3.7. The Ritz–Galerkin Method

In the theory of classical solutions it is natural to use approximation procedures like Finite Difference Methods which are based on a point wise evaluation of functions and differential operators. When dealing with weak solutions, this approach cannot be taken over because point values of functions in $H^m(\Omega)$ are in general not well defined. In fact, the formulation of the original problem as abstract variational problem suggests a different strategy to convert the infinite dimensional setup into a finite dimensional one which then allows a numerical treatment. If the abstract variational problem is to find u in a Hilbert space V such that for some $l \in V'$

$$a(u, v) = l(v) \qquad \forall v \in V$$

then a discretization is obtained by replacing V with a finite dimensional subspace V_h. (Here, $h > 0$ is related to the dimension of V_h and will later play the role of a discretization length.) To get a numerical approximation, the idea of the *Ritz–Galerkin Method* is to find $u_h \in V_h$ such that

$$a(u_h, v) = l(v) \qquad \forall v \in V_h \tag{3.36}$$

which is a finite dimensional problem. By introducing a basis $\{\psi_1, \psi_2, ..., \psi_N\}$ of V_h we can reformulate (3.36) as

$$a(u_h, \psi_i) = l(\psi_i), \quad i = 1, \ldots, N.$$

Using

$$u_h = \sum_{j=1}^{N} z_j \psi_j$$

we get the system of equations

$$\sum_{j=1}^{N} a(\psi_j, \psi_i) z_j = l(\psi_i), \quad i = 1, \ldots, N$$

which we may write in the form

$$Az = b$$

with the matrix $A_{ij} = a(\psi_j, \psi_i)$ and right hand side $b_i = l(\psi_i)$. In engineering science the matrix A is called the *stiffness matrix*. Note that A is positive definite if a is V–elliptic since

$$z^T A z = \sum_{i,j} z_i A_{ij} z_j = a\left(\sum_j z_j \psi_j, \sum_i z_i \psi_i\right) = a(u_h, u_h) \geq C_2 \|u_h\|_V^2$$

In particular, $Az = 0$ implies $u_h = 0$ and thus also $z = 0$ which shows that A is invertible. Consequently, the Ritz–Galerkin solution always exists if a is elliptic. Moreover, if a is symmetric, i.e. $a(u, v) = a(v, u)$ for all $u, v \in V$ then A is a symmetric matrix. The basic estimate on the quality of a numerical approximation u_h based on the sub space V_h is the following lemma.

Lemma 3.31. (Céa Lemma)
Let a be a bilinear, V-elliptic form and let u and u_h be the solutions of the variational problem in V and $V_h \subset V$. Then

$$\|u - u_h\|_V \leq \frac{C_1}{C_2} \inf_{v_h \in V_h} \|u - v_h\|_V$$

where C_1 is the constant of continuity and C_2 the constant of ellipticity of a.

Proof: By definition we have $a(u, v) = l(v)$ for all $v \in V$ and $a(u_h, v) = l(v)$ for all $v \in V_h$. Since $V_h \subset V$ we can subtract the two equations which yields

$$a(u - u_h, v) = 0 \qquad \forall v \in V_h.$$

Now, let $v_h \in V_h$. With $v = v_h - u_h \in V_h$ we have $a(u - u_h, v_h - u_h) = 0$ and

$$\begin{aligned} C_2\|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &\leq C_1\|u - u_h\|_V\|u - v_h\|_V \end{aligned}$$

Hence, it follows that $C_2\|u - u_h\|_V \leq C_1\|u - v_h\|_V$ for all $v_h \in V_h$. ∎

The Céa Lemma gives information about the convergence of the Ritz–Galerkin method. Obviously, the distance between approximation u_h and exact solution u vanishes if the distance of u to the space V_h becomes arbitrarily small. (The distance of a point u to a set V_h is defined by $d(u, V_h) := \inf\{\|u - v_h\|_V : v_h \in V_h\}$.) Consequently, convergence can only be obtained if the spaces V_h become larger for $h \to 0$ and eventually exhaust the Hilbert space V.

Corollary 3.32. *Let $V_h \subset V$ $(h > 0)$ be a family of subspaces with the property*

$$d(u, V_h) \xrightarrow[h \to 0]{} 0 \qquad \forall u \in V.$$

If a is a V–elliptic bilinear form and if u is the solution of an abstract variational problem related to a, the Ritz–Galerkin approximation u_h converges

$$\|u - u_h\|_V \xrightarrow[h \to 0]{} 0.$$

Proof: Immediate consequence of Lemma 3.31. ∎

The Ritz–Galerkin Method turns into an efficient solution algorithm if V_h has good approximation properties and if the stiffness matrix A remains either small or sparse. The latter property is obtained if the basis is chosen almost a–orthogonal so that each row of A has only a few non zero entries. Apart from that, it is useful if the basis is related to the problem under consideration. More precisely, if V is a function space, the basis functions should be selected in such a way that high resolution is possible where the solution is expected to develop important structures. In the next section, a few exemplary choices of V_h are discussed.

3.8. Examples of Finite Element spaces

3.8.1. First remarks. In a first example, we consider a Helmholtz problem on the unit interval with homogeneous Dirichlet conditions

$$\begin{aligned} -u'' + u &= f && \text{in } (0,1) \\ u &= 0 && \text{on } \{0,1\} \end{aligned} \tag{3.37}$$

The corresponding weak formulation is to find $u \in H_0^1(0,1)$ such that

$$\langle u, v \rangle_1 = \langle f, v \rangle_0 \qquad \forall v \in H_0^1(0,1).$$

As finite dimensional subspaces we choose

$$V_h = \{\theta_n \,:\, n = 1, \dots, N\} \qquad h = 1/N$$

where θ_n are orthonormal basis functions in $H_0^1(0,1)$

$$\theta_n(x) = \left(\frac{2}{1 + n^2\pi^2}\right)^{\frac{1}{2}} \sin(n\pi\, x).$$

(Here, h does not correspond to a discretization length. For a derivation of the orthonormal basis see Exercise C.4.) This particular choice is very well suited for the Helmholtz problem because θ_n are eigenfunctions of the Helmholtz operator. Indeed,

$$-\theta_n'' + \theta_n = (1 + n^2\pi^2)\theta_n \qquad \forall n \in \mathbb{N}.$$

Another side effect is that the nested sequence of spaces $(V_{1/N})_{N\in\mathbb{N}}$ exhausts $H_0^1(0,1)$ because each element can be represented in terms of the basis vectors. Finally, the orthonormality of θ_n greatly simplifies the structure of the stiffness matrix. We have

$$A_{ij} = a(\theta_j, \theta_i) = \langle \theta_j, \theta_i \rangle_1 = \delta_{ij}$$

so that A is the identity matrix. Hence, the Ritz–Galerkin solution is just

$$u_h = \sum_{n=1}^{N} z_n \theta_n \qquad \text{with} \qquad z_n = b_n = \langle f, \theta_n \rangle_0 \,.$$

Since $\theta_n \in C^\infty(0,1)$, the approximation u_h is a smooth function and we can check in which sense it approximates a solution of the problem in the classical formulation. Since θ_n are eigenfunctions of the Helmholtz operator, we have

$$-u_h'' + u_h = \sum_{n=1}^{N} (1 + n^2\pi^2) \langle f, \theta_n \rangle_0 \, \theta_n = \sum_{n=1}^{N} \langle f, \psi_n \rangle_0 \, \psi_n$$

where $\psi_n = \sqrt{2}\sin(n\pi\, x)$ forms an orthonormal basis of $\mathbb{L}^2(0,1)$ (see Exercise C.4). Note that

$$\sum_{n=1}^{N} \langle f, \psi_n \rangle_0 \, \psi_n(x) = 2 \sum_{n=1}^{N} \int_0^1 f(y) \sin(n\pi\, y)\, dy \sin(n\pi\, x)$$

is the truncated Fourier sine series of f. In particular, we see that the weak solution of the Helmholtz problem is the limiting case of classical solutions to Helmholtz problems with modified right hand sides. (There is a close relation to the *separation of variables* approach in Section 1.3.1.) Rewriting the approximate solution in terms of ψ_n we find

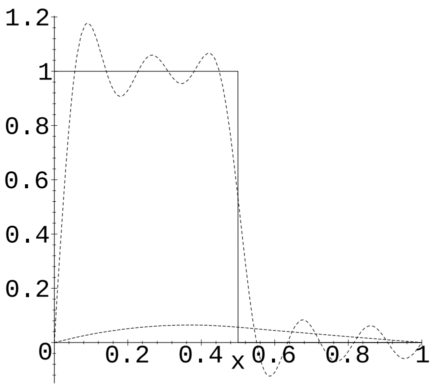
$$u_h = \sum_{n=1}^{N} \frac{1}{1+n^2\pi^2} \langle f, \psi_n \rangle_0 \, \psi_n$$

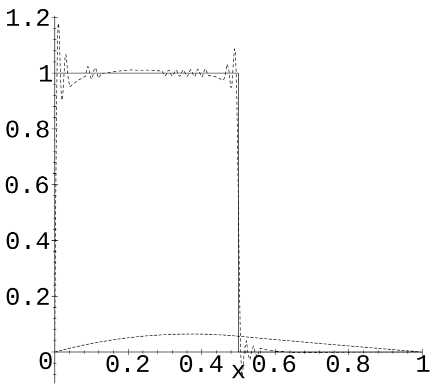
which shows that u_h is obtained from the Fourier sine series of f by damping highly oscillatory terms (large n) which are responsible for a possible rough behavior of f. Consequently, u_h is much smoother than the right hand side which demonstrates the smoothing properties of the inverse Dirichlet Helmholtz operator.

For the case

$$f(x) = \begin{cases} 1 & 0 \le x \le \frac{1}{2} \\ 0 & \frac{1}{2} < x \le 1 \end{cases}$$

with $N = 10$ and $N = 100$, the approximation u_h, the truncated Fourier sine series of f and f itself are shown in the figures below.





$\dim V_h = 10$ (left) and $\dim V_h = 100$ (right)

The simplicity of our first example is related to the availability of a basis which is orthonormal with respect to the bilinear form of the abstract variational problem. As a consequence, $A_{ij} = a(\theta_j, \theta_i)$ is the identity matrix so that the linear problem which determines u_h is trivial. However, for more general geometries or other differential operators the construction of an orthonormal basis is as difficult as solving the original problem so that we have to resort to other basis functions. As a model problem for the general case we consider the Poisson equation on the unit square with homogeneous Dirichlet boundary conditions,

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega = [0,1]^2 \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{3.38}$$

The restriction to the 2D case is only for simplicity. It should be clear, how the ideas can be generalized to other dimensions.

We know from Section 3.1.4 that the bilinear form corresponding to (3.38) is $a(u, v) = \langle \nabla u, \nabla v \rangle_0$ on $H_0^1(\Omega)$. Although a sine basis is again available (see Section 1.3.1), we choose V_h as the set of all functions $v \in C^0(\Omega)$, which are linear in each triangle of the triangulation shown in Fig. 3.1 and which vanish on the boundary of Ω.

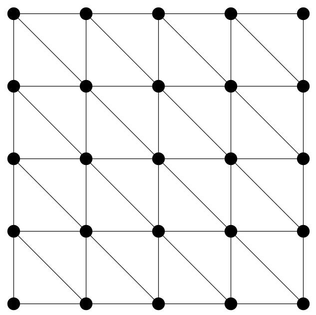


Figure 3.1: Triangulation of the Unit Square

More precisely, in each triangle, a function $v \in V_h$ has the form

$$v(x, y) = a + bx + cy$$

and the three values at the corners of a triangle determine the parameters a, b and c. Consequently, v is given by the values at the N interior grid points (x_j, y_j), $j = 1, \ldots, N$. The dimension of V_h is equal to N and we choose a basis ψ_i, $i = 1, \ldots, N$ with $\psi_i(x_j, y_j) = \delta_{ij}$. Since the basis functions should be linear on each triangle, the support of an exemplary basis function at point number C is non zero on all triangles apart from II and V (see Figure 3.2).

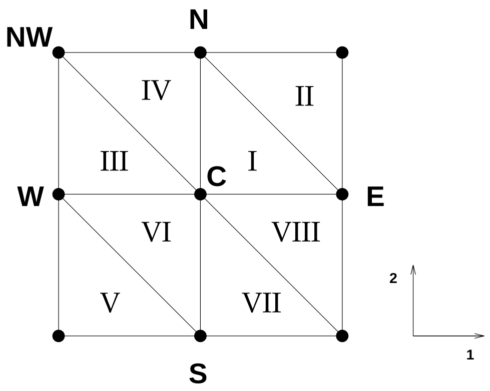


Figure 3.2: Ordering of triangles around point C

Since the values of ψ_C drop linearly from one at C to zero at the neighboring points, the partial derivatives are either zero or constant $\pm 1/h$ on the neighboring triangles ($h > 0$ denotes the spacing of grid points). In the following table, we list the derivatives of the basis function ψ_C on the triangles depicted in Figure 3.2

	I	II	III	IV	V	VI	VII	VIII
$\partial_1\psi_C$	$-h^{-1}$	0	h^{-1}	0	0	h^{-1}	0	$-h^{-1}$
$\partial_2\psi_C$	$-h^{-1}$	0	0	$-h^{-1}$	0	h^{-1}	h^{-1}	0
$\|\nabla\psi_C\|^2$	$2h^{-2}$	0	h^{-2}	h^{-2}	0	$2h^{-2}$	h^{-2}	h^{-2}

The directions 1 and 2 are chosen as indicated in Figure 3.2. Using the fact that the area of a single triangle is $h^2/2$, we get

$$a(\psi_C,\psi_C) = \int\limits_{I,\ldots,VIII} |\nabla\psi_C|^2\,dxdy = 8h^{-2}\frac{1}{2}h^2 = 4.$$

The basis functions are not a–orthogonal because the constant derivatives of neighboring basis functions overlap giving contributions to the bilinear form. For example, ψ_N and ψ_C have overlapping support on the triangles I and IV. We find

$$\begin{aligned} a(\psi_C,\psi_N) = &\int\limits_{I+IV} (\nabla\psi_C)\cdot(\nabla\psi_N)dxdy \\ &= \int\limits_{I+IV} \partial_2\psi_C\partial_2\psi_N dxdy = \int\limits_{I+IV} (-h^{-1})h^{-1}\mathrm{d}x\mathrm{d}y = -1 \end{aligned}$$

Similarly, we obtain

$$a(\psi_C,\psi_E) = a(\psi_C,\psi_S) = a(\psi_C,\psi_W) = a(\psi_C,\psi_N) = -1$$

For the remaining corners, one computes

$$a(\psi_C,\psi_{NW}) = \int\limits_{III+IV} (\partial_1\psi_C\partial_1\psi_{NW} + \partial_2\psi_C\partial_2\psi_{NW})\,\mathrm{d}x\mathrm{d}y = 0$$

and the same for the other points NE, ES and WS. Hence, in the i^{th} row of the matrix $A_{ij} = a(\psi_j,\psi_i)$ there are five non zero entries referring to the point i and its neighboring points in north, south west and east direction. Using the stencil notation to describe the action of A in the neighborhood of the point i, we recover the standard 5 point stencil

$$\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$$

as in the Finite Difference approach. Consequently, in this particular case, the matrix A is the same as in the standard Finite Difference scheme (which is not true in general). Finally, the right hand side of the Ritz–Galerkin Method is obtained by

$$b_i = l(\psi_i) = \int_\Omega f(x)\psi_i(x)\,dx.$$

Observe that, in contrast to the finite difference approach, we use integral averages of f instead of point values.

The resulting system $Az = b$ now has to be solved numerically in order to find the Ritz Galerkin approximation. Note that due to the properties of a, the matrix A is symmetric and positive definite. The system consists of N equations where N is the number of interior nodes. Since the basis functions all have very little overlap, the resulting matrix A is sparse.

The finite dimensional space V_h of piecewise linear functions on Ω is a typical example of a *Finite Element Space.* In general, finite element spaces are characterized by:

(i) The partition of the domain Ω:
The domain Ω is represented as the union of simple subdomains (triangles, rectangles, etc.). The single subdomains are called *elements.*

(ii) The basis functions:
Typically, the basis functions are defined on each element as polynomials which match at the interfaces up to a certain order of differentiability. The polynomial building blocks are called *shape functions.*

Definition 3.33. *A partition $\mathcal{T} = \{T_1, T_2, ..., T_M\}$ of Ω into elements elements is called admissible, if $\bar{\Omega} = \bigcup_{i=1}^{M} T_i$ and if $T_i \cap T_j \neq \emptyset$ the intersection must be a common vertex or a common edge of the elements. If $h > 0$ is the maximal side length of all elements of $\mathcal{T}$, we also write $\mathcal{T}_h$ instead of $\mathcal{T}$.*

Note that a situation as shown in Fig. 3.3 is ruled out in an admissible partition.

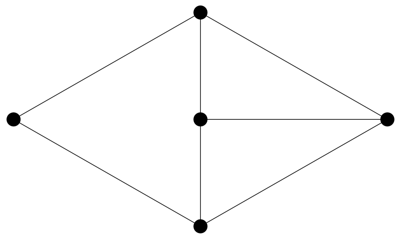


Figure 3.3: Non–admissible partition into three triangles

The reason for the exclusion is that on such a configuration, for example, a piecewise linear function is, in general, not continuous.

3.8.2. A complete example. To give a detailed analysis of a finite element space, we focus on the simple situation of *linear, triangular elements* for Ritz Galerkin approximations in $V = H_0^1(\Omega)$ (i.e. the elements are triangles and the shape functions are linear polynomials). To ensure that we can find a partition of Ω into triangles, we assume that $\Omega \subset \mathbb{R}^2$ is a bounded set with a polygonal boundary. However, the triangles need not be completely regular, as in the second example of the last section. In general, we allow any distortion of the triangles as long as the partition (respectively *triangulation*) is admissible (see Figure 3.4 for examples).

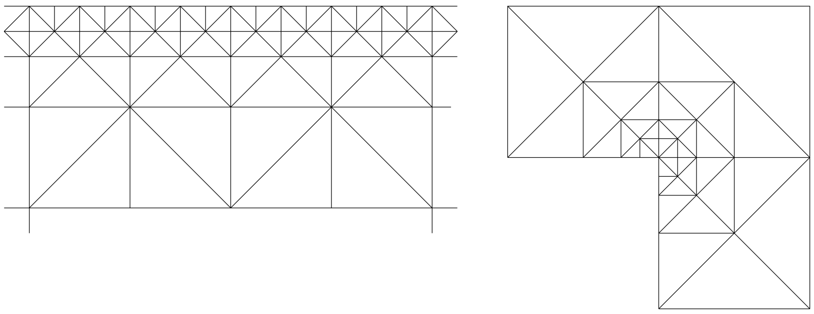


Figure 3.4: Admissible triangulations

The basic finite dimensional space V_h we investigate in the following is defined as

$$V_h := \left\{ u \in C(\bar{\Omega}) \,:\, u|_{\partial\Omega} = 0, \quad u|_{T_i} \text{ is linear for all } T_i \in \mathcal{T}_h \right\}. \tag{3.39}$$

Lemma 3.34. *V_h defined in (3.39) is a linear subspace of $H_0^1(\Omega)$. Every function $v \in V_h$ is uniquely characterized by its values at the interior nodes (i.e. the corners of $T_i \in \mathcal{T}_h$ in the interior of Ω). The dimension of V_h is equal to the number of interior nodes.*

Proof: Since, by definition, each $v \in V_h$ is continuous and vanishes on $\partial\Omega$, it suffices to check $v \in H^1(\Omega)$. Obviously, $v \in \mathbb{L}^2(\Omega)$ since v is continuous and Ω is bounded. To show that also $\frac{\partial v}{\partial x_i}$ are in $\mathbb{L}^2$, we first have to calculate the generalized derivatives. Taking any test function $\phi \in C_0^\infty(\Omega)$, we have

$$\begin{aligned}\left\langle v, \frac{\partial \phi}{\partial x_i} \right\rangle = \int_\Omega v \frac{\partial \phi}{\partial x_i}\, dx = \sum_{T_j \in \mathcal{T}_h} \int_{T_j} v \frac{\partial \phi}{\partial x_i}\, dx \\ = \sum_{T_j \in \mathcal{T}_h} \left(- \int_{T_j} \frac{\partial v}{\partial x_i} \phi\, dx + \int_{\partial T_j} v \phi n_i \, d\sigma \right).\end{aligned}$$

Since ϕ vanishes on $\partial\Omega$, edges of triangles which coincide with parts of the boundary of Ω do not have to be considered. All other integrals along edges of the triangles appear twice in the sum with opposite orientation of the outer normal n but the same values of ϕ and v (due to continuity). Taking the cancellation into account, we find

$$\left\langle v, \frac{\partial \phi}{\partial x_i} \right\rangle = - \sum_{T_j \in \mathcal{T}_h} \int_{T_j} \frac{\partial v}{\partial x_i} \phi \, dx$$

so that the weak derivative of v is, by definition, the piecewise constant function

$$\frac{\partial v}{\partial x_i}(x) = \sum_{T_j \in \mathcal{T}_h} \mathcal{X}_{T_j}(x) \frac{\partial}{\partial x_i} \left(v|_{T_j} \right)$$

which is certainly square integrable on Ω.

Since a linear function on a triangle $T_i \in \mathcal{T}_h$ is uniquely determined by its values at the corners, it follows from $\bar{\Omega} = \bigcup_{i=1}^M T_i$ that $v \in V_h$ is determined by its values at the interior nodes $P_1, \ldots, P_N$.

To show that $\dim V_h = N$, we construct basis functions $\psi_i \in V_h$ by the condition

$$\psi_i(P_j) = \delta_{ij} \qquad i, j = 1, \ldots, N. \tag{3.40}$$

Certainly, on each triangle we can find a linear function satisfying (3.40) at its corners. To show that the resulting function on Ω is continuous, we need that $\mathcal{T}_h$ is an admissible triangulation. This assumption guarantees that each edge of $T_i \in \mathcal{T}_h$ is either part of the boundary or an edge of a neighboring triangle. On a common edge of two neighboring triangles, however, both linear functions are uniquely determined by the linear combination of the two values at the endpoints of the edge which gives continuity. With $\{\psi_i : i = 1, \ldots, N\}$ being a well defined subset of V_h we can now check that it is a basis. Given any $v \in V_h$, the superposition

$$x \mapsto \sum_{i=1}^{N} v(P_i)\psi_i(x)$$

is obviously contained in V_h and coincides with v on all nodes of the partition. Due to the uniqueness result, we thus have

$$v = \sum_{i=1}^{N} v(P_i)\psi_i.$$

The required linear independence of the functions ψ_i follows from the uniqueness of the trivial representation of the zero function. ■

The basis functions $\psi_i \in V_h$ constructed in the proof of Lemma 3.34 are characterized by the condition (3.40). A basis consisting of functions which satisfy such a condition is called *nodal basis*. We also remark that the mapping

$$u \mapsto I_h u := \sum_{i=1}^{N} u(P_i)\psi_i \qquad u \in C(\bar{\Omega})$$

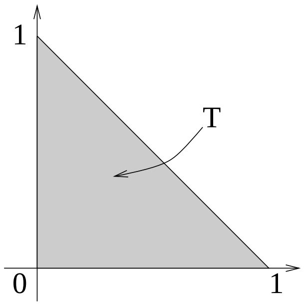
describes a linear interpolation of the nodal values $u(P_i)$ of u.

If we have a family of admissible partitions $\mathcal{T}_h$, it is clear that the dimension of V_h increases for $h \to 0$. In order to apply the Céa Lemma, however, it is necessary to show that V_h comes arbitrarily close to any function $u \in H_0^1(\Omega)$. The idea how to prove that is essentially the following: First, we use the fact that any $u \in H_0^1(\Omega)$ can be approximated by a more regular function $\tilde{u} \in S \cap H_0^1(\Omega)$ (where S can be, for example, $C_0^\infty(\Omega), C^k(\bar{\Omega})$ or $H^2(\Omega)$). For the smooth function $\tilde{u}$ we then show that the distance to the linear interpolation $I_h\tilde{u}$ becomes arbitrarily small in the H^1–norm for $h \to 0$. Altogether, this implies that any $u \in H_0^1(\Omega)$ can be approximated by elements of V_h to any accuracy if h is small enough.

To carry out this program, we choose $S = H^2(\Omega)$. The reason for this choice is that we can get more precise convergence estimates of the Ritz–Galerkin method under the additional assumption that u is H^2–regular (i.e. that $u \in$

$H^2(\Omega) \cap H_0^1(\Omega)$ which is indeed the case if the domain and the data have some extra smoothness).

We start our considerations with the so called *reference triangle* T shown in the figure below.



The main tool to obtain estimates is the Bramble-Hilbert Lemma 3.12.

Lemma 3.35. *Let T be the reference triangle and $u \in H^2(T)$. Then*

$$\|u\|^2_{H^2(T)} \leq C\left(|u(0)|^2 + |u(e_1)|^2 + |u(e_2)|^2 + |u|^2_{2,2,T}\right).$$

Similarly, for $u \in H^2(hT)$ with $h > 0$, we find

$$\|\nabla^\beta u\|^2_{\mathbb{L}^2(hT)} \leq Ch^{2-2|\beta|}\left(|u(0)|^2 + |u(he_1)|^2 + |u(he_2)|^2 + h^2|u|^2_{2,2,hT}\right).$$

Proof: According to Theorem C.2, $H^2(T)$ is embedded in the set of bounded continuous functions on T. In particular, for any $y \in T$, the evaluation functional $\delta_y(u) = u(y)$ is well defined and we have

$$|u(y)| = |\delta_y(u)| \leq \|u\|_{C_B(T)} \leq C\|u\|_{H^2(T)}.$$

Consequently,

$$\mu_0(u) := |u(0)|, \qquad \mu_1(u) := |u(e_1)|, \qquad \mu_2(u) := |u(e_2)|$$

are continuous semi-norms on $H^2(T)$. Moreover, any linear polynomial Q on T which satisfies $\mu_i(Q) = 0$ for all $i = 0, 1, 2$ is necessarily identically zero. Thus we can apply the Bramble-Hilbert Lemma which yields the estimate

$$\|u\|^2_{H^2(T)} \leq C\left(|u|^2_{2,2,T} + \sum_{i=0}^{2} \mu_i^2(u)\right)$$

and thus gives the first part of the Lemma.

The second estimate on the scaled triangle hT is reduced to the previous one by the transformation $v(y) = u(x)$ with $y = hx$ and $dy = h^2\,dx$.

$$\int_{hT} |\nabla_x^\beta u|^2\,dx = \int_T |\nabla_x^\beta u|^2 h^2\,dy = \int_T h^{-2|\beta|}|\nabla_y^\beta v|^2 h^2\,dy. \tag{3.41}$$

Using the result for the reference triangle we get

$$\int_{hT} |\nabla_x^\beta u|^2\,dx \leq Ch^{2-2|\beta|}\left(|v(0)|^2 + |v(e_1)|^2 + |v(e_2)|^2 + |v|^2_{2,2,T}\right).$$

Finally, by using (3.41) with $|\beta| = 2$, we can translate the right hand side back into terms involving u

$$\int_{hT} |\nabla_x^\beta u|^2 \, dx \le C h^{2-2|\beta|} \left(|u(0)|^2 + |u(he_1)|^2 + |u(he_2)|^2 + h^2 |u|^2_{2,2,hT} \right).$$

∎

In order to extend the result to an estimate which applies uniformly to all triangles T_i of the partition $\mathcal{T}_h$, we consider T_i as an affine transformation $\Phi(h_i T)$ of a scaled reference triangle. An estimate as in Lemma 3.35 then follows from uniform bounds on all the involved transformations.

Setting $\Phi(x) = Bx + a$ for $x \in T$, the origin is mapped to a and the edges e_1 and e_2 turn into $\eta = Be_1$ and $\xi = Be_2$.

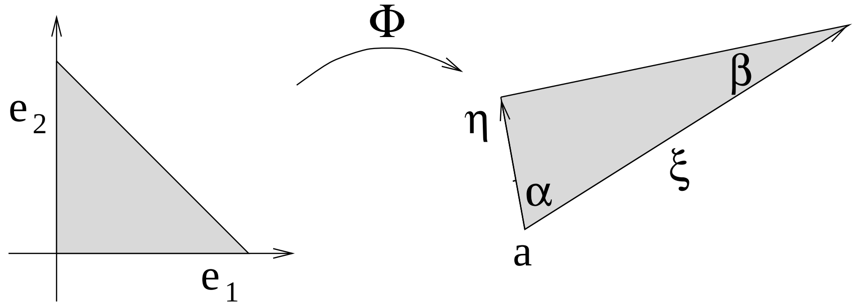


Figure 3.5: Affine linear deformation of the reference triangle.

Since the behavior under scaling is already understood, we just have to investigate distortions. Without loss of generality, we assume $|\eta| \le |\xi| = 1$. Then, all entries of the matrix

$$B = \begin{pmatrix} \eta_1 & \xi_1 \\ \eta_2 & \xi_2 \end{pmatrix}$$

are less than one and the norm of B just depends on the chosen matrix norm (in the matrix norm associated to the Euclidean norm, we have $\|B\| \le 2$). To estimate the norm of the inverse we use the representation

$$B^{-1} = \frac{1}{\det B} \begin{pmatrix} \xi_2 & -\xi_1 \\ -\eta_2 & \eta_1 \end{pmatrix}.$$

so that $\|B^{-1}\| \le 2/|\det B|$. The determinant of B is given by

$$\det B = \eta_1 \xi_2 - \eta_2 \xi_1$$

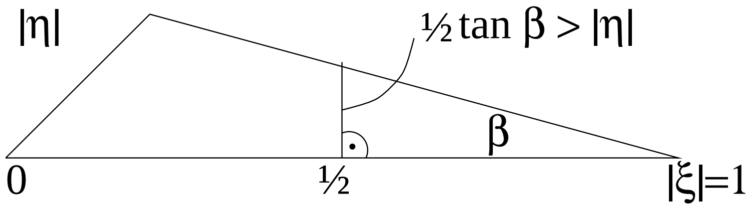
and can be related to geometric properties of the image triangle due to the relation

$$(\eta_1 \xi_2 - \eta_2 \xi_1)^2 = |\eta|^2 |\xi|^2 - |\eta \cdot \xi|^2 = |\eta|^2 |\xi|^2 (1 - \cos^2 \alpha) = |\eta|^2 |\xi|^2 \sin^2 \alpha$$

which implies

$$|\det B| = |\eta| \sin \alpha.$$

(det B describes the volume change under the mapping Φ and thus is twice the area of the image triangle.) Note that $\det B$ is small either if α or the ratio of the two edges $|\eta|/|\xi| = |\eta|$ is close to zero. To get a simple description of these extreme cases, we relate a possible bad ratio of the edges to a condition on the interior angle β (see Figure 3.5). In the case $|\eta| < 1/2$ we can estimate $|eta$ from below by $\frac{1}{2}\tan\beta$ as depicted in the figure below.



Hence, $|\eta| \geq \frac{1}{2}\min\{1, \tan\beta\}$ so that

$$|\det B| \geq \frac{1}{2}\min\{1, \tan\beta\}\sin\alpha$$

and

$$\|B^{-1}\| \leq \frac{4}{\min\{1, \tan\beta\}\sin\alpha}.$$

Consequently, if α_h denotes the smallest interior angle of all triangles in $\mathcal{T}_h$, we have a uniform estimate

$$\det B, \|B\|, (\det B)^{-1}, \|B^{-1}\| \leq C(\alpha_h).$$

Lemma 3.36. *Let $T_i \in \mathcal{T}_h$ be a triangle with corners $P_0^{(i)}, P_1^{(i)}, P_2^{(i)}$. Then, for any $u \in H^2(T_i)$ and $|\beta| \leq 2$ we have*

$$\|\nabla^\beta u\|^2_{\mathbb{L}^2(T_i)} \leq \tilde{C}(\alpha_h)h^{2-2|\beta|}\left(\sum_{j=0}^{2}|u(P_j^{(i)})|^2 + h^2|u|^2_{2,2,T_i}\right)$$

where $\tilde{C}(\alpha_h)$ just depends on the smallest interior angle α_h of all triangles in $\mathcal{T}_h$.

Proof: Let $\Phi(x) = Bx + a$ be the affine linear transformation introduced above. Then T_i is the image of the scaled reference triangle h_iT under Φ where $h_i \leq h$. According to Lemma 3.35 we have for $v \in H^2(h_iT)$

$$\|\nabla^\beta v\|^2_{\mathbb{L}^2(h_iT)} \leq Ch^{2-2|\beta|}\left(|v(0)|^2 + |v(he_1)|^2 + |v(he_2)|^2 + h^2|v|^2_{2,2,h_iT}\right).$$

The effect of the additional distortion due to Φ leads to additional constants which just depend on α_h (due to our control on B). We demonstrate the estimate for the case $|\beta| = 2$.

With $v(y) = u(\Phi(y))$ for $y \in h_iT$, we obtain a function $v \in H^2(h_iT)$ to which we can apply the above estimate. Hence, the main ingredients of the proof are transformation of variables in integrals and the chain rule.

$$\int_{T_i}\left|\frac{\partial^2 u}{\partial x_i \partial x_j}\right|^2 dx = \int_{h_iT}\left|\frac{\partial^2 u}{\partial x_i \partial x_j}(\Phi(y))\right|^2 |\det B|\, dy.$$

Since

$$\frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2}{\partial x_i \partial x_j} v \circ \Phi^{-1} = \sum_{k,l=1}^{n} B_{ki}^{-1} B_{lj}^{-1} \frac{\partial^2 u}{\partial y_k \partial y_l}$$

Cauchy Schwarz inequality yields

$$\left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 \leq \sum_{k,l=1}^{n} \left(B_{ki}^{-1} B_{lj}^{-1} \right)^2 \sum_{k,l=1}^{n} \left| \frac{\partial^2 u}{\partial y_k \partial y_l} \right|^2 .$$

Finally, since $|B_{ki}^{-1}| \leq \|B^{-1}\| \leq C(\alpha_h)$ and $|\det B| \leq C(\alpha_h)$ we get with our result for the scaled triangle $h_i T$

$$\int_{T_i} \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 dx \leq \hat{C}(\alpha_h) h^{2-2|\beta|} \left(|v(0)|^2 + |v(he_1)|^2 + |v(he_2)|^2 + h^2 |v|_{2,2,h_i T}^2 \right) .$$

To obtain the required estimate, $|v|_{2,2,h_i T}^2$ has to be transformed back to the distorted triangle

$$\sum_{k,l=1}^{n} \int_{h_i T} \left| \frac{\partial^2 u}{\partial y_k \partial y_l} \right|^2 dy$$

$$= \sum_{k,l=1}^{n} \frac{1}{\det B} \int_{T_i} \left| \sum_{i,j=1}^{n} B_{ik} B_{jl} \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 dx \leq \frac{\|B\|^2}{\det B} |u|_{2,2,T_i}^2$$

Since $|\det B|^{-1}, \|B\| \leq C(\alpha_h)$, the proof is complete. ■

Now we are in the position to estimate the quality of linear interpolation of H^2–functions.

Theorem 3.37. *Let $\mathcal{T}_h$ be an admissible triangulation with interior nodes $P_1, \ldots, P_N$. and let $\{\psi_1, \ldots, \psi_N\}$ be the nodal basis of V_h. For $u \in H^2(\Omega) \cap H_0^1(\Omega)$ we set*

$$I_h u := \sum_{i=1}^{N} u(P_i) \psi_i$$

and obtain for $m = 0, 1$

$$\|u - I_h u\|_{H^m(\Omega)} \leq \tilde{C}(\alpha_h) h^{2-m} \|u\|_{H^2(\Omega)}$$

where α_h is the smallest interior angle of all triangles in $\mathcal{T}_h$.

Proof: For abbreviation, we set $w = u - I_h u$. Note that w vanishes on all nodes of the partition. We have, using Lemma 3.36, for $|\beta| = m < 2$

$$\|w\|_{H^m(\Omega)}^2 = \sum_{T_i \in \mathcal{T}_h} \|w\|_{H^m(T_i)}^2 \leq \sum_{T_i \in \mathcal{T}_h} \tilde{C}(\alpha_h) h^{4-2m} |w|_{2,2,T_i}^2 .$$

Since all second order derivatives of $I_h u$ vanish on T_i, we find $|w|_{2,2,T_i}^2 = |u|_{2,2,T_i}^2$ so that the result follows. ■

If we have a family of triangulations $\{\mathcal{T}_h : h > 0\}$, the result of Theorem 3.37 applies simultaneously to all $\mathcal{T}_h$ if we can assure that the minimal interior angles are uniformly bounded away from zero. A family of triangulations satisfying this condition is called *quasi-uniform.* Thus, if V_h is constructed from a quasi-uniform family of triangulations we can apply the Céa Lemma and obtain convergence of the Ritz-Galerkin method.

Theorem 3.38. *Let $\{\mathcal{T}_h : h > 0\}$ be a quasi-uniform family of admissible triangulations of Ω with associated spaces $V_h \subset H_0^1(\Omega)$ consisting of continuous piecewise linear functions. Then, for any $u \in H_0^1(\Omega)$ we have*

$$\inf_{v \in V_h} \|u - v\|_{H^1(\Omega)} = d(u, V_h) \xrightarrow[h \to 0]{} 0.$$

If, moreover, $u \in H^2(\Omega) \cap H_0^1(\Omega)$ we get

$$d(u, V_h) \leq Ch \|u\|_{H^2(\Omega)}$$

respectively

$$\inf_{v \in V_h} \|u - v\|_{\mathbb{L}^2(\Omega)} \leq Ch^2 \|u\|_{H^2(\Omega)}.$$

Proof: Theorem 3.37 shows the last two estimates. For the first one, density of $H^2(\Omega) \cap H_0^1(\Omega)$ in $H_0^1(\Omega)$ together with Theorem 3.37 is used. ■

A situation as described in the last theorem is typical in the theory of finite elements: the interpolation is of higher order, if the function, which is to be approximated, possesses more smoothness. In view of the Ritz-Galerkin method, it is therefore of interest, to classify abstract variational problems according to possible extra smoothness of solutions.

Definition 3.39. *Let a be a V–elliptic bilinear form where $H_0^m(\Omega) \subset V \subset H^m(\Omega)$ and $m \geq 1$. The abstract variational problem to find $u \in V$ such that*

$$a(u, v) = \langle f, v \rangle_0, \quad \forall v \in V$$

is called H^s–regular ($s \geq 2m$), if the unique solution $u \in V$ is also in $H^s(\Omega)$ such that

$$\|u\|_s \leq c\|f\|_{s-2m}$$

for all $f \in H^{s-2m}(\Omega)$

For second order problems with homogeneous Dirichlet conditions, one can find regularity theorems of the following kind.

Theorem 3.40. *Let a be a H_0^1–elliptic bilinear form with smooth coefficients*

(i) *If Ω is convex, then the Dirichlet problem in H^2-regular.*
(ii) *Let $s \geq 2$. If Ω has a C^s–boundary, then the Dirichlet problem is H^s-regular.*

Proof: See [**9**]. ■

Up to now, we have assumed that Ω has a polygonal boundary such that admissible triangulations can be constructed. This assumption, is not compatible with $\partial\Omega$ being very smooth. Using the part on convex sets, however, we derive

Theorem 3.41. *Let $\{\mathcal{T}_h : h > 0\}$ be a quasi–uniform family of admissible triangulations of the convex set Ω. Moreover, let a be a $H_0^1(\Omega)$–elliptic bilinear form with smooth coefficients and u as the unique solution of the corresponding abstract variational problem $a(u,v) = \langle f, v\rangle_0$ for all $v \in H_0^1(\Omega)$. Then, the Ritz–Galerkin approximation $u_h \in V_h$ satisfies*

$$\|u - u_h\|_1 \leq ch\|f\|_0$$

Proof: Since the domain Ω is convex, we have H^2–regularity and the estimate $\|u\|_2 \leq c_1\|f\|_0$ holds. Due to Theorem 3.38 we have $d(u, V_h) \leq Ch\|u\|_{H^2}$ so that the Céa Lemma yields the desired result. ■

While Theorem 3.41 yields $\mathcal{O}(h)$ convergence in the H^1-norm, one can expect that a higher order of convergence is obtained if we measure the distance between u and u_h in a weaker norm. Looking at the result of Theorem 3.38 we see that the approximation of H^2-functions by linear elements is indeed one order better in the weaker $\mathbb{L}^2$-norm. However, this result cannot be used directly to get a higher order of convergence because the Céa Lemma only holds for the H^1-distance. One can nevertheless show that second order convergence is obtained in the $\mathbb{L}^2$-norm if the *adjoint problem* is H^2-regular.

Definition 3.42. *Let a be a bilinear form on a Hilbert space V. Then,*

$$a^*(u,v) := a(v,u) \qquad \forall u, v \in V$$

is called the adjoint bilinear form to a. The corresponding abstract variational problem to find $u \in V$ such that

$$a^*(u,v) = l(v) \qquad \forall v \in V$$

is called adjoint problem.

Theorem 3.43. *Under the conditions of Theorem 3.41 we have*

$$\|u - u_h\|_0 \leq ch^2\|f\|_0$$

Proof: Since $a(u,u) = a^*(u,u)$, H_0^1-ellipticity of a implies H_0^1-ellipticity of a^*. Of course, if a has smooth coefficients, the same holds for a^* so that Theorem 3.40 shows that the adjoint problem is H^2-regular. Hence, for any $g \in \mathbb{L}^2(\Omega)$, the solution w of the adjoint problem with linear form $l(v) = \langle g, v\rangle_0$ satisfies $\|w\|_2 \leq C\|g\|_0$. According to Theorem 3.38 we then find

$$\inf_{v_h \in V_h} \|w - v_h\|_1 \leq Ch\|g\|_0. \tag{3.42}$$

(Note that we use C as abbreviation for any constant, i.e. the value of C in different estimates is not necessarily the same.) If u is the solution of

$$a(u,v) = \langle f, v\rangle_0 \qquad \forall v \in H_0^1(\Omega) \tag{3.43}$$

and u_h is the corresponding Ritz-Galerkin approximation, i.e. $a(u_h, v) = \langle f, v\rangle_0$ for all $v \in V_h$, we have, as in the Céa Lemma

$$a(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h. \tag{3.44}$$

The main idea of the proof rests on the relation

$$\|u - u_h\|_0 = \sup_{\|g\|_0=1} |\langle g, u - u_h\rangle_0|.$$

Using the definition of the adjoint problem together with (3.44), we get

$$\begin{aligned} |\langle g, u-u_h\rangle_0| &= |a^*(w, u-u_h)| = |a(u-u_h, w)| \\ &= |a(u-u_h, w-v_h)| \le C\|u-u_h\|_1\|w-v_h\|_1 \end{aligned}$$

Since v_h is an arbitrary element of V_h we can use (3.42) to estimate

$$|\langle g, u-u_h\rangle_0| \le Ch\|u-u_h\|_1\|g\|_0$$

so that

$$\|u-u_h\|_0 \le Ch\|u-u_h\|_1.$$

Finally, we apply Theorem 3.41 which yields

$$\|u-u_h\|_1 \le Ch\|f\|_0$$

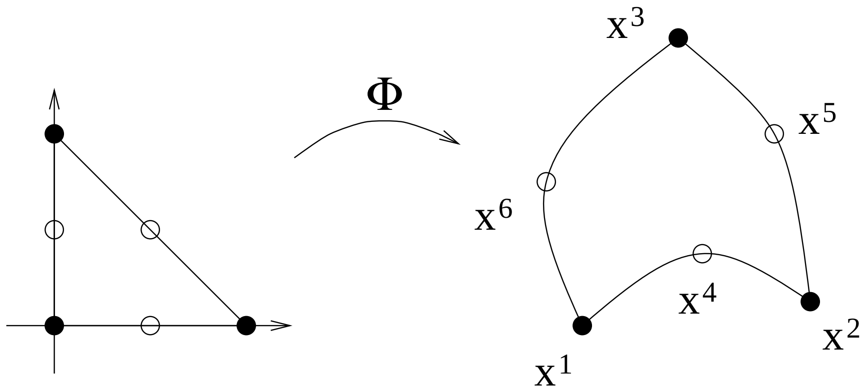
and the proof is complete. ■

3.8.3. Final remarks.

3.8.3.1. *Natural boundary conditions.* Replacing the Dirichlet condition, for example, by natural boundary conditions leads to some obvious modifications in the example of the last section. The nodal basis now includes also piecewise linear functions which are equal to one on boundary nodes. As a consequence, the dimension of the finite element space based on a partition $\mathcal{T}_h$ is larger in comparison with the corresponding space for Dirichlet conditions.

3.8.3.2. *Non-polygonal domains.* If Ω is not a polygonal domain, one can still apply the finite element method. A general approach is to use *isoparametric* finite elements. The idea is to replace the affine linear transformation Φ in Figure 3.5 by a more general, quadratic mapping

$$\Phi(x_1,x_2) = \begin{pmatrix} a_1 + a_2x_1 + a_3x_2 + a_4x_1^2 + a_5x_1x_2 + a_6x_2^2 \\ a_7 + a_8x_1 + a_9x_2 + a_{10}x_1^2 + a_{11}x_1x_2 + a_{12}x_2^2 \end{pmatrix}, \qquad \forall (x_1,x_2) \in T$$



The coefficients $a_1, \dots, a_{12}$ are uniquely determined by the images of corners and edge centers $x^1, \dots x^6$ of the reference triangle. The shape function u on the isoparametric element is defined as the composition of the standard polynomial shape function v on the reference triangle and the mapping Φ^{-1} (i.e. $u = v \circ \Phi^{-1}$). A typical application of isoparametric elements is shown in Figure 3.6. The point x^4 is located on $\partial\Omega$ so that only the corresponding edge is not a linear segment.

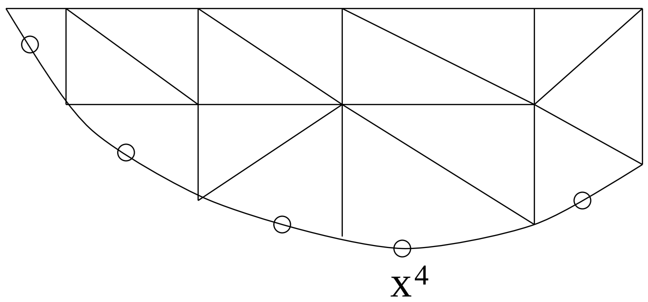


Figure 3.6: Application of isoparametric elements

3.8.3.3. *Non-conforming Finite Elements.* The condition $V_h \subset V$ in the Ritz-Galerkin method characterizes the *conforming Finite Element method.* Discretizations which are based on $V_h \not\subset V$ are called *non-conforming.* This situation appears if the shape functions on the elements are not connected in a sufficiently regular way (e.g. piecewise constant functions are not in H^1 because jumps lead to weak derivatives which are not square integrable). Nevertheless, the approach is appealing especially for higher order problems where the construction of finite dimensional subspaces of sufficiently smooth functions is difficult. If one decides to use non-conforming elements, an immediate problem arises in setting up the stiffness matrix because the original approach to evaluate the bilinear form at the basis functions can no longer be applied. Hence, an additional approximation $a \to a_h$ is required beside the discretization $V \to V_h$. In the case where V_h is based on a partition $\mathcal{T}_h$ and smooth (say polynomial) shape functions which are not connected regularly enough, one can set for example

$$a_h(u,v) := \sum_{T_i \in \mathcal{T}_h} \int_{T_i} \sum_{0\le|\sigma|,|\gamma|\le m} a_{\sigma\gamma} \nabla^\gamma u \nabla^\sigma v \, dx.$$

However, the quality of the approximation $a \to a_h$ now also has to be investigated in order to ensure consistency of the discretization with the original problem.

3.8.3.4. *Other elements.* A generalization of linear triangular elements is obtained if we allow higher order polynomials as shape functions, i.e.

$$\mathcal{P}_d := \left\{ x \mapsto \sum_{|\alpha|\le d} c_\alpha x^\alpha \ : \ c_\alpha \in \mathbb{R} \right\}$$

with $d > 1$. Since $\mathcal{P}_d$ is invariant under affine transformations, it suffices to consider shape functions for the reference triangle. The case of general triangles can always be reduced to the reference situation.

Lemma 3.44. *Let $d \ge 0$. In the reference triangle T we define nodes on $d+1$ parallel lines as shown in Figure 3.7*

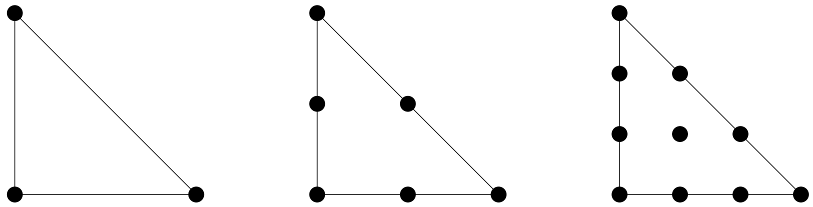


Figure 3.7: Nodes for linear, quadratic and cubic triangular elements

In this way we obtain $s = 1 + 2 + \cdots + (d+1)$ points which we call $z_1, \ldots, z_s$. To any $u \in C(T)$ there exists exactly one polynomial $Q \in \mathcal{P}_d$ which solves the interpolation problem

$$Q(z_i) = u(z_i) \qquad \forall i = 1, \ldots, s.$$

Proof: The proof is done by an induction argument. For $d = 0$, the claim is trivially true so that we can assume it holds for $d-1$. Our first observation is that for the $d+1$ points on the x-axis, we can find a unique polynomial Q_0 which takes the required values $u(z_i)$ in $z_1, \ldots, z_{d+1}$ (Lagrange interpolation). Moreover, due to induction assumption, there exists a polynomial Q_1 of degree $d-1$ which satisfies

$$Q(z_i) = \frac{1}{y_i}(u(z_i) - Q_0(z_i)) \qquad \forall i = d+2, \ldots, s$$

since the remaining nodes $z_{d+2}, \ldots z_s$ are located also on a triangle (which can be obtained by a simple scaling form the reference triangle). Obviously, $Q(x,y) = Q_0(x,y) + yQ_1(x,y)$ solves the interpolation problem. ■

The concept of nodal bases can now easily be taken over.

Definition 3.45. (nodal basis)
Let V_h be a Finite Element space such that the functions in V_h are uniquely determined by their values at a finite number of points. Those functions which are equal to one at exactly one of these points form a basis of V_h. The basis is called nodal basis.

Similar to (3.39) we can now define C^0-elements based on an admissible triangulation $\mathcal{T}_h$ and shape functions from $\mathcal{P}_d$ with $d \geq 1$, i.e.

$$V_h := \left\{ u \in C(\bar{\Omega}) \; : \; u|_{\partial\Omega} = 0, \quad u|_{T_i} \in \mathcal{P}_d \; \forall T_i \in \mathcal{T}_h \right\}.$$

In order to construct a nodal basis of V_h, the $s = (d+1)(d+2)/2$ points in the reference triangle are mapped to each triangle $T_i \in \mathcal{T}_h$ under a suitable affine transformation. In particular, each edge of T_i carries exactly $d+1$ equally spaced points. In a first step, the prescription of zero and one as point values for a nodal basis function gives rise to uniquely defined polynomials on all triangles (due to Lemma 3.44). Considering two neighboring triangles, it remains to show that the corresponding polynomials match continuously at a common edge. Since the restrictions of both polynomials to the edge are again polynomials of degree d (in edge-coordinates) which coincide on $d+1$ separate points they must be identical. Consequently, the nodal basis function is automatically continuous and thus contained in V_h. Similar to Theorem 3.37 we have the result

Theorem 3.46. *Let $d \geq 2$ and $\mathcal{T}_h$ a quasi-uniform triangulation of Ω. Then we have for the interpolation using piecewise polynomials of degree $d-1$*

$$\|u - I_h u\|_{m,h} \leq c \cdot h^{d-m} |u|_{t,\Omega}, \quad u \in H^d(\Omega), \quad 0 \leq m \leq d$$

where the norm $\|\cdot\|_{m,h}$ is defined by

$$\|v\|_{m,h} = \sqrt{\sum_{T_j \in \mathcal{T}_h} \|v\|^2_{m,T_j}}.$$

Proof: The proof is similar to the one of Theorem 3.37. We refer to [**2**] for more details. ∎

If a Finite Element space in H^2 is required (e.g. for the Ritz-Galerkin method applied to a fourth order problem), overall continuity is not enough. Instead, one can show as in Lemma 3.34 that the functions have to be at least in $C^1(\bar{\Omega})$. The construction of so called C^1-elements is much more complicated than that of C^0-elements. As an example, we consider the construction of triangular C^1–elements using polynomials of degree five as shape functions, which was given by Argyris: for polynomials of degree five, we have to determine $(d+1)(d+2)/2|_{d=5} = 21$ degrees of freedom. First, we prescribe for each triangle the values of the polynomial at the three corners up to the second order derivatives, which fixes 18 degrees of freedom. For the *Argyris element*, we determine the resulting 3 degrees of freedom by the values of the normal derivatives at the midpoints of the three edges (see Fig. 3.8).

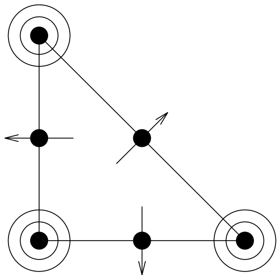


Figure 3.8: Argyris Element (C^1–element)

To see that the *Argyris element* yields a globally C^1–element, we consider the restriction of the polynomials on the edges of the triangles, which are one–dimensional polynomials of degree 5, i.e. we have 6 degrees of freedom. Moreover, the polynomials are determined by the values at the two nodes up to the second derivative. Hence, we have continuity and continuity of the tangential derivatives along the edges. The normal derivative is a polynomial of degree 4, i.e. we have five degrees of freedom. These degrees of freedom are fixed by the values up to the first order derivatives at the nodes and the value at the midpoints of the edges. The normal derivative is therefore uniquely determined and we have continuity of the normal derivative.

Note, that there is a variety of other C^1-elements (see [**2, 4**]).

Finally, let us consider a rectangular C^0-element.

Figure 3.9: Rectangle to define C^0–elements

While in triangular elements the shape functions are polynomials from the full class $\mathcal{P}_d$ for some $d \geq 1$ (so called *complete polynomials*), the shape functions for rectangular elements typically have a tensorial structure, i.e. they are contained in

$$\mathcal{Q}_d = \left\{ (x, y) \mapsto \sum_{0 \leq i,j \leq d} c_{ij} x^i y^j \ : \ c_{ij} \in \mathbb{R} \right\}.$$

In particular, the case $d = 1$ leads to bilinear shape functions

$$u(x, y) = a + bx + cy + dxy.$$

If the corners of all rectangles of a partition $\mathcal{T}_h$ are taken as nodes, the global continuity of the nodal basis functions follows from the fact that the restriction of a bilinear polynomial to an edge of a rectangle is linear and thus uniquely determined by the values at the endpoints of the edge. Again, one can find estimates for the interpolation property of bilinear rectangular elements similar to Theorem 3.37 (see [**2**]).

3.9. Exercises

Exercise 3.1. On $\Omega \subset \mathbb{R}^n$ with $n > 1$ let $L(x, \nabla) = \sum_{|\alpha| \le k} a_\alpha(x) \nabla^\alpha$ be a general linear differential operator.

a) Show that L can only be elliptic if k is even.
b) Show that, if $k = 2m$ and a_α are sufficiently smooth, L can be written in divergence form $L(x, \nabla)u = \sum_{0 \le |\sigma|, |\gamma| \le m} \nabla^\sigma (a_{\sigma\gamma}(x) \nabla^\gamma u)$.

Exercise 3.2. Let $L(x, \nabla) = \sum_{|\alpha| \le m} a_\alpha(x) \nabla^\alpha$ be defined on Ω with C^m coefficients. The formal adjoint L^* of L is defined by $\langle L^*\phi, \psi \rangle := \langle \phi, L\psi \rangle$ for all $\phi, \psi \in \mathscr{D}(\Omega)$.

a) Show that L^* is a linear differential operator.
b) Show that L^*L is in divergence form.
c) Use b) to show that the biharmonic operator $\Delta\Delta$ is in divergence form.

Exercise 3.3. Let $\Omega \subset \mathbb{R}^n$ be a bounded C^1 domain, $f \in C(\Omega)$ and

$$L(x, \nabla)u = \sum_{0 \le |\sigma|, |\gamma| \le m} (-1)^{|\sigma|} \nabla^\sigma (a_{\sigma\gamma}(x) \nabla^\gamma u)$$

with smooth coefficients. Show that a weak solution $u \in H_0^m(\Omega)$ of the Dirichlet problem

$$\begin{aligned} L(x, \nabla)u &= f \qquad \text{in } \Omega \\ \nabla^\alpha u &= 0 \qquad \text{on } \partial\Omega,\ |\alpha| \le m - 1 \end{aligned} \tag{3.45}$$

which satisfies $u \in C^{2m}(\Omega) \cap C^m(\bar{\Omega})$ is also a classical solution (i.e. satisfies (3.45) point-wise).

Exercise 3.4. Let L be a second order, uniformly elliptic differential operator in divergence form with constant coefficients. Show that the corresponding bilinear form is $H_0^1(\Omega)$–elliptic if a_{00} (the coefficient of the zeroth order term in L) is positive, respectively $a_{00} \ge 0$ if Ω is bounded.

Exercise 3.5. Consider the bilinear form $a : H_0^1(0,1) \times H_0^1(0,1) \mapsto \mathbb{R}$ defined by

$$a(u, v) = \int_0^1 x^2 u'(x) v'(x)\, dx.$$

- What is the corresponding (ordinary) differential operator L?
- Show that L is elliptic but not uniformly elliptic.
- Show that a is not $H_0^1(0,1)$–elliptic.

- Show that the functional
$$\Theta(u) = a(u,u)/2 - \int_0^1 u\,dx$$
has no minimum on $H_0^1(0,1)$.

Exercise 3.6. Let $D = \{x \in \mathbb{R}^2 : |x| < 1\}$ be the unit disk in $\mathbb{R}^2$ and $\Omega := D\backslash\{0\}$ the disk without the origin. Consider the Dirichlet problem $\Delta u = 1$ in Ω and $u = 0$ on $\partial\Omega$.

a) Show that the Dirichlet problem has no classical solution $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$.
b) Show that the Dirichlet problem has a unique weak solution $u \in H_0^1(\Omega)$.
c) What is the weak solution? (Hint: Using the fact that $\delta(u) := u(0)$ defined on $C_0^\infty(D)$ cannot be extended to a continuous functional on $H_0^1(\Omega)$ show that $H_0^1(\Omega) = H_0^1(D)$.)
d) Show that the assumption $|\delta(u)| \le C\|u\|_{H^1(D)}$ for all all $C_0^\infty(D)$ functions u (i.e. that δ can be extended to H_0^1) leads to the wrong result that $H_0^1(\mathbb{R}^2)$ is continuously embedded in the space of bounded continuous functions on $\mathbb{R}^2$ (this explains the hint in (c) a little further).

Exercise 3.7. Consider the boundary value problem on $\Omega = (0,1)$
$$\begin{aligned} u'' + p(x)u' + q(x)u &= f(x) \qquad &&\text{in } \Omega \\ u &= 0 &&\text{on } \partial\Omega \end{aligned}$$
where $p \in C^1[0,1]$ and $q \in C[0,1]$. Prove that a unique solution exists if $p' \ge 2q$.

Exercise 3.8. Let $\Omega \subset \mathbb{R}^n$ be a bounded connected domain with C^1–boundary. Let $\emptyset \ne \Gamma_D \subset \partial\Omega$ be open and define
$$H^1_{\Gamma_D} := \{u \in H^1(\Omega) : Bu = 0 \quad \text{on } \Gamma_D\}$$
where $B : H^1(\Omega) \mapsto \mathbb{L}^2(\partial\Omega)$ is the trace operator.

- Derive the classical boundary value problem which corresponds to finding $u \in H^1_{\Gamma_D}(\Omega)$ such that
$$\langle \nabla u, \nabla v\rangle_0 = \langle f, v\rangle_0 \qquad \forall v \in H^1_{\Gamma_D}(\Omega)$$
- Show that $a(u,v) = \langle \nabla u, \nabla v\rangle_0$ is $H^1_{\Gamma_D}(\Omega)$ elliptic. (Hint: use Bramble Hilbert Lemma with a suitable semi norm.)