.
\]

(2.4.73b)

**Proof.** The proof of (2.4.72a) by taking into account

\[
|D^\alpha \hat{v}(\hat{x})| \leq \|D\hat{v}(\hat{x})\| = \sup_{\|\xi\| \leq 1} |D\hat{v}(\hat{x})\xi|,
\]

where $D\hat{v}(\hat{x})$ stands for the total differential of $\hat{v}$, and by means of the chain rule

\[
D\hat{v}(\hat{x})\xi = Dv(x)(B\xi)
\]

as well as the transformation rule for multiple integrals

\[
\int_{\hat{T}} \|D(v(F\hat{x}))(\hat{x})\|^2 \, d\hat{x} = |\det(B_T^{-1})| \int_{T} \|Dv(x)\|^2 \, dx.
\]

The proof of (2.4.72b) can be done analogously. The proofs of (2.4.73a),(2.4.73b) are left as exercises. \hfill \Box

The following **Lemma of Bramble-Hilbert** is an important tool in the a priori error analysis of finite element methods.

(2.4.74) **Theorem:** Let $\Omega \subset \mathbb{R}^d$ be a simply connected Lipschitz domain. For $k \in \mathbb{N}$ let $\ell \in H^k(\Omega)^*$ such that $P_k(\Omega) \subset \text{Ker}(\ell)$. Then, there exists a positive constant $C$ depending on $\Omega$ such that for all $u \in H^{k+1}(\Omega)$

\[
|\ell(u)| \leq C \|\ell\|_{H^k(\Omega)^*} \|u\|_{k+1,\Omega}.
\]

(2.4.75)
Proof. For $p \in P_k(\Omega)$ there holds
\[ |\ell(u)| = |\ell(u + p)| \leq \|\ell\|_{H^k(\Omega)^*} \|u\|_{k+1,\Omega}, \]
where $[u] := \{w \in H^{k+1}(\Omega) \mid w - u \in P_k(\Omega)\}$. We claim the existence of a positive constant $C$ depending on $\Omega$ such that
\[ \|u\|_{k+1,\Omega} \leq C \|u\|_{k+1,\Omega}, \]
which together with the previous inequality proves the assertion. For the proof of (2.4.76) let $n_k := \dim(P_k(\Omega))$ and let $\ell_i, 1 \leq i \leq n_k$, be a basis of the dual space of $P_k(\Omega)$ such that
\[ \ell_i(p) = 0, \quad 1 \leq i \leq n_k, \quad p \in P_k(\Omega) \implies p = 0. \]
The Hahn-Banach extension theorem implies the existence of functionals $\tilde{\ell}_i \in H^{k+1}(\Omega)^*, 1 \leq i \leq n_k$, with $\tilde{\ell}_i|_{P_k(\Omega)} = \ell_i$. By a contradiction argument we show
\[ \|u\|_{k+1,\Omega} \leq C(\Omega) \left( |u|_{k+1,\Omega} + \sum_{i=1}^{n_k} |\tilde{\ell}_i(u)| \right), \quad u \in H^{k+1}(\Omega). \]
Indeed, if (2.4.78) does not hold true, there exists a sequence $\{u_k\}_N, u_k \in H^{k+1}(\Omega), k \in \mathbb{N}$, such that
\[ \|u_k\|_{k+1,\Omega} = 1, \quad k \in \mathbb{N}, \quad \lim_{k \to \infty} \left( |u_k|_{k+1,\Omega} + \sum_{i=1}^{n_k} |\tilde{\ell}_i(u_k)| \right) = 0. \]
Since the sequence $\{u_k\}_N$ is bounded in $H^{m+1}(\Omega)$ and since $H^{k+1}(\Omega)$ is compactly embedded in $H^k(\Omega)$, there exist a subsequence $m' \subset \mathbb{N}$ and a function $u \in H^m(\Omega)$ such that
\[ \|u_k - u\|_{k,\Omega} \to 0 \quad (k \in m', \quad k \to \infty). \]
On the other hand, due to (2.4.79) we have
\[ \|u_k\|_{k+1,\Omega} \to 0 \quad (k \in m', \quad k \to \infty). \]
Hence, in view of (2.4.80) it follows that $u_k \to u (k \in m', \quad k \to \infty)$ in $H^{k+1}(\Omega)$ such that
\[ \|D^\alpha u\|_{0,\Omega} = \lim_{k \to \infty} \|D^\alpha u_k\|_{0,\Omega} = 0, \quad |\alpha| = k + 1. \]
Consequently, we have $D^\alpha u = 0, |\alpha| = k + 1$. The fact that $\Omega$ is simply connected implies $u \in P_k(\Omega)$. Due to (2.4.79) we obtain
\[ \tilde{\ell}_i(u) = \lim_{k \to \infty} \tilde{\ell}_i(u_k) = 0, \quad 1 \leq i \leq n_k. \]
Since $\tilde{\ell}_i(u) = \ell_i(u), \quad (2.4.77)$ implies $u = 0$ thus contradicting (2.4.79). \qed

The combination of Lemma (2.4.71) and the Lemma of Bramble-Hilbert (2.4.74) results in the following estimate of the local interpolation error:
Lemma: Let \((\hat{T}, P_k(\hat{T}), \hat{\Sigma})\) be the simplicial Lagrangian reference element of type \((k)\), where \(k \in \mathbb{N}\) is such that \(H^{k+1}(\hat{T})\) is continuously embedded in \(C(\hat{T})\). Then, for all elements \((T, P_k(T), \Sigma_T)\) that are affine equivalent to \((\hat{T}, P_k(\hat{T}), \hat{\Sigma})\) there holds

\[
|v - I_T v|_{m,T} \leq C \frac{h_T^{k+1}}{\rho_T} |v|_{k+1,T}, \quad v \in H^{k+1}(T), \quad m \in \{0,1\},
\]

where \(C\) is a positive constant depending only on the reference element.

Proof. The proof relies on the polynomial preserving property of the local interpolation operator \(\hat{I}_{\hat{T}}\), i.e., \(\hat{I}_{\hat{T}} = \hat{p}\). Denoting by \(\hat{I}\) the continuous embedding \(\hat{I} : H^{k+1}(\hat{T}) \to H^1(\hat{T})\), it follows that for \(\hat{v} \in H^{k+1}(\hat{T})\) and \(\hat{p} \in P_k(\hat{T})\) there holds

\[
\hat{v} - \hat{I}_{\hat{T}} \hat{v} = (\hat{I} - \hat{I}_{\hat{T}})(\hat{v} + \hat{p}).
\]

Hence, the Lemma of Bramble-Hilbert (2.4.74) implies

\[
|\hat{v} - \hat{I}_{\hat{T}} \hat{v}|_{m,\hat{T}} \leq C |\hat{v}|_{k+1,\hat{T}}.
\]

Moreover, due to Lemma (2.4.71) and in view of \(\hat{v} - \hat{I}_{\hat{T}} \hat{v} = v - I_T v\) we obtain

\[
|v - I_T v|_{m,T} \leq C \|B_T^{-1}\|^m |\det(B_T)|^{1/2} |\hat{v} - \hat{I}_{\hat{T}} \hat{v}|_{m,\hat{T}}
\]

and

\[
|\hat{v}|_{k+1,\hat{T}} \leq C \|B_T\|^{k+1} |\det(B_T)|^{-1/2} |v|_{k+1,T}.
\]

Taking account of (2.4.73a),(2.4.73b), the combination of the previous estimates gives the assertion.

Now, let \(\mathcal{H}\) be a null sequence of positive real numbers and let \(\{T_h(\Omega)\}_{\mathcal{H}}\) be a family of geometrically conforming simplicial triangulations of \(\Omega\). Then, (2.4.82) shows that the estimate (2.4.70) of the global interpolation error depends on the triangulations in such a way that the upper bound in (2.4.70) is unbounded, if \(h_T/\rho_T \to \infty, T \in T_h(\Omega)\) as \(h \to 0\). In order to exclude this case, we have to impose a further condition on the family of triangulations \(\{T_h(\Omega)\}_{\mathcal{H}}\).

Definition: The family \(\{T_h(\Omega)\}_{\mathcal{H}}\) of geometrically conforming simplicial triangulations is called regular, if there exists a constant \(\sigma > 0\), independent of \(h_T, T \in T_h(\Omega)\), such that for all \(T \in T_h(\Omega), h \in \mathcal{H}\)

\[
\frac{h_T}{\rho_T} \leq \sigma.
\]

We are now in a position to provide an a priori estimate of the global discretization error in the \(\| \cdot \|_{1,\Omega}\) norm.
Theorem: Assume that \( \{ T_h(\Omega) \} \) is a regular family of geometrically conforming simplicial triangulations of \( \Omega \). Further, suppose that \( \{ V_h \} \) is the associated family of finite element spaces on the basis of simplicial Lagrangian finite elements of type \( (k) \) where \( k \in \mathbb{N}_0 \) is such that the embeddings \( H^{k+1}(\Omega) \rightarrow C(\Omega) \) and \( H^{k+1}(\hat{T}) \rightarrow C(\hat{T}) \) are continuous. If \( u \in V \) is the solution of (2.4.1) such that \( u \in V \cap H^{k+1}(\Omega) \) and if \( u_h \in V_h, h \in \mathcal{H} \), are the solutions of (2.4.2), then there exists a constant \( C > 0 \), depending only on the local geometry of the triangulations, such that

\[
\| u - u_h \|_{1,\Omega} \leq C h^k |u|_{k+1,\Omega}.
\]

Proof. The proof is an immediate consequence of (2.4.69), (2.4.70), (2.4.82) and (2.4.84).

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: The variational equation (2.4.1) with

\[
\ell(v) = (f,v)_{0,\Omega}, \quad f \in L^2(\Omega)
\]

is said to be \( k \)-regular, \( k \geq 2 \), if its solution satisfies \( u \in V \cap H^{k+1}(\Omega) \) and if there exists a constant \( C > 0 \) such that

\[
\| u \|_{k,\Omega} \leq C \| f \|_{0,\Omega}.
\]

Corollary: Under the assumptions of Theorem (2.4.85) suppose that the solution \( u \) of (2.4.1) is \((k + 1)\)-regular in the sense of Definition (2.4.87). Then, there exists a constant \( C > 0 \), depending only on the local geometry of the triangulations, such that

\[
\| u - u_h \|_{1,\Omega} \leq C h^k \| f \|_{0,\Omega}.
\]

Proof. The proof follows directly from (2.4.86) and (2.4.88).

In view of \( \| v \|_{0,\Omega} \leq \| v \|_{1,\Omega}, v \in V \), and under the assumptions of Theorem (2.4.85), it follows readily from (2.4.86) that \( \| u - u_h \|_{0,\Omega} = O(h^k) \). However, the interpolation estimate \( \| I_h u - u \|_{0,\Omega} = O(h^{k+1}) \) from Lemma 2.4.81 suggests that this estimate is not optimal. Indeed, under additional assumptions the optimal order \( O(h^{k+1}) \) can be verified. The theoretical background for such an optimal a priori estimate of the global discretization error in the \( L^2 \)-norm will be provided by the Lemma of Aubin-Nitsche, which is also known as Nitsche’s trick.

Theorem: Let \( V \) be a Hilbert space with inner product \( (\cdot,\cdot)_V \) and let \( a(\cdot,\cdot) : V \times V \rightarrow \mathbb{R} \) be a bounded, \( V \)-elliptic bilinear form and \( \ell \in V^* \). Moreover, let \( H \) be another Hilbert space with inner product \( (\cdot,\cdot)_H \) such that
2.4 Conforming finite element methods

$V$ is continuously and densely embedded in $H$ and let $V_h \subset V$. Finally, let $u \in V$ and $u_h \in V_h$ be the unique solutions of (2.4.1) and (2.4.2). Then, there holds:

(i) For any $g \in H$ the adjoint variational equation

\[(2.4.91) \quad a(v,z_g) = (g,v)_H \quad , \quad v \in V\]

has a unique solution $z_g \in V$.

(ii) There exists a positive constant $C$ such that for the global discretization error $u - u_h$ there holds

\[(2.4.92) \quad \|u - u_h\|_H \leq C \|u - u_h\|_V \left( \sup_{g \in H} \frac{1}{\|g\|_H} \inf_{\varphi_h \in V_h} \|z_g - \varphi_h\|_V \right) .\]

Proof. In view of $(g,v)_H = g(v), g \in H \subset V^*$, the Lemma of Lax-Milgram (2.3.12) implies that the adjoint variational equation (2.4.91) admits a unique solution which proves (i). For the proof of (ii) let $v = u - u_h \in V$ in (2.4.91). Then, we have

\[a(u - u_h, z_g) = (g, u - u_h)_H .\]

On the other hand, choosing $v = v_h = \varphi_h \in V_h$ in (2.4.1),(2.4.2), we obtain

\[a(u - u_h, \varphi_h) = 0 \quad , \quad \varphi_h \in V_h .\]

The combination of both equations yields

\[a(u - u_h, z_g - \varphi_h) = (g, u - u_h)_H ,\]

whence

\[(g, u - u_h)_H \leq C \|u - u_h\|_V \inf_{\varphi_h \in V_h} \|z_g - \varphi_h\|_V .\]

Finally, the estimate (2.4.92) results from

\[\|u - u_h\|_H = \sup_{g \in H, g \neq 0} \frac{|(g, u - u_h)_H|}{\|g\|_H} .\]

The following a priori estimate of the global discretization error in the $L^2$ norm is a direct consequence of the Lemma of Aubin-Nitsche (2.4.90):

*(2.4.93) Theorem:* Under the assumptions of Theorem (2.4.85) suppose that the adjoint variational equation (2.4.91) is $(k+1)$-regular. Then, there exists a positive constant $C$, depending only on the local geometry of the triangulations, such that

\[(2.4.94) \quad \|u - u_h\|_{0,\Omega} \leq C h^{k+1} |u|_{k+1,\Omega} .\]
Proof. The proof follows from the Lemma of Aubin-Nitsche (2.4.90) with \( H = L^2(\Omega) \). In particular, due to the Corollary of Theorem 2.4.85 and the \((k + 1)\) regularity of the adjoint variational equation we have

\[
\inf_{\varphi_h \in V_h} \| z_g - \varphi_h \|_{1,\Omega} \leq \| z_g - I_h z_g \|_{1,\Omega} \leq C h \| z_g \|_{k+1,\Omega} \leq C h \| g \|_{0,\Omega} .
\]

Hence, from (2.4.92) we deduce

\[
\| u - u_h \|_{0,\Omega} \leq C h \| u - u_h \|_{1,\Omega} .
\]

Theorem (2.4.85) allows a further estimate of the right-hand side which implies (2.4.94).

In Lemma (2.4.81), the error with respect to the approximation of a Sobolev space function by its finite element interpolant is measured in a coarser Sobolev norm than the norm associated with the Sobolev space where the function lives in. In general, the converse does not hold true, i.e., it is not possible to estimate the norm of a Sobolev space function from above in a coarser norm. However, for finite element functions such estimates can be provided which are known as inverse estimates.

(2.4.95) Theorem: Let \( \{ T_h(\Omega) \}_H \) be a regular family of geometrically conforming simplicial triangulations and let \( \{ V_h \}_H \) be the associated family of simplicial Lagrangian finite element spaces based on simplicial Lagrangian finite elements of type \((k)\) where \( k \in \mathbb{N}_0 \) is such that \( H^{k+1}(\Omega) \) is continuously embedded in \( C(\Omega) \). For \( \ell \in \mathbb{N} \) such that \( 0 \leq k + 1 \leq \ell \) there exists a constant \( C > 0 \), depending on \( k \) and \( \ell \) as well as on \( \sigma \) in (2.4.84), such that for \( v_h \in V_h \) there holds

\[
\| v_h \|_{\ell,\Omega} \leq C h^{k+1-\ell} \| v_h \|_{k+1,\Omega} .
\]

Proof. In a first step, for the reference element we show

\[
\| \hat{v} \|_{\ell,\hat{K}} \leq C \| \hat{v} \|_{k+1,\hat{K}} , \quad \hat{v} \in \hat{P}_K .
\]

We assume \( \hat{P}_K = \text{span}\{ \hat{p}_1, ..., \hat{p}_{n_k} \} \) and choose \( (\hat{f}_i)_{i=1}^{n_k} \) as a basis of the dual space, i.e.,

\[
\hat{f}_i : \hat{P}_K \to \mathbb{R} , \quad 1 \leq i \leq n_k ,
\]

\[
\hat{p}_j \mapsto \hat{f}_i(\hat{p}_j) = \delta_{ij} , \quad 1 \leq i, j \leq n_k .
\]

Then, we have the equivalence of the \( \| \cdot \|_{k+1,\hat{K}} \) norm and the norm given by

\[
\| \| \hat{v} \|_{k+1,\hat{K}} := |\hat{v}|_{k+1,\hat{K}} + \sum_{i=1}^{n_k} |\hat{f}_i(\hat{v})| .
\]

Due to the finite dimension of \( V_h \) this also holds true for the norms \( \| \cdot \|_{k+1,\hat{K}} \) and \( \| \cdot \|_{\ell,\hat{K}} \). We thus obtain the estimate
\[ |\hat{v}|_{l,K} \leq |\hat{v} - \hat{N}_K \hat{v}|_{l,K} + |\hat{N}_K \hat{v}|_{l,K} \leq \]
\[ \leq ||\hat{v} - \hat{N}_K \hat{v}||_{l,K} \leq C ||\hat{v} - \hat{N}_K \hat{v}||_{k+1,K} = \]
\[ = C \left( |\hat{v} - \hat{N}_K \hat{v}|_{k+1,K} + \sum_{i=1}^{n_K} |\hat{f}_i(\hat{v} - \hat{N}_K \hat{v})| \right) \leq \]
\[ \leq C \left( |\hat{v}|_{k+1,K} + |\hat{N}_K \hat{v}|_{k+1,K} \right), \]

which proves (2.4.97). Finally, the inverse estimate follows by a scaling argument according to

\[ |v|_{l,K} \leq C \|B^{-1}_{K}\|_{l} |\det(B_k)|^{1/2} |\hat{v}|_{l,K} \leq \]
\[ \leq C \|B^{-1}_{K}\|_{l} |\det(B_k)|^{1/2} |\hat{v}|_{k+1,K} \leq \]
\[ = C \|B^{-1}_{K}\|_{l} \|B_K\|^{k+1} |v|_{k+1,K} \leq \]
\[ \leq C h_{K}^{k+1-\ell} |v|_{k+1,K}. \]

As far as a priori estimates in the \( L^\infty \) norm are concerned, using an inverse estimate we prove the following result which, however, is not optimal:

\[(2.4.98) \quad \textbf{Theorem:} \quad \text{Under the assumptions of Theorem (2.4.95) there exists a positive constant } C, \text{ depending only on the local geometry of the triangulations, such that} \]

\[(2.4.99) \quad \|u - u_h\|_{\infty,\Omega} \leq C h^k |u|_{k+1,\Omega}. \]

\textbf{Proof.} Using the global interpolation operator \( I_h \), we split the error according to

\[ \|u - u_h\|_{\infty,\Omega} \leq \|u - I_h u\|_{\infty,\Omega} + \|I_h u - u_h\|_{\infty,\Omega} \]

and estimate the two terms on the right-hand side separately.

(i) In order to estimate \( \|u - I_h u\|_{\infty,\Omega} \) we use the affine equivalence. For \( \hat{v} \in H^{k+1}(\hat{K}) \), the Lemma of Bramble-Hilbert (2.4.74) implies

\[ \|\hat{v} - \hat{I}_K \hat{v}\|_{\infty,\hat{K}} \leq C |\hat{v}|_{k+1,\hat{K}}. \]

A scaling argument yields

\[ \|u - I_K u\|_{\infty,K} \leq \|\hat{u} - \hat{I}_K \hat{u}\|_{\infty,\hat{K}} \leq C |\hat{u}|_{k+1,\hat{K}} \leq C h^k |u|_{k+1,K}, \]

whence
(2.4.100) \[ \| u - I_h u \|_{\infty, \Omega} = \max_{K \in T_h} \| u - I_K u \|_{\infty, K} \leq \]
\[ \leq C h^k \max_{K \in T_h} |u|_{k+1, K} \leq C h^k |u|_{k+1, \Omega}. \]

(ii) For the estimation of \( \| I_h u - u_h \|_{\infty, \Omega} \), the Sobolev embedding theorem (1.2.11) and the inverse inequality (2.4.96) imply
\[ \| u_h - I_h u \|_{\infty, \Omega} \leq C \| u_h - I_h u \|_{1, \Omega} \leq C h^{-1} \| u_h - I_h u \|_{0, \Omega}. \]

Taking advantage of the a priori estimate (2.4.94) and the interpolation estimate \( \| u - I_h u \|_{0, \Omega} \leq h^{k+1} |u|_{k+1, \Omega} \) it follows that
\[ \| u_h - I_h u \|_{0, \Omega} \leq C \| u_h - u \|_{0, \Omega} + \| u - I_h u \|_{0, \Omega} \leq C h^{k+1} |u|_{k+1, \Omega}, \]
and hence,
\[ \| u_h - I_h u \|_{0, \Omega} \leq C h^k |u|_{k+1, \Omega}. \]

The estimates (2.4.100) and (2.4.101) allow to conclude. \( \square \)

The derivation of an optimal a priori estimate in the \( L^\infty \) norm requires the use of \textit{weighted Sobolev norms.} (cf. Chapter 3.3. in Ciarlet (2002)).

(2.4.102) \textbf{Theorem:} Assume that the conditions of Theorem (2.4.98) hold true for \( k = 1 \). Further, suppose that the solution \( u \) of (2.4.1) satisfies \( u \in V \cap W^{2, \infty}(\Omega) \). Then, there exists a positive constant \( C \), depending only on the local geometry of the triangulations, such that
\[ \| u - u_h \|_{\infty, \Omega} \leq C h^2 |\log h|^{3/2} \| D^2 u \|_{\infty, \Omega}. \]

We note that the estimate (2.4.103) is optimal in \( h \).

\textbf{Proof.} We refer to Theorem 3.3.7 in Ciarlet (2002). \( \square \)

2.5 Numerical quadrature and nonconforming methods

We consider the variational equation
\[ a(u, v) = \ell(v), \quad v \in V \subset H^1(\Omega), \]
and assume that the assumptions of the Lemma of Lax-Milgram (2.3.12) are satisfied. In particular, (2.5.1) admits a unique solution \( u \in V \). Nonconformity in the finite element approximation of (2.5.1) may occur due to several reasons. One is the construction of approximations \( u_h \) in subspaces \( V_h \subset V \) by means of bilinear forms \( a_h(\cdot, \cdot) : V_h \times V_h \to \mathbb{R} \) and functionals \( \ell_h : V_h \to \mathbb{R} \) that are not simply the restrictions of \( a(\cdot, \cdot) \) and \( \ell(\cdot) \) to \( V_h \). Another one is to compute approximations \( u_h \) in finite dimensional spaces \( V_h \) which are
not subspaces of $V$. In both cases, besides the approximation error we have additionally to consider a consistency error. A priori estimates of the global discretization error are addressed by Strang’s first and second lemma.

We suppose that $H$ is a null sequence of positive real numbers and that \{V_h\}_{h \in H} is a family of conforming finite element spaces $V_h \subset V, h \in H$. We approximate the bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ and the functional $\ell(\cdot) : V \to \mathbb{R}$ in (2.5.1) by bounded bilinear forms and bounded linear functionals

$$\tag{2.5.2} a_h(\cdot, \cdot) : V_h \times V_h \to \mathbb{R}, \quad \ell_h(\cdot) : V_h \to \mathbb{R}, \quad h \in H.$$ 

On this basis, we approximate (2.5.1) by the finite dimensional variational equation

$$\tag{2.5.3} a_h(u_h, v_h) = \ell_h(v_h) \quad , \quad v_h \in V_h , \quad h \in H.$$ 

\textbf{(2.5.4) Definition:} The sequence \{a_h(\cdot, \cdot)\}_{h \in H} of bilinear forms $a_h(\cdot, \cdot) : V_h \times V_h \to \mathbb{R}$ is said to be uniformly $V_h$-elliptic, if there exists a positive constant $\alpha$ such that uniformly in $h \in H$ there holds

$$\tag{2.5.5} a_h(u_h, u_h) \geq \alpha \|u_h\|_V^2 \quad , \quad u_h \in V_h.$$ 

Under the assumption of uniform $V_h$-ellipticity, the variational equations (2.5.3) admit unique solutions $u_h \in V_h$. The following result, known as Strang’s first lemma, can be viewed as a generalization of Céa’s Lemma (2.4.10).

\textbf{(2.5.6) Theorem:} Let \{a_h(\cdot, \cdot)\}_{h \in H} be a uniformly $V_h$-elliptic sequence of bilinear forms and let $u \in V$ and $u_h \in V_h, h \in H$ be the unique solutions of (2.5.1) and (2.5.3). Then, there exists a constant $C > 0$, independent of $h \in H$, such that for $h \in H$

$$\tag{2.5.7} \|u - u_h\|_V \leq C \left( \inf_{v_h \in V_h} \|u - v_h\|_V + \right.$$ 

$$+ \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_V} + \sup_{w_h \in V_h} \frac{\|\ell(w_h) - \ell_h(w_h)\|}{\|w_h\|_V} \right).$$

\textbf{Proof.} Using the uniform $V_h$-ellipticity, for $v_h \in V_h$ we obtain

$$\alpha \|u_h - v_h\|_V^2 \leq a_h(u_h - v_h, u_h - v_h) \pm a(u - v_h, u_h - v_h) =$$

$$= a(u - v_h, u_h - v_h) + \left( a(v_h, u_h - v_h) - a_h(v_h, u_h - v_h) \right) +$$

$$+ \left( \ell_h(u_h - v_h) - \ell(u_h - v_h) \right).$$

In view of the boundedness of the bilinear forms $a(\cdot, \cdot)$, i.e., $|a(u, v)| \leq C\|u\|_V\|v\|_V$, division by $\|u_h - v_h\|_V$ results in
The triangle inequality yields
\[ \| u - u_h \|_V \leq \| u - v_h \|_V + \| u_h - v_h \|_V. \]

We conclude by estimating the second term on the right-hand side by means of (2.5.8).

Strang’s first lemma shows that the upper bound for the global discretization error consists of two parts, the approximation error
\[ \inf_{v_h \in V_h} \| u - v_h \|_V, \]
and the consistency error
\begin{align*}
\inf_{v_h \in V_h} \sup_{w_h \in V_h} & \left| \frac{a(v_h, w_h) - a_h(v_h, w_h)}{\| w_h \|_V} \right|, \\
\sup_{w_h \in V_h} & \left| \ell(w_h) - \ell_h(w_h) \right| \| w_h \|_V .
\end{align*}

We now assume that \( \{ V_h \}_{h \in \mathcal{H}} \) is a sequence of finite dimensional linear spaces equipped with grid dependent norms \( \| \cdot \|_h \), \( h \in \mathcal{T}_h \), on \( V_h + V \). Further, we suppose that the functional \( \ell \) is well defined on \( V_h + V \).

We consider a sequence of approximating bilinear forms
\[ a_h(\cdot, \cdot) : (V_h + V) \times (V_h + V) \to \mathbb{R}. \]
We assume that the sequence \( \{ a_h(\cdot, \cdot) \}_{h \in \mathcal{H}} \) is uniformly \( V_h \)-elliptic, i.e.,
\[ a_h(v_h, v_h) \geq \tilde{\alpha} \| v_h \|_{V_h}^2, \quad v_h \in V_h, \quad h \in \mathcal{H}, \]
and uniformly bounded on \( V_h + V \), i.e.,
\[ |a_h(u, v)| \leq C \| u \|_h \| v \|_h, \quad u, v \in V_h + V, \quad h \in \mathcal{H}, \]
where \( \alpha \) and \( C \) are positive constants independent of \( h \in \mathcal{H} \).

We approximate (2.5.1) by the finite dimensional variational equation
\[ a_h(u_h, v_h) = \ell(v_h), \quad v_h \in V_h, \quad h \in \mathcal{H}. \]

Strang’s second lemma provides an upper bound for the global discretization error \( u - u_h \) in the grid dependent norm \( \| \cdot \|_h \).
Theorem: Let \( \{a_h(\cdot, \cdot)\}_{h \in \mathcal{H}} \) be a uniformly bounded and uniformly \( V_h \)-elliptic sequence of bilinear forms and assume that \( u \in V \) and \( u_h \in V_h, h \in \mathcal{H} \), are the unique solutions of the variational equations (2.5.1) and (2.5.14). Then, there exists a constant \( C > 0 \), independent of \( h \in \mathcal{H} \), such that

\[
\|u - u_h\|_h \leq C \left( \inf_{v_h \in V_h} \|u - v_h\|_h + \sup_{w_h \in V_h} \frac{|a_h(u, w_h) - \ell(w_h)|}{\|w_h\|_h} \right).
\]

Proof. The uniform \( V_h \)-ellipticity implies that for \( v_h \in V_h \)

\[
\alpha \|u_h - v_h\|_h^2 \leq a_h(u_h - v_h, u_h - v_h) = a_h(u - v_h, u_h - v_h) + \left( \ell(u_h - v_h) - a_h(u, u_h - v_h) \right),
\]

and hence, by the uniform boundedness we obtain

\[
\|u_h - v_h\|_h \leq a_h(u_h - v_h, u_h - v_h) \leq C \|u - v_h\|_h + \frac{\|\ell(u_h - v_h) - a_h(u, u_h - v_h)\|}{\|u_h - v_h\|_h}
\]

(2.5.17)

\[
\leq C \|u - v_h\|_h + \sup_{w_h \in V_h} \frac{\|\ell(w_h) - a_h(u, w_h)\|}{\|w_h\|_h}.
\]

Using the triangle inequality

\[
\|u - u_h\|_h \leq \|u - v_h\|_h + \|u_h - v_h\|_h,
\]

the assertion follows from (2.5.17).

As in case of Strang’s first lemma, the upper bound for the global discretization error consists of the approximation error

\[
\inf_{v_h \in V_h} \|u - v_h\|_h
\]

(2.5.18)

and the consistency error

\[
\sup_{w_h \in V_h} \frac{\|\ell(w_h) - a_h(u, w_h)\|}{\|w_h\|_h}.
\]

(2.5.19)

As an application of Strang’s first lemma, we consider the approximation of (2.5.1) for

\[
a(u, v) := \sum_{i,j=1}^{d} \int_{\Omega} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx, \quad u, v \in V \subset H^1(\Omega), \tag{2.5.20a}
\]

\[
\ell(v) := \int_{\Omega} f v \, dx, \quad v \in V, \tag{2.5.20b}
\]
evaluating the integrals in (2.5.20a) and (2.5.20b) by numerical integration using a quadrature formula. We assume \( \Omega \subset \mathbb{R}^d \) to be a bounded polyhedral domain. As far as the data of the problem are concerned, we suppose that for some \( p \geq 2 \) and \( k \geq d/p \) there holds
\[
(2.5.21) \quad a_{ij} \in W^{k,\infty}(\Omega) , \quad 1 \leq i, j \leq d , \quad f \in W^{k,p}(\Omega) ,
\]
and that \((a_{ij})_{i,j=1}^d\) is a symmetric and uniformly positive definite matrix-valued function. Moreover, we assume that the solution \( u \) of (2.5.1) satisfies
\[
(2.5.22) \quad u \in V \cap H^{k+1}(\Omega) .
\]

(2.5.23) **Definition:** Let \( T_h(\Omega) \) be a geometrically conforming triangulation of \( \Omega \) and
\[
(2.5.24) \quad b^T_i \in T , \quad \omega^T_i \in \mathbb{R} , \quad T \in T_h(\Omega) , \quad 1 \leq i \leq L , \quad L \in \mathbb{N} .
\]
Given a function \( \varphi : \overline{\Omega} \to \mathbb{R} \), for the approximation of the integral \( \int_T \varphi dx \), we refer to
\[
(2.5.25) \quad Q_T(\varphi) := \sum_{i=1}^L \omega^T_i \varphi(b^T_i)
\]
as a quadrature formula. The points \( b^T_i \) and the numbers \( \omega^T_i , 1 \leq i \leq L \), are called the nodes and the weights of the quadrature formula and the quantity
\[
(2.5.26) \quad E_T(\varphi) := \int_T \varphi dx - Q_T(\varphi)
\]
is said to be the quadrature error.

If \( V_h \subset V \) is a simplicial Lagrangian finite element space of type \((k)\), the affine equivalence of the finite elements allows the construction of (2.5.25) based on a quadrature formula
\[
(2.5.27) \quad \hat{Q}_{\hat{T}}(\hat{\varphi}) := \sum_{i=1}^L \hat{\omega}^T_i \hat{\varphi}(\hat{b}^T_i)
\]
for the reference element \( \hat{T} \) with nodes \( \hat{b}^T_i \in \hat{T} \) and weights \( \hat{\omega}^T_i , 1 \leq i \leq L \).

(2.5.28) **Lemma:** Let \( \hat{Q}_{\hat{T}} \) be a quadrature formula for the reference element \( \hat{T} \) of the form (2.5.27). For an element \( T \in T_h(\Omega) \) such that \( T = F_T(\hat{T}) \), where \( F_T(\hat{x}) = B_T \hat{x} + b_T, \hat{x} \in \hat{T} \), we obtain a quadrature formula \( Q_T \) with nodes and weights given by
\[
(2.5.29) \quad b^T_i = F_T(\hat{b}^T_i) , \quad \omega^T_i = \det(B_T) \hat{\omega}^T_i , \quad 1 \leq i \leq L ,
\]
which satisfies
\[
(2.5.30) \quad E_T(\varphi) = \det(B_T) \hat{E}_{\hat{T}}(\hat{\varphi}) , \quad \hat{\varphi} = \varphi \circ F_T .
\]
Proof. The assertions readily follow from
\[ \int_T \varphi \, dx = \det(B_T) \int_{\hat{T}} \hat{\varphi} \, d\hat{x}. \]
\[ \square \]

On the basis of a quadrature formula of the form (2.5.25) we define bilinear forms \( a_h(\cdot, \cdot) : V_h \times V_h \to \mathbb{R} \) and linear functionals \( \ell_h(\cdot) : V_h \to \mathbb{R}, h \in \mathcal{H} \), according to
\begin{align*}
(2.5.31a) \quad a_h(u_h, v_h) &:= \sum_{T \in \mathcal{T}_h(\Omega)} \sum_{\ell=1}^L \omega_{\ell, T} \sum_{i,j=1}^d (a_{ij} \frac{\partial u_h}{\partial x_i} \frac{\partial v_h}{\partial x_j})(b_{\ell, T}) , \\
(2.5.31b) \quad \ell_h(v_h) &:= \sum_{T \in \mathcal{T}_h(\Omega)} \sum_{\ell=1}^L \omega_{\ell, T}^K (fv_h)(b_{\ell, T}^K). 
\end{align*}

The following result provides sufficient conditions for the uniform \( V_h \)-ellipticity of the sequence \( \{a_h(\cdot, \cdot)\}_{h \in \mathcal{H}} \).

(2.5.32) Theorem: Let \( \hat{Q}_{\hat{T}} \) be a quadrature formula for the reference element \( \hat{T} \) with positive weights \( \hat{\omega}_{\ell, \hat{T}}^T, 1 \leq \ell \leq L \), and the property that for all \( q \in \mathbb{N} \) there holds
\begin{align*}
(2.5.33a) \quad \hat{P}_{\hat{T}} \subset P_q(\hat{T}), \\
(2.5.33b) \quad the \quad quadrature \quad formula \quad is \quad exact \quad for \quad polynomials \quad \hat{p} \in P_{2q-2}(\hat{T}) \quad or \quad \\
(2.5.33c) \quad \bigcup_{\ell=1}^L \{\hat{b}_{\ell, T}^K\} \quad contains \quad a \quad P_{q-1}(\hat{T}) \quad unisolvent \quad subset. 
\end{align*}

Then, the sequence \( \{a_h(\cdot, \cdot)\}_{h \in \mathcal{H}} \) is uniformly \( V_h \)-elliptic.

Proof. Observing \( v_h|_T = p_T \in P_T \) and using the ellipticity condition 2.1.2b, we find
\begin{align*}
(2.5.34) \quad \sum_{\ell=1}^L \omega_{\ell, T} \sum_{i,j=1}^d (a_{ij} \frac{\partial \hat{v}_h}{\partial x_i} \frac{\partial \hat{v}_h}{\partial x_j})(b_{\ell, T}) &= \\
= \sum_{\ell=1}^L \omega_{\ell, T} \sum_{i,j=1}^d (a_{ij} \frac{\partial p_T}{\partial x_i} \frac{\partial p_T}{\partial x_j})(b_{\ell, T}) &\geq \\
\geq \alpha \sum_{\ell=1}^L \omega_{\ell, T} \sum_{i=1}^d \left( \frac{\partial p_T}{\partial x_i}(b_{\ell, T}) \right)^2 &\geq \|D_{p_T}(b_{\ell, T})\|^2.
\end{align*}
Now, observing

\[ D\hat{p}_T(\hat{b}_{\ell, T})\xi = Dp(b_{\ell, T})(B_T\xi) \quad , \quad 1 \leq \ell \leq L , \]

we have

\[ \| D\hat{p}_T(\hat{b}_{\ell, T})\| \leq \| B_T \| \| Dp(b_{\ell, T})\| \quad , \quad 1 \leq \ell \leq L , \]

and hence, using Lemma 2.4.71, we get

\[ (2.5.35) \sum_{\ell=1}^{L} \omega_{\ell, T} \sum_{i=1}^{d} (\partial p_T(\partial x_i)(b_{\ell, T}))^2 \geq \| B_T \|^{-2} \sum_{\ell=1}^{L} \omega_{\ell, T} \sum_{i=1}^{d} (\partial \hat{p}_T(\hat{b}_{\ell, T}))^2 = \| B_T \|^{-2} \sum_{\ell=1}^{L} \omega_{\ell, T} \sum_{i=1}^{d} (\partial \hat{p}_T(\hat{b}_{\ell, T}))^2 . \]

(i) Let us first assume that the quadrature formula \( \hat{Q}_T \) is exact for polynomials \( \hat{p} \in P_{2q-2}(T) \). Since

\[ \sum_{i=1}^{d} (\partial \hat{p}_T(\hat{b}_{\ell, T}))^2 \in P_{2q-2}(T) , \]

we then have

\[ (2.5.36) |\hat{p}_T|_{1, T}^2 = \int_T \sum_{i=1}^{d} (\partial \hat{p}_T(\hat{b}_{\ell, T}))^2 d\hat{x} = \sum_{i=1}^{d} (\partial \hat{p}_T(\hat{b}_{\ell, T}))^2 . \]

Inserting (2.5.36) into (2.5.35) and using again Lemma 2.4.71, it follows that

\[ (2.5.37) \sum_{\ell=1}^{L} \omega_{\ell, T} \sum_{i=1}^{d} (\partial p_T(\partial x_i)(b_{\ell, T}))^2 \geq \det(B_T) \| B_T \|^{-2} \sum_{\ell=1}^{L} \omega_{\ell, T} \sum_{i=1}^{d} (\partial \hat{p}_T(\hat{b}_{\ell, T}))^2 = \det(B_T) \| B_T \|^{-2} |\hat{p}_T|_{1, T}^2 \geq \left( \| B_T \| \| B_T^{-1} \| \right)^{-2} |p_T|_{1, T}^2 . \]

Due to the shape regularity of \( T_h(\Omega), h \in H \),

\[ \| B_T \| \| B_T^{-1} \| \leq \frac{\hat{h}_{\ell, T}}{\rho_T} \frac{h_T}{\rho_T} \leq C . \]

Combining (2.5.33a),(2.5.35) and (2.5.37), we deduce the existence of a positive constant \( \tilde{\alpha} \), independent of \( h \in H \), such that
2.5 Numerical quadrature and nonconforming methods

\[(2.5.38) \sum_{\ell=1}^{L} \omega_{\ell,T} \sum_{i,j=1}^{d} (a_{ij} \frac{\partial v_{h}}{\partial x_{i}} \frac{\partial v_{h}}{\partial x_{j}})(b_{\ell,T}) \geq \tilde{\alpha} |v_{h}|_{1,T}^{2}, \quad v_{h} \in V_{h}.\]

Forming the sum over all \(T \in T_{h}(\Omega)\) in (2.5.38), gives the assertion.

(ii) What remains to be shown is that the assertion also holds true, if we assume that the union \(\bigcup_{\ell=1}^{L} \{\hat{b}_{\ell,T}\}\) contains a \(P_{q-1}(\hat{T})\)-unisolvent subset. We claim that in this case (2.5.36) has to be replaced by

\[(2.5.39) \sum_{i=1}^{d} \left( \frac{\partial \hat{p}_{T}}{\partial \hat{x}_{i}}(\hat{b}_{\ell,T}) \right)^{2} \geq \hat{C} |\hat{p}_{T}|_{1,T}^{2},\]

where \(\hat{C}\) is a positive constant. The proof then proceeds in the same way as before. In order to verify (2.5.39), it suffices to show that

\[
\left( \sum_{\ell=1}^{L} \hat{\omega}_{\ell,\hat{T}} \sum_{i=1}^{d} \left( \frac{\partial \hat{p}_{T}}{\partial \hat{x}_{i}}(\hat{b}_{\ell,T}) \right)^{2} \right)^{1/2}
\]

provides a norm on the quotient space \(\hat{P}_{q-1}(\hat{T})/P_{0}(\hat{T})\), since so does \(\cdot |_{1,\hat{T}}\), and we may conclude taking advantage of the equivalence of norms on finite dimensional spaces. For that purpose, we assume

\[
\sum_{\ell=1}^{L} \hat{\omega}_{\ell,\hat{T}} \sum_{i=1}^{d} \left( \frac{\partial \hat{p}_{T}}{\partial \hat{x}_{i}}(\hat{b}_{\ell,T}) \right)^{2} = 0 .
\]

Then, the positivity of the weights \(\hat{\omega}_{\ell,\hat{T}}, 1 \leq \ell \leq L\), yields

\[
\frac{\partial \hat{p}_{T}}{\partial \hat{x}_{i}}(\hat{b}_{\ell,T}) = 0, \quad 1 \leq i \leq d, \quad 1 \leq \ell \leq L .
\]

But for each \(i \in \{1,\ldots,d\}\)

\[
\frac{\partial \hat{p}_{T}}{\partial \hat{x}_{i}} \in P_{q-1}(\hat{T}),
\]

and hence,

\[
\frac{\partial \hat{p}_{T}}{\partial \hat{x}_{i}} \equiv 0 ,
\]

since it vanishes on a \(P_{q-1}(\hat{T})\)-unisolvent subset. \(\Box\)

Strang’s first lemma (2.5.6) allows to derive an a priori estimate. Under the regularity assumption (2.5.21), for the approximation error we obtain

\[(2.5.40) \inf_{v_{h} \in V_{h}} \|u - v_{h}\|_{1,\Omega} \leq \|u - \Pi_{h}u\|_{1,\Omega} \leq C h^{k} |u|_{k+1,\Omega} .\]
Hence, we have to provide sufficient conditions such that the consistency errors (2.5.10a), (2.5.10b) do not deteriorate this order, i.e., there exist constants \( C_1 \) and \( C_2 \), independent of \( h \in \mathcal{H} \), such that

\[
\begin{align*}
(2.5.41a) & \quad \sup_{w_h \in V_h} \frac{|a(\Pi_h u, w_h) - a_h(\Pi_h u, w_h)|}{\|w_h\|_V} \leq C_1 h^k, \\
(2.5.41b) & \quad \sup_{w_h \in V_h} \frac{|\ell(w_h) - \ell_h(w_h)|}{\|w_h\|_V} \leq C_2 h^k.
\end{align*}
\]

(2.5.42) **Theorem:** Let \( u \in V \) be the unique solution of (2.5.1). Moreover, for \( p \geq 2 \) and \( k > \max((d-2)/2, d/p) \) assume that the conditions (2.5.21), (2.5.22) are satisfied. Further, let \( T_h(\Omega) \) be a geometrically conforming triangulation of \( \Omega \) and let \( (\hat{T}, \hat{P}_k, \hat{\Sigma}_T) \) a Lagrangian finite element of type \((k)\), where \( \hat{T} \) stands for the reference element, \( \hat{P}_k = P_k(\hat{T}) \) and \( \hat{\Sigma}_T \) is the set of degrees of freedom. We suppose that \( \hat{Q}_T \) is a quadrature formula with positive weights \( \hat{\omega}_l, 1 \leq l \leq L \), which is exact for polynomials \( \hat{p} \in P_{2k-2}(\hat{T}) \). If \( V_h \) is the associated Lagrangian finite element space of type \((k)\) and \( u_h \in V_h \) is the unique solution of (2.5.3) with \( a_h(\cdot, \cdot), h \in \mathcal{H} \), and \( l_h(\cdot), h \in \mathcal{H} \), according to (2.5.31a), (2.5.31b), there exists a constant \( C > 0 \), independent of \( h \in \mathcal{H} \), such that

\[
(2.5.43) \quad \|u - u_h\|_{1, \Omega} \leq Ch^k \left( |u|_{k+1, \Omega} + \sum_{i,j=1}^d \|a_{ij}\|_{k,\infty, \Omega} \|u\|_{k+1, \Omega} \right).
\]

**Proof.** Due to (2.5.40) it suffices to prove (2.5.41a), (2.5.41b). Using the Lemma of Bramble-Hilbert and the **generalized Leibniz formula**

\[
(2.5.44) \quad |uv|_{m,\Omega} \leq C(m, d) \sum_{j=0}^m |u|_{m-j,\Omega} |v|_{j,\infty,\Omega},
\]

which holds true for \( u \in H^m(\Omega), v \in W^{m,\infty}(\Omega), m \in \mathbb{N}_0 \), for \( p, q \in P_k(T) \) we obtain

\[
|E_T(a_{ij} \frac{\partial q}{\partial x_i} \frac{\partial p}{\partial x_j})| \leq Ch^k \|a_{ij}\|_{k,\infty, \Omega} |p|_{1,T} \|q\|_{k,T}.
\]

This implies the following estimate of the consistency error (2.5.10a)

\[
\begin{align*}
& \sup_{w_h \in V_h} \frac{|a(\Pi_h u, w_h) - a_h(\Pi_h u, w_h)|}{\|w_h\|_{1, \Omega}} \leq \\
& \leq Ch^k \sum_{i,j=1}^d \|a_{ij}\|_{k,\infty, \Omega} \|u\|_{k+1, \Omega}.
\end{align*}
\]

On the other hand, as far the consistency error (2.5.10b) is concerned, the **generalized Cauchy-Schwarz inequality**
where obtained by tensor product construction, there holds

\[ \sum_{i=1}^{m} |a_i b_i c_i| \leq \left( \sum_{i=1}^{m} |a_i|^r_1 \right)^{1/r_1} \left( \sum_{i=1}^{m} |b_i|^r_2 \right)^{1/r_2} \left( \sum_{i=1}^{m} |c_i|^r_3 \right)^{1/r_3} , \]

where \( 1/r_1 = 1/2 - 1/p, r_2 = p, r_3 = 2 \) and \( m = \text{card}(T_h(\Omega)) \), implies the estimate

\[ \sup_{w_h \in V_h} \frac{|\ell(\hat{w}_h) - \ell_h(\hat{w}_h)|}{\|w_h\|_{1,\Omega}} \leq C h^k \|\Omega\|^{1/2-1/p} \|f\|_{k,p,\Omega} . \]

For details we refer to chapter 4.1 in Ciarlet (2002).

**Example (Quadrature formulas for simplicial elements):** Quadrature formulas for simplicial elements \( T \) are based on polynomial approximations of the integrand where the nodes \( b^T_\ell \) and the weights \( \omega^T_\ell \), \( 1 \leq \ell \leq L \), are chosen such that \( E_T(p) = 0, p \in P_k(T), k \in \mathbb{N} \). For \( P_k(T) = \text{span}\{p_1, \cdots, p_{n_k}\}, n_k = \binom{k+d}{d} \), this leads to the moment equations

\[ Q_T(p_i) = I_T(p_i) \quad , \quad 1 \leq i \leq n_k \]

in the unknowns \( b^T_\ell \) and \( \omega^T_\ell \). For the construction of quadrature formulas one can either determine polynomials that satisfy (2.5.46) or one can try to solve the system of nonlinear equations (2.5.46). In case \( d = 2 \) the requirement (2.5.46) leads to the conditions

\[ L \geq \begin{cases} \frac{(k+2)(k+4)/8}{(k+1)(k+3)/8 + [(k+1)/4]} & , \text{k odd} \\ \frac{(k+2)(k+4)/8}{(k+1)(k+3)/8} & , \text{k even} \end{cases} \]

where \( q := \min\{m \in \mathbb{N} | m \leq q\}, q \in \mathbb{Z} \).

For even \( k \), a quadrature formula with \( L = (k+2)(k+4)/8 \) has positive weights, whereas this property is not guaranteed for odd \( k \). Examples of quadrature formulas for simplicial elements can be found in in Stroud (1971)

**Example (Quadrature formulas for quadrilateral elements):** Let \( \hat{T} \) be the quadrilateral reference element \( T = [0,1]^d \) and let \( b_\ell, \omega_\ell, 1 \leq \ell \leq k+1, k \in \mathbb{N}_0 \), be the nodes and weights of the Gauss-Legendre quadrature formula

\[ Q_{[0,1]}(\hat{\varphi}) = \sum_{\ell=1}^{k+1} \omega_\ell \varphi(b_\ell) , \]

where \( E_{[0,1]}(\hat{p}) = 0 \), \( \hat{p} \in P_{2k+1}([0,1]) \). Then, for the quadrature formula

\[ Q_{\hat{T}}(\varphi) = \sum_{\ell_1, \cdots, \ell_d \in \{1, \cdots, k+1\}} \prod_{j=1}^{d} \omega_{\ell_j} \varphi(b_{\ell_1}, \cdots, b_{\ell_d}) , \]

obtained by tensor product construction, there holds \( E_{\hat{T}}(\hat{p}) = 0 \), \( \hat{p} \in Q_{2k+1}(\hat{T}) \).

As an example for the application of Strang’s second lemma we consider nonconforming finite elements in the discretization of second order elliptic boundary value problems with respect to a family of shape-regular simplicial triangulations \( T_h(\Omega), h \in \mathcal{H} \), of the computational domain \( \Omega \subset \mathbb{R}^d \). We restrict ourselves to the lowest order Crouzeix-Raviart element.
(2.5.47) **Definition:** Let $T$ be a simplex in $\mathbb{R}^d$ and

$$P_T := P_1(T),$$

$$\Sigma_T := \{ p(m_i(T)) \mid 1 \leq i \leq d+1 \},$$

where $m_i(T), 1 \leq i \leq d+1,$ are the midpoints of the edges $(d = 2)$ and the centers of gravity of the faces $(d = 3),$ respectively. Then, $CR_1(T) := (T, P_T, \Sigma_T)$ is called the lowest order Crouzeix-Raviart element. It is also referred to as the nonconforming $P1$ element.

(2.5.48) **Lemma:** The Crouzeix-Raviart element $CR_1(T)$ is unisolvent.

**Proof.** The easy proof is left as an exercise. □

(2.5.49) **Definition:** We denote by $CR_1(\Omega, T_h(\Omega))$ the Crouzeix-Raviart finite element space composed by the Crouzeix-Raviart elements $CR_1(T)$, $T \in T_h(\Omega)$ according to

$$CR_1(\Omega, T_h(\Omega)) := \{ v_h \in L^2(\Omega) \mid v_h|_T \in P_1(T), T \in T_h(\Omega),$$

$$v_h \text{ is continuous in } m(F), F \in \mathcal{F}_h(\Omega) \},$$

where $m(F), F \in \mathcal{F}_h(\Omega)$ are the midpoints (centers of gravity) of interior edges (faces). The subspace $CR_{1,0}(\Omega, T_h(\Omega))$ is defined by means of

$$CR_{1,0}(\Omega, T_h(\Omega)) := \{ v_h \in CR_1(\Omega, T_h(\Omega)) \mid v_h(m(F)) = 0, F \in \mathcal{F}_h(\partial \Omega) \}.$$

Obviously, in general

$$CR_1(\Omega, T_h(\Omega)) \not\subset H^1(\Omega), \quad CR_{1,0}(\Omega, T_h(\Omega)) \not\subset H^1_0(\Omega).$$

We consider the numerical solution of Poisson’s equation under homogeneous Dirichlet boundary conditions

(2.5.50) $$a(u, v) := \int_{\Omega} \nabla u \cdot \nabla v \, dx = \ell(v) := \int_{\Omega} fv \, dx, \ v \in H^1_0(\Omega)$$

by the nonconforming finite element approximation

(2.5.51) $$a_h(u_h, v_h) = \ell(v_h), \quad v_h \in CR_{1,0}(\Omega, T_h(\Omega)), \ h \in \mathcal{H}.$$ 

Here, the mesh dependent bilinear form

$$a_h(\cdot, \cdot) : H^1_0(\Omega) \oplus CR_{1,0}(\Omega, T_h(\Omega)) \times H^1_0(\Omega) \oplus CR_{1,0}(\Omega, T_h(\Omega)) \to \mathbb{R}$$

is given by

(2.5.52) $$a_h(u, v) := \sum_{T \in T_h(\Omega)} \int_T \nabla u \cdot \nabla v \, dx,$$
where \( u, v \in H^1_0(\Omega) \oplus CR_{1,0}(\Omega, T_h(\Omega)) \). We associate with (2.5.52) the mesh dependent norm

\[
\|v\|_h := \sqrt{a_h(v,v)}, \quad v \in CR_{1,0}(\Omega, T_h(\Omega)) \times H^1_0(\Omega).
\]

We will derive a quasi-optimal a priori error estimate in the \( \| \cdot \|_h \)-norm by the application of Strang’s second lemma.

\[\text{(2.5.54) Theorem: Let} \quad u \in H^1(\Omega) \cap H^s(\Omega) \quad s = 2 \quad (d = 2), \quad s = 3 \quad (d = 3), \quad \text{be the solution of} \quad (2.5.50) \quad \text{and} \quad u_h \in CR_{1,0}(\Omega, T_h(\Omega)), h \in \mathcal{H}, \quad \text{its nonconforming P1-approximations. Then, there exists a constant} \quad C > 0, \quad \text{depending only on the shape regularity of the triangulations} \quad T_h(\Omega), h \in \mathcal{H}, \quad \text{such that} \quad \]

\[
\|u - u_h\|_h \leq C h |u|_{s, \Omega}.
\]

\[\text{Proof. It is easily verified that the family} \quad (a_h(\cdot, \cdot))_{h \in \mathcal{H}} \quad \text{of bilinear forms is uniformly} \quad V_h\text{-elliptic} \quad (V_h := CR_{1,0}(\Omega, T_h(\Omega))). \quad \text{According to Strang’s second lemma (2.5.15) we have to estimate the approximation error} \quad \]

\[
\inf_{v_h \in CR_{1,0}(\Omega, T_h)} \|u - v_h\|_h
\]

and the consistency error

\[
\sup_{w_h \in CR_{1,0}(\Omega, T_h)} \frac{a_h(u, w_h) - \ell(w_h)}{\|w_h\|_h}.
\]

(i) Estimation of the approximation error

Since \( S_{1,0}(\Omega, T_h(\Omega)) \subset CR_{1,0}(\Omega, T_h(\Omega)) \), in view of Theorem ?? we have

\[
\inf_{v_h \in CR_{1,0}(\Omega, T_h)} \|u - v_h\|_h \leq \inf_{v_h \in S_{1,0}(\Omega, T_h)} \|u - v_h\|_h \leq C h |u|_{s, \Omega}.
\]

(ii) Estimation of the consistency error

Observing that \(- \Delta u(x) = f(x) \quad \text{f.a.a.} \quad x \in \Omega, \)

for \( w_h \in CR_{1,0}(\Omega, T_h(\Omega)) \) we obtain by Green’s formula

\[
L_u(w_h) := a_h(u, w_h) - \ell(w_h) =
\]

\[
= \sum_{T \in T_h(\Omega)} \left( \int_T \nabla u \cdot \nabla w_h \, dx - \int_T f w_h \, dx \right) =
\]

\[
= \sum_{T \in T_h(\Omega)} \left( \int_T (- \Delta u - f) w_h \, dx + \int_{\partial T} \frac{\partial u}{\partial n} w_h \, d\sigma \right) =
\]

\[
= \sum_{T \in T_h(\Omega)} \sum_{F \in F_h(T)} \int_F \frac{\partial u}{\partial n} w_h \, d\sigma.
\]
Denoting by \( w_h|_F \) the integral mean

\[
w_h|_F := \frac{1}{\text{meas}(F)} \int_F w_h\, d\sigma , \quad F \in \mathcal{F}_h(\Omega),
\]

it follows that

\[
(2.5.60) \quad L_u(w_h) = \sum_{T \in \mathcal{T}_h(\Omega)} \sum_{F \in \mathcal{F}_h(T)} \int_{\nu \cdot \nabla I_h u \cdot (w_h - w_h|_F)} d\sigma .
\]

Moreover, since \( \nu \cdot \nabla I_h u \in P_0(T_\nu) \), \( 1 \leq \nu \leq 2 \), we have

\[
\int_{F \cap T_\nu} \nu \cdot \nabla I_h u \cdot (w_h - w_h|_F) \, d\sigma = 0 ,
\]

and hence, (2.5.60) gives rise to

\[
L_u(w_h) = \sum_{T \in \mathcal{T}_h(\Omega)} \sum_{F \in \mathcal{F}_h(T)} \int_{F} \nu \cdot \nabla (u - I_h u) \cdot (w_h - w_h|_F) \, d\sigma .
\]

The Cauchy-Schwarz inequality yields

\[
|L_u(w_h)| \leq \sum_{T \in \mathcal{T}_h(\Omega)} \sum_{F \in \mathcal{F}_h(T)} \left( \int_{F} |\nu \cdot \nabla (u - I_h u)|^2 \, d\sigma \int_{F} |w_h - w_h|_F|^2 \, d\sigma \right)^{1/2} .
\]

We will estimate \( I_1 \) and \( I_2 \) separately. For the estimation of \( I_1 \) we take advantage of the affine equivalence of the Crouzeix-Raviart elements. For the reference element \( \hat{T} \) we find by the trace theorem (1.2.23) and the Bramble-Hilbert lemma (2.4.74)

\[
\int_{\partial \hat{T}} |\nu \cdot \nabla (u - I_h u)|^2 \, d\sigma \leq C \|u - I_h u\|_{2,\hat{T}}^2 \leq C \|u\|_{s,\hat{T}}^2 .
\]

By a standard scaling argument

\[
(2.5.61) \quad \int_{\partial T} |\nu \cdot \nabla (u - I_h u)|^2 \, d\sigma \leq C \, h \, |u|_{s,T}^2 .
\]

For the estimation of \( I_2 \) we consider again the reference element \( \hat{T} \) and find by the Bramble-Hilbert lemma (2.4.74)

\[
\int_{\hat{F}} |\hat{w}_h - \hat{w}_h|_{\hat{F}}|^2 \, d\hat{\sigma} \leq C |\hat{w}_h|_{1,\hat{T}}^2 , \quad \hat{w}_h \in P_1(\hat{T}) , \quad \hat{F} \in \mathcal{F}(\hat{T}) ,
\]
For $w_h \in CR_{1,0}(\Omega, T_h(\Omega))$ and $F \in F_h(T), T \in T_h(\Omega)$, a standard scaling argument yields

\[(2.5.62) \quad \int_F |w_h - w_h|_F^2 \, d\sigma \leq C \ h |w_h|_{1,T}^2 .\]

Using (2.5.61) and (2.5.62), we finally obtain

\[
|L_u(w_h)| \leq 3 \ C \ h \ \sum_{T \in T_h(\Omega)}|u|_{s,T} |w_h|_{1,T} \leq \ C \ h \left( \sum_{T \in T_h(\Omega)}|u|_{s,T}^2 \ \sum_{T \in T_h(\Omega)}|w_h|_{1,T}^2 \right)^{1/2} = \ C \ h \ |u|_{s,\Omega} \ |w_h|_h .
\]

\[\square\]

We remark that the definition of nonconforming Q1 elements for quadrilateral triangulations is not obvious, since the analogue of the Crouzeix-Raviart element with the degrees of freedom associated with the midpoints of the edges (faces) is not unisolvent. One has to modify $P_T = Q_1(T)$ by appropriate rotations which gives rise to the so-called rotated bilinear elements, also known as Rannacher-Turek elements. For details we refer to Rannacher and Turek (1992).

### 2.6 Adaptive finite element methods

The theory and application of adaptive finite element methods based on a posteriori error estimates has reached some state of maturity, as it is documented by a variety of monographs on this subject (cf., e.g., Ainsworth and Oden (2000), Babuska and Strouboulis (2001), Bangerth and Rannacher (2003), Estep et al. (1996), Neittaanmäki and Repin (2004), Verfürth (1996). There are different concepts such as residual and hierarchical type a posteriori error estimators, error estimators based on local averaging, functional type error majorants, and the goal-oriented weighted dual approach. In this section, we will focus on residual type a posteriori error estimators and follow the exposition in Verfürth (1996).

We shall deal with the following model problem: Let $\Omega$ be a bounded simply-connected polygonal domain in Euclidean space $\mathbb{R}^2$ with boundary $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N, \Gamma_D \cap \Gamma_N = \emptyset$, and consider the elliptic boundary value problem

\[
\begin{align*}
(2.6.1a) \quad -\Delta u &= f \quad \text{in} \ \Omega , \\
(2.6.1b) \quad u &= 0 \quad \text{on} \ \Gamma_D , \\
\nu \cdot \nabla u &= g \quad \text{on} \ \Gamma_N , 
\end{align*}
\]
where \( f \in L^2(\Omega) \) and \( g \in L^2(\Gamma_N) \). Setting
\[
V := \{ v \in H^1(\Omega) \mid v|_{\Gamma_D} = 0 \},
\]
the weak formulation of (2.6.1a)-(???) is as follows: Find \( u \in V \) such that
\[
(2.6.2) \quad a(u, v) = \ell(v), \quad v \in V,
\]
where
\[
a(v, w) := \int_{\Omega} \nabla v \cdot \nabla w \, dx, \quad v, w \in V,
\]
\[
\ell(v) := \int_{\Omega} f v \, dx + \int_{\Gamma_N} g v \, d\sigma, \quad v \in V.
\]
Given a geometrically conforming simplicial triangulation \( T_h(\Omega) \) of \( \Omega \), we denote by
\[
V_h := \{ v_h \in H^1_{0,D}(\Omega) \mid v_h|_T \in P_1(T), \ T \in T_h(\Omega) \}
\]
the trial space of continuous, piecewise linear finite elements with respect to \( T_h(\Omega) \). In the sequel we will refer to \( \mathcal{N}_h(D) \) and \( \mathcal{E}_h(D) \), \( D \subseteq \Omega \) as the sets of vertices and edges of \( T_h(\Omega) \) in \( D \). Moreover, we denote by \( h_T \) the diameter of an element \( T \in T_h(\Omega) \) and by \( h_E \) the length of an edge \( E \in \mathcal{E}_h(\Omega \cup \Gamma_N) \). For two quantities \( A, B \in \mathbb{R}_+ \) we will use the notation \( A \lesssim B \), if there exists a positive constant \( C \) depending only on the local geometry of the triangulations such that \( A \leq CB \). We will write \( A \approx b \), if \( A \lesssim B \) and \( B \lesssim A \).

The conforming P1 approximation of (2.6.2) reads as follows: Find \( u_h \in V_h \) such that
\[
(2.6.3) \quad a(u_h, v_h) = \ell(v_h), \quad v_h \in V_h.
\]
We are interested in the a posteriori estimation of the global discretization error
\[
e := u - \tilde{u}_h.
\]
It is easy to see that \( e \in V \) satisfies the error equation
\[
(2.6.4) \quad a(e, v) = r(v), \quad v \in V,
\]
where \( r(\cdot) \) stands for the residual with respect to the computed approximation \( u_h \in V_h \) according to
\[
(2.6.5) \quad r(v) := \int_{\Omega} f v \, dx + \int_{\Gamma_N} g v \, d\sigma - a(u_h, v), \quad v \in V.
\]
We are interested in a cheaply computable \textit{a posteriori error estimator} $\eta_h$ consisting of elementwise error contributions $\eta_T, T \in \mathcal{T}_h(\Omega)$ and edgewise error contributions $\eta_E, E \in \mathcal{E}_h(\Omega \cup \Gamma_N)$ in the sense that

\begin{equation}
\eta^2_h = \sum_{T \in \mathcal{T}_h(\Omega)} \eta^2_T + \sum_{E \in \mathcal{E}_h(\Omega \cup \Gamma_N)} \eta^2_E,
\end{equation}

which up to higher order terms (h.o.t.) provides a lower and an upper bound for $e$ according to

\begin{equation}
\gamma \eta_h - \text{h.o.t.} \leq |e|_{1,\Omega} \leq \Gamma \eta_h + \text{h.o.t.}
\end{equation}

with constants $0 < \gamma \leq \Gamma$ depending only on the local geometry of the triangulation $\mathcal{T}_h(\Omega)$.

We may use the local error terms $\eta_K$ and $\eta_E$ as a criterion for a \textit{local refinement} of the elements and edges of the triangulation. Among several possible criteria, the one based on mean values strategy is as follows: Compute

\[
\bar{\eta}_T := \frac{1}{n_T} \sum_{T \in \mathcal{T}_h(\Omega)} \eta_T,
\]

\[
\bar{\eta}_E := \frac{1}{n_E} \sum_{E \in \mathcal{E}_h(\Omega \cup \Gamma_N)} \eta_E,
\]

where $n_T := \text{card}(\mathcal{T}_h(\Omega))$ and $n_E := \text{card}(\mathcal{E}_h(\Omega \cup \Gamma_N))$.

Mark an element $T \in \mathcal{T}_h(\Omega)$ and an edge $E \in \mathcal{E}_h(\Omega \cup \Gamma_N)$ for refinement, if

\[
\eta_T \geq \sigma \bar{\eta}_T,
\]

\[
\eta_E \geq \sigma \bar{\eta}_E,
\]

where $0 < \sigma \leq 1$ is some appropriate safety factor, e.g., $\sigma = 0.9$.

(2.6.8) \textbf{Definition:} \textit{An a posteriori error estimator} $\eta_h$ satisfying

\begin{equation}
|e|_{1,\Omega} \lesssim \eta_h + \text{h.o.t.}
\end{equation}

is called reliable, since it ensures a sufficient refinement in the sense that the $H^1$-seminorm of the error $e$ will be bounded by a quantity of the same order of magnitude as a user-prescribed accuracy, if this accuracy is tested by $\eta_h$. On the other hand, an a posteriori error estimator for which

(2.6.10) \quad $\eta_h - \text{h.o.t.} \lesssim |e|_{1,\Omega}$

is said to be efficient, since it prevents too much refinement.

A \textit{residual a posteriori error estimator} can be derived by viewing the residual as an element of the dual space $V^*$ and evaluating it with respect to the dual norm. An important tool in the construction of an upper bound for the error is \textit{Clément's quasi-interpolation operator}. 
(2.6.11) **Definition:** For \( p \in \mathcal{N}_h(\Omega \cup \Gamma_N) \) we denote by \( \varphi_p \) the nodal basis function in \( V_h \) with supporting point \( p \) and we refer to \( D_p \) as the set

\[
D_p := \bigcup \{ T \in \mathcal{T}_h(\Omega) \mid p \in \mathcal{N}_h(T) \}.
\]

We refer to \( \pi_p \) as the \( L^2 \)-projection onto \( P_1(D_p) \), i.e.,

\[
(\pi_p(v),w)_{0,D_p} = (v,w)_{0,D_p} , \quad w \in P_1(D_p).
\]

Then, Clément’s quasi-interpolation operator \( P_C : L^2(\Omega) \to V_h \) is defined as follows

\[
P_Cv := \sum_{p \in \mathcal{N}_h(\Omega \cup \Gamma_N)} \pi_p(v) \varphi_p.
\]

In order to establish local approximation properties of Clément’s interpolation operator, for \( T \in \mathcal{T}_h(\Omega) \) and \( E \in \mathcal{E}_h(\Omega \cup \Gamma_N) \) we introduce the sets

(2.6.12a) \( D_T := \bigcup \{ T' \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(T') \cap \mathcal{N}_h(T) \neq \emptyset \} \),

(2.6.12b) \( D_E := \bigcup \{ T' \in \mathcal{T}_h(\Omega) \mid \mathcal{N}_h(E) \cap \mathcal{N}_h(T') \neq \emptyset \} \).

Using the affine equivalence of the elements and the Bramble-Hilbert Lemma one can show the following result.

(2.6.13) **Theorem:** For \( T \in \mathcal{T}_h(\Omega) \) and \( E \in \mathcal{E}_h(\Omega \cup \Gamma_N) \) let \( D_T^{(1)} \) and \( D_E^{(1)} \) be given by (2.6.12a),(2.6.12b) and let \( P_C \) be Clément’s interpolation operator as given by Definition (2.6.11). Then, for all \( v \in V \) there holds

(2.6.14a) \( \| P_C v \|_{0,T} \lesssim \| v \|_{0,D_T} \),

(2.6.14b) \( \| P_C v \|_{0,E} \lesssim \| v \|_{0,D_E} \),

(2.6.14c) \( \| \nabla P_C v \|_{0,T} \lesssim \| \nabla v \|_{0,D_T^{(1)}} \),

(2.6.14d) \( \| v - P_C v \|_{0,T} \lesssim h_T \| v \|_{1,D_T} \),

(2.6.14e) \( \| v - P_C v \|_{0,E} \lesssim h_{1/2}^{E} \| v \|_{1,D_E} \).

Further, there holds

(2.6.15a) \( \left( \sum_{T \in \mathcal{T}_h(\Omega)} \| v \|^2_{\mu,D_T^{(1)}} \right)^{1/2} \lesssim \| v \|_{\mu,\Omega} , \quad 0 \leq \mu \leq 1 \),

(2.6.15b) \( \left( \sum_{E \in \mathcal{E}_h(\Omega \cup \Gamma_N)} \| v \|^2_{\mu,D_E^{(1)}} \right)^{1/2} \lesssim \| v \|_{\mu,\Omega} , \quad 0 \leq \mu \leq 1 \).

**Proof.** We refer to Verfürth (1996). \( \square \)
We have now provided all prerequisites to establish an upper bound for the error $e$ measured in the $H^1$-seminorm by a residual a posteriori error estimator

\begin{equation}
\eta_h := \left( \sum_{T \in \mathcal{T}_h(\Omega)} \eta_T^2 + \sum_{E \in \mathcal{E}_h(\Omega \cup \Gamma_N)} \eta_E^2 \right)^{1/2}.
\end{equation}

Here, the element residuals $\eta_T$ and the edge residuals $\eta_E$ are given by

\begin{align}
\eta_T &:= h_T \| f_h \|_{0,T}, \\
\eta_E &:= \begin{cases} 
    h_E^{1/2} \| \nabla u_h \|_{0,E}, & E \in \mathcal{E}_h(\Omega), \\
    h_E^{1/2} \| g_h - \nu \cdot \nabla u_h \|_{0,E}, & E \in \mathcal{E}_h(\Gamma_N), 
\end{cases}
\end{align}

where $f_h, g_h$ stand for the elementwise resp. edgewise constant functions with $f_h|_T := \int_T f \, dx$ and $g_h|_E := \int_E g \, d\sigma$ and $[\nabla u_h] := u_h|_{T_+} - u_h|_{T_-}$ stands for the jump of $u_h$ across an interior edge $E = T_+ \cap T_-, T_\pm \in \mathcal{T}_h(\Omega)$.

The a posteriori error analysis further invokes the data oscillations

\begin{equation}
\text{osc}_h := \left( \sum_{T \in \mathcal{T}_h(\Omega)} \text{osc}_T^2 + \sum_{E \in \mathcal{E}_h(\Gamma_N)} \text{osc}_E^2 \right)^{1/2},
\end{equation}

where $\text{osc}_T$ and $\text{osc}_E$ are given by

$$
\text{osc}_T := h_T \| f - f_h \|_{0,T}, \quad \text{osc}_E := h_E \| g - g_h \|_{0,E}.
$$

\begin{theorem}
Let $\eta_h$ and $\text{osc}_h$ be the residual error estimator and the data oscillations as given by (2.6.16) and (2.6.18). Then, there holds

\begin{equation}
|e|_{1,\Omega}^2 \lesssim \eta_h^2 + \text{osc}_h^2.
\end{equation}

\end{theorem}

\begin{proof}
Setting $v = e$ in (2.6.4), we obtain

\begin{equation}
|e|_{1,\Omega}^2 \leq a(e, e) = r(e) = r(PCe) + r(e - PCe).
\end{equation}

Due to Galerkin orthogonality, we have $r(PCe) = 0$. On the other hand, for the second term on the right-hand side of (2.6.21), Green’s formula yields

$$
r(e - PCe) = \int_{\Omega} f (e - PCe) \, dx + \int_{\Gamma_N} g (e - PCe) \, d\sigma +
\sum_{T \in \mathcal{T}_h(\Omega)} \int_T -\Delta u_h (e - PCe) \, dx + \sum_{T \in \mathcal{T}_h(\Omega) \partial T} \int \nu \cdot \nabla u_h (e - PCe) \, d\sigma =
\sum_{T \in \mathcal{T}_h(\Omega)} \int_T f_h (e - PCe) \, dx + \sum_{E \in \mathcal{E}_h(\Omega) \partial E} \int [\nu \cdot \nabla u_h] (e - PCe) \, d\sigma +
$$
\begin{align*}
&+ \sum_{E \in \mathcal{E}(\Gamma_N)} \int_E (g - g_h) \, (e - P_C e) \, d\sigma + \\
&+ \sum_{E \in \mathcal{E}(\Gamma_N)} \int_E (g - g_h) \, (e - P_C e) \, d\sigma + \\
&+ \sum_{T \in \mathcal{T}_h(\Omega)} \int_T (f - f_h) \, (e - P_C e) \, dx .
\end{align*}

In view of (2.6.14a),(2.6.14b) and (2.6.15a),(2.6.15b), it follows that

\begin{equation}
(2.6.22) \quad r(e - P_C e) \leq C \left( \left( \sum_{T \in \mathcal{T}_h(\Omega)} h_T^2 \| f_h \|_{0,T}^2 \right)^{1/2} |e|_{1,\Omega} + \\
+ \left( \sum_{E \in \mathcal{E}_h(\Omega)} h_E \| \nu \cdot \nabla u_h \|_{0,E}^2 \right)^{1/2} |e|_{1,\Omega} + \\
+ \left( \sum_{E \in \mathcal{E}(\Gamma_N)} h_E \| g_h - \nu \cdot \nabla u_h \|_{0,E}^2 \right)^{1/2} |e|_{1,\Omega} + \\
+ \left( \sum_{T \in \mathcal{T}_h(\Omega)} h_T^2 \| f - f_h \|_{0,T}^2 \right)^{1/2} |e|_{1,\Omega} + \\
+ \left( \sum_{E \in \mathcal{E}(\Gamma_N)} h_E \| g - g_h \|_{0,E}^2 \right)^{1/2} |e|_{1,\Omega} \right) .
\end{equation}

We deduce by using (2.6.21),(2.6.22) in (2.6.20).

For the construction of a lower bound we will now show local efficiency of the residual error estimator in the sense that up to data oscillations the local contributions provide lower bounds for the error measured in the $H^1$-seminorm on some patch associated with the elements and edges. For this purpose we need appropriate localized polynomial functions defined on the elements and the edges, respectively. Such functions are given by the element bubble functions $\psi_T$ and the edge bubble functions $\psi_E$.

\begin{definition}
(2.6.23) Let $\lambda_i^T, 1 \leq i \leq 3$, be the barycentric coordinates of $T \in \mathcal{T}_h(\Omega)$. The element bubble function is defined by means of

\begin{equation}
(2.6.24) \quad \psi_T := 27 \, \lambda_1^T \, \lambda_2^T \, \lambda_3^T .
\end{equation}

Note that $\text{supp}(\psi_T) = T$ and $\psi_T \big|_{\partial T} = 0$ , $T \in \mathcal{T}_h(\Omega)$. On the other hand, for $E \in \mathcal{E}_h(\Omega \cup \Gamma_N)$ and $T \in \mathcal{T}_h(\Omega)$ such that $E \subset \partial T$ and $p_i^T \in \mathcal{N}_h(E)$ , $1 \leq i \leq 2$, we introduce the edge bubble function according to

\begin{equation}
(2.6.25) \quad \psi_E := 4 \, \lambda_i^T \, \lambda_2^T .
\end{equation}

Note that $\psi_E \big|_{E'} = 0$ for $E' \in \mathcal{E}_h(T), E' \neq E$.\end{definition}
The bubble functions $\psi_T$ and $\psi_E$ have the following important properties that can be easily verified taking advantage of the affine equivalence of the elements.

**Lemma:** Let $k \in \mathbb{N}_0$. For the element and edge bubble functions and polynomials there holds

\begin{align*}
\tag{2.6.27a} & \|p_h\|_{0,T}^2 \lesssim (p_h, p_h \psi_T)_{0,T}, \quad p_h \in P_k(T), \\
\tag{2.6.27b} & \|p_h\|_{0,E}^2 \lesssim (p_h, p_h \psi_E)_{0,E}, \quad p_h \in P_k(E), \\
\tag{2.6.27c} & |p_h \psi_T|_{1,T} \lesssim h_T^{-1} \|p_h\|_{0,T}, \quad p_h \in P_k(T), \\
\tag{2.6.27d} & \|p_h \psi_T\|_{0,T} \lesssim \|p_h\|_{0,T}, \quad p_h \in P_k(T), \\
\tag{2.6.27e} & \|p_h \psi_E\|_{0,E} \lesssim \|p_h\|_{0,E}, \quad p_h \in P_k(E).
\end{align*}

**Proof.** We refer to Verfürth (1996). \hfill \Box

We are now able to prove the local efficiency of the estimator $\eta_h$. This will be done through a series of lemmas.

**Lemma:** For $T \in \mathcal{T}_h(\Omega)$ there holds

\begin{equation}
\tag{2.6.29} h_T^2 \|f_h\|_{0,T} \lesssim |e|_{1,T}^2 + \text{osc}_T^2(f).
\end{equation}

**Proof.** We set $\varphi := h_T f_h \psi_T$. In view of (2.6.27a), observing $\Delta u_h|_T = 0$ and applying Green’s formula, we obtain

\begin{equation}
\tag{2.6.30} h_T^2 \|f_h\|_{0,T} \lesssim (f_h + \Delta u_h, \varphi)_{0,T} = (f_h, \varphi)_{0,T} - a(u_h, \varphi).
\end{equation}

where we have used that $\psi_T|_{\partial T} = 0$. Since $\varphi$ is an admissible test function in (2.6.2), we further have

\begin{equation}
\tag{2.6.31} a(u, \varphi) - (f, \varphi)_{0,T} = 0.
\end{equation}

Using (2.6.31) in (2.6.30), it follows that

\begin{equation}
\tag{2.6.32} h_T^2 \|f_h\|_{0,T} \lesssim (a(u - u_h, \varphi) + (f_h - f, \varphi)_{0,T} \leq |e|_{1,T} \|\varphi\|_{1,T} + \|f - f_h\|_{0,T} \|\varphi\|_{0,T}.
\end{equation}

We deduce (2.6.29) from (2.6.27c) and (2.6.27d). \hfill \Box

**Lemma:** Let $E \in \mathcal{E}_h(\Omega)$ such that $E = T_+ \cap T_-, T_{\pm} \in \mathcal{T}_h(\Omega)$ and $\omega_E := T_+ \cup T_-$. Then, we have

\begin{equation}
\tag{2.6.33} h_E \|\nabla u_h\|_{0,E}^2 \lesssim |e|_{1,\omega_E}^2 + h_T^2 \|f_h\|_{0,\omega_E}^2 + \text{osc}_E^2(\omega_E)(f),
\end{equation}

where $\text{osc}_E^2(\omega_E)(f) = \text{osc}_{T_+}^2(f) + \text{osc}_{T_-}^2(f)$. 

Proof. We set \( \varphi := h_E(\nu \cdot [\nabla u_h]) \psi_E \). Using (2.6.27b), Green’s formula yields
\[
h_E \| \nu \cdot [\nabla u_h] \|_{0, E}^2 \lesssim (\nu \cdot [\nabla u_h], \varphi)_{0, E} = a(u_h, \varphi),
\]
where we have taken into account that \( \psi_E|_{\partial \Omega_E} = 0 \). Incorporating (2.6.2) with the admissible test function \( v = \varphi \), we obtain
\[
h_E \| \nu \cdot [\nabla u_h] \|_{0, E}^2 \lesssim a(u_h - u, \varphi) + (f_h, \varphi)_{0, \Omega_E} + (f - f_h, \varphi)_{0, \Omega_E} \leq \leq e_{1, \Omega_E} |\varphi|_{1, \Omega_E} + (\|f_h\|_{0, \Omega_E} + \|f - f_h\|_{0, \Omega_E}) \|\varphi\|_{0, \Omega_E}.
\]
Observing \( h_E \approx h_{T_h} \) and using (2.6.27c),(2.6.27d), straightforward estimation of the terms on the right-hand side in the previous inequality allows to conclude. \( \square \)

(2.6.35) Lemma: Let \( E \in \mathcal{E}_h(\Gamma_N) \) and \( T \in \mathcal{T}_h(\Omega) \) such that \( E \in \mathcal{E}_h(T) \). Then, there holds
\[
h_E \|g_h - \nu \cdot \nabla u_h\|_{0, E}^2 \lesssim |e_{1, T}|^2 + \nu_T^2 + \text{osc}_f^2(f) + \text{osc}_E^2(g).
\]
Proof. We set \( \varphi := h_E(g_h - \nu \cdot \nabla u_h) \psi_E \). Applying (2.6.27b) and observing \( \psi_E|_{E'} = 0, E' \in \mathcal{E}_h(T), E' \not= E \), Green’s formula gives
\[
h_E \|g_h - \nu \cdot \nabla u_h\|_{0, E}^2 \lesssim (g_h - \nu \cdot \nabla u_h, \varphi)_{0, E} = (g_h, \varphi)_{0, E} - a(u_h, \varphi).
\]
Choosing the admissible test function \( v = \varphi \) in (2.6.2), we have
\[
a(u, \varphi) - (f, \varphi)_{0, T} - (g, \varphi)_{0, E} = 0.
\]
Adding (2.6.38) to the right-hand side in (2.6.37), it follows that
\[
h_E \|g_h - \nu \cdot \nabla u_h\|_{0, E}^2 \lesssim a(u - u_h, \varphi) - (f_h, \varphi)_{0, T} + (f_h - f, \varphi)_{0, T} + (g_h - g, \varphi)_{0, E} \leq |e_{1, T}| |\varphi|_{1, T} + \|f_h\|_{0, T} \|\varphi\|_{0, T} + \|g - g_h\|_{0, E} \|\varphi\|_{0, E}.
\]
The assertion follows from an application of (2.6.27c)-(2.6.27e) to the right-hand side in (2.6.39). \( \square \)

The previous results can be used to establish the efficiency of the error estimator.

(2.6.40) Theorem: Let \( \eta_h \) and \( \text{osc}_h \) be the residual error estimator and the data oscillations as given by (2.6.16) and (2.6.18). Then, there holds
\[
|e_{1, \Omega}|^2 - \text{osc}_h^2 \lesssim e_h^2.
\]
Proof. The assertion follows easily by collecting the estimates (2.6.29),(2.6.34) and (2.6.36). \( \square \)