Chapter II
Conforming Finite Elements

The mathematical treatment of the finite element method is based on the variational formulation of elliptic differential equations. Solutions of the most important differential equations can be characterized by minimal properties, and the corresponding variational problems have solutions in certain function spaces called Sobolev spaces. The numerical treatment involves minimization in appropriate finite-dimensional linear subspaces. A suitable choice for these subspaces, both from a practical and from a theoretical point of view, are the so-called *finite element spaces*.

For linear differential equations, it suffices to work with Hilbert space methods. In this framework, we immediately get the existence of so-called *weak solutions*. Regularity results, to the extent they are needed for the finite element theory, will be presented without proof.

This chapter contains a theory of the simple methods which suffice for the treatment of scalar elliptic differential equations of second order. The aim of this chapter are the error estimates in §7 for the finite element solutions. They refer to the L_2-norm and to the Sobolev norm $\| \cdot \|_1$. Some of the more general results presented here will also be used later in our discussion in Chapter III of other elliptic problems whose treatment requires additional techniques.

The paper of Courant [1943] is generally considered to be the first mathematical contribution to a finite element theory, although a paper of Schellbach [1851] had appeared already a century earlier. If we don't take too narrow a view, finite elements also appear in some work of Euler. The method first became popular at the end of the sixties, when engineers independently developed and named the method. The long survey article of Babuška and Aziz [1972] laid a broad foundation for many of the deeper functional analytic tools, and the first textbook on the subject was written by Strang and Fix [1973].

Independently, the method of finite elements became an established technique in engineering sciences for computations in structural mechanics. The developments there began around 1956, e.g., with the paper of Turner, Clough, Martin, and Topp [1956] who also created the name *finite elements* and the paper by Argyris [1957]. The book by Zienkiewicz [1971] also had great impact. An interesting review of the history was presented by Oden [1991].

§ 1. Sobolev Spaces

In the following, let Ω be an open subset of $\mathbb{R}^n$ with piecewise smooth boundary.

The *Sobolev spaces* which will play an important role in this book are built on the function space $L_2(\Omega)$. $L_2(\Omega)$ consists of all functions u which are square-integrable over Ω in the sense of Lebesgue. We identify two functions u and v whenever $u(x) = v(x)$ for $x \in \Omega$, except on a set of measure zero. $L_2(\Omega)$ becomes a Hilbert space with the scalar product

$$(u, v)_0 := (u, v)_{L_2} = \int_\Omega u(x)v(x)dx \qquad (1.1)$$

and the corresponding norm

$$\|u\|_0 = \sqrt{(u, u)_0}. \qquad (1.2)$$

1.1 Definition. $u \in L_2(\Omega)$ possesses the *(weak) derivative $v = \partial^\alpha u$ in $L_2(\Omega)$* provided that $v \in L_2(\Omega)$ and

$$(\phi, v)_0 = (-1)^{|\alpha|}(\partial^\alpha \phi, u)_0 \quad \text{for all } \phi \in C_0^\infty(\Omega). \qquad (1.3)$$

Here $C^\infty(\Omega)$ denotes the space of infinitely differentiable functions, and $C_0^\infty(\Omega)$ denotes the subspace of such functions which are nonzero only on a compact subset of Ω.

If a function is differentiable in the classical sense, then its weak derivative also exists, and the two derivatives coincide. In this case (1.3) becomes Green's formula for integration by parts.

The concept of the weak derivative carries over to other differential operators. For example, let $u \in L_2(\Omega)^n$ be a vector field. Then $v \in L_2(\Omega)$ is the divergence of u in the weak sense, $v = \operatorname{div} u$ for short, provided $(\phi, v)_0 = -(\operatorname{grad} \phi, u)_0$ for all $\phi \in C_0^\infty(\Omega)$.

Introduction to Sobolev Spaces

1.2 Definition. Given an integer $m \geq 0$, let $H^m(\Omega)$ be the set of all functions u in $L_2(\Omega)$ which possess weak derivatives $\partial^\alpha u$ for all $|\alpha| \leq m$. We can define a scalar product on $H^m(\Omega)$ by

$$(u, v)_m := \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v)_0$$

with the associated norm

$$\|u\|_m := \sqrt{(u, u)_m} = \sqrt{\sum_{|\alpha| \leq m} \|\partial^\alpha u\|^2_{L_2(\Omega)}}. \tag{1.4}$$

The corresponding semi-norm

$$|u|_m := \sqrt{\sum_{|\alpha| = m} \|\partial^\alpha u\|^2_{L_2(\Omega)}} \tag{1.5}$$

is also of interest.

We shall often write H^m instead of $H^m(\Omega)$. Conversely, we will write $\| \cdot \|_{m,\Omega}$ instead of $\| \cdot \|_m$ whenever it is important to distinguish the domain.

The letter H is used in honor of David Hilbert.

$H^m(\Omega)$ is complete with respect to the norm $\| \cdot \|_m$, and is thus a Hilbert space. We shall make use of the following result which is often used to introduce the Sobolev spaces without recourse to the concept of weak derivative.

1.3 Theorem. *Let $\Omega \subset \mathbb{R}^n$ be an open set with piecewise smooth boundary, and let $m \geq 0$. Then $C^\infty(\Omega) \cap H^m(\Omega)$ is dense in $H^m(\Omega)$.*

By Theorem 1.3, $H^m(\Omega)$ is the completion of $C^\infty(\Omega) \cap H^m(\Omega)$, provided that Ω is bounded. This suggests a corresponding generalization for functions satisfying zero boundary conditions.

1.4 Definition. We denote the completion of $C_0^\infty(\Omega)$ w.r.t. the Sobolev norm $\| \cdot \|_m$ by $H_0^m(\Omega)$.

Obviously, the Hilbert space $H_0^m(\Omega)$ is a closed subspace of $H^m(\Omega)$. Moreover, $H_0^0(\Omega) = L_2(\Omega)$, and we have the following inclusions:

$$L_2(\Omega) = H^0(\Omega) \supset H^1(\Omega) \supset H^2(\Omega) \supset \cdots$$
$$\| \qquad\qquad \cup \qquad\qquad \cup$$
$$H_0^0(\Omega) \supset H_0^1(\Omega) \supset H_0^2(\Omega) \supset \cdots$$

The above Sobolev spaces are based on $L_2(\Omega)$ and the L_2-norm. Analogous Sobolev spaces can be defined for arbitrary L_p-norms with $p \neq 2$. They are useful in the study of *nonlinear* elliptic problems. We denote the spaces analogous to H^m and H_0^m by $W^{m,p}$ and $W_0^{m,p}$, respectively.

Friedrichs' Inequality

In spaces with generalized homogeneous boundary conditions, i.e. in H_0^m, the semi-norm (1.5) is equivalent to the norm (1.4).

1.5 Poincaré–Friedrichs Inequality. *Suppose Ω is contained in an n-dimensional cube with side length s. Then*

$$\|v\|_0 \leq s|v|_1 \quad \text{for all } v \in H_0^1(\Omega). \tag{1.6}$$

Proof. Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, it suffices to establish the inequality for $v \in C_0^\infty(\Omega)$. We may assume that $\Omega \subset W := \{(x_1, x_2, \ldots, x_n); \ 0 < x_i < s\}$, and set $v = 0$ for $x \in W \backslash \Omega$. Then

$$v(x_1, x_2, \ldots, x_n) = v(0, x_2, \ldots, x_n) + \int_0^{x_1} \partial_1 v(t, x_2, \ldots, x_n) dt.$$

The boundary term vanishes, and using the Cauchy–Schwarz inequality gives

$$|v(x)|^2 \leq \int_0^{x_1} 1^2 dt \int_0^{x_1} |\partial_1 v(t, x_2, \ldots, x_n)|^2 dt$$

$$\leq s \int_0^s |\partial_1 v(t, x_2, \ldots, x_n)|^2 dt.$$

Since the right-hand side is independent of x_1, it follows that

$$\int_0^s |v(x)|^2 dx_1 \leq s^2 \int_0^s |\partial_1 v(x)|^2 dx_1.$$

To complete the proof, we integrate over the other coordinates to obtain

$$\int_W |v|^2 dx \leq s^2 \int_W |\partial_1 v|^2 dx \leq s^2 |v|_1^2.$$

□

The Poincaré–Friedrichs inequality is often called Friedrichs' inequality or the Poincaré inequality for short.

1.6 Remark. The proof of the Poincaré–Friedrichs inequality only requires zero boundary conditions on a part of the boundary. If $\Gamma = \partial\Omega$ is piecewise smooth, it suffices that the function vanishes on a part of the boundary Γ_D, where Γ_D is a set with positive $(n-1)$-dimensional measure. – Moreover, if zero Dirichlet boundary conditions are prescribed on the whole boundary, then it is sufficient that Ω is located between two hyperplanes whose distance apart is s. □

Applying Friedrichs' inequality to derivatives, we see that

$$|\partial^\alpha v|_0 \leq s|\partial_1 \partial^\alpha v|_0 \quad \text{for } |\alpha| \leq m-1, \ v \in H_0^m(\Omega).$$

Now induction implies

1.7 Theorem. *If Ω is bounded, then $|\cdot|_m$ is a norm on $H_0^m(\Omega)$ which is equivalent to $\|\cdot\|_m$. If Ω is contained in a cube with side length s, then*

$$|v|_m \leq \|v\|_m \leq (1+s)^m |v|_m \quad \textit{for all } v \in H_0^m(\Omega). \tag{1.7}$$

Possible Singularities of H^1 functions

It is well known that $L_2(\Omega)$ also contains unbounded functions. Whether such functions also belong to higher order Sobolev spaces depends on the dimension of the domain. We illustrate this with the most important space $H^1(\Omega)$.

1.8 Remark. If $\Omega = [a, b]$ is a real interval, then $H^1[a, b] \subset C[a, b]$, i.e., each element in $H^1[a, b]$ has a representer which lies in $C[a, b]$.

Proof. Let $v \in C^\infty[a, b]$ or more generally in $C^1[a, b]$. Then for $|x - y| \leq \delta$, the Cauchy–Schwarz inequality gives

$$|v(x) - v(y)| = \left| \int_x^y Dv(t)dt \right| \leq \left| \int_x^y 1^2 dt \right|^{1/2} \cdot \left| \int_x^y [Dv(t)]^2 dt \right|^{1/2} \leq \sqrt{\delta} \, \|v\|_1.$$

Thus, every Cauchy sequence in $H^1[a, b] \cap C^\infty[a, b]$ is equicontinuous and bounded. The theorem of Arzelà-Ascoli implies that the limiting function is continuous. $\square$

The analogous assertion already fails for a two-dimensional domain Ω. The function

$$u(x, y) = \log\log \frac{2}{r}, \tag{1.8}$$

where $r^2 = x^2 + y^2$, is an unbounded H^1 function on the unit disk $D := \{(x, y) \in \mathbb{R}^2; \; x^2 + y^2 < 1\}$. The fact that u lies in $H^1(D)$ follows from

$$\int_0^{1/2} \frac{dr}{r \log^2 r} < \infty.$$

For an n-dimensional domain with $n \geq 3$,

$$u(x) = r^{-\alpha}, \quad \alpha < (n-2)/2, \tag{1.9}$$

is an H^1 function with a singularity at the origin. Clearly, the singularity in (1.9) becomes stronger with increasing n.

The fact that functions in H^2 over a domain in $\mathbb{R}^2$ are continuous will be established in §3 in connection with an imbedding and a trace theorem.
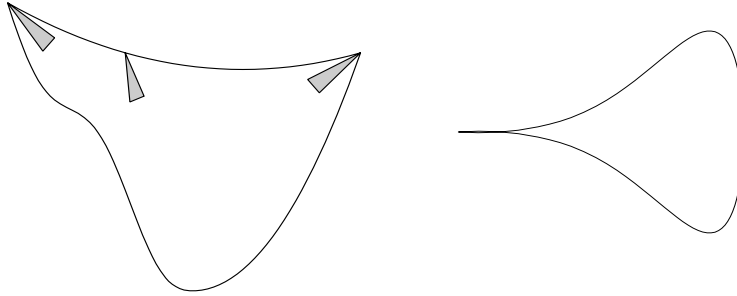
Fig. 6. Domains which satisfy and fail to satisfy the cone condition, respectively

Compact Imbeddings

A continuous linear mapping $L : U \to V$ between normed linear spaces U and V is called *compact* provided that the image of the unit ball in U is a relatively compact set in V. In particular, if $U \subset V$ and the imbedding $J : U \hookrightarrow V$ is compact, we call it a *compact imbedding*.

By the theorem of Arzelà-Ascoli, the C^1 functions v for which

$$\sup_{\Omega} |v(x)| + \sup_{\Omega} |\nabla v(x)| \tag{1.10}$$

is bounded by a given number form a relatively compact subset of $C^0(\Omega)$. The quantity (1.10) is a norm on C^1. In this sense, $C^1(\Omega)$ is compactly imbedded in $C^0(\Omega)$. The analogous assertion also holds for Sobolev spaces, although as we have seen, H^1 functions can exhibit singularities.

1.9 Rellich Selection Theorem. *Given $m \geq 0$, let Ω be a Lipschitz domain,[1] and suppose that it satisfies a cone condition (see Fig. 6), i.e., the interior angles at each vertex are positive, and so a nontrivial cone can be positioned in Ω with its tip at the vertex. Then the imbedding $H^{m+1}(\Omega) \hookrightarrow H^m(\Omega)$ is compact.*

---

[1] A function $f : \mathbb{R}^n \supset D \to \mathbb{R}^m$ is called *Lipschitz continuous* provided that for some number c, $\|f(x) - f(y)\| \leq c\|x - y\|$ for all $x, y \in D$. A hypersurface in $\mathbb{R}^n$ is a *graph* whenever it can be represented in the form $x_k = f(x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$, with $1 \leq k \leq n$ and some suitable domain in $\mathbb{R}^{n-1}$. A domain $\Omega \subset \mathbb{R}^n$ is called a *Lipschitz domain* provided that for every $x \in \partial\Omega$, there exists a neighborhood of $\partial\Omega$ which can be represented as the graph of a Lipschitz continuous function.

Problems

1.10 Let Ω be a bounded domain. With the help of Friedrichs' inequality, show that the constant function $u = 1$ is not contained in $H_0^1(\Omega)$, and thus $H_0^1(\Omega)$ is a proper subspace of $H^1(\Omega)$.

1.11 Let $\Omega \subset \mathbb{R}^n$ be a sphere with center at the origin. Show that $u(x) = \|x\|^s$ possesses a weak derivative in $L_2(\Omega)$ if $2s > 2 - n$ or if $s = 0$ (the trivial case).

1.12 A variant of Friedrichs' inequality. Let Ω be a domain which satisfies the hypothesis of Theorem 1.9. Then there is a constant $c = c(\Omega)$ such that

$$\|v\|_0 \le c\big(|\bar{v}| + |v|_1\big) \quad \text{for all } v \in H^1(\Omega) \tag{1.11}$$

$$\text{with} \quad \bar{v} = \frac{1}{\mu(\Omega)} \int_\Omega v(x) dx.$$

Hint: This variant of Friedrichs' inequality can be established using the technique from the proof of the inequality 1.5 only under restrictive conditions on the domain. Use the compactness of $H^1(\Omega) \hookrightarrow L_2(\Omega)$ in the same way as in the proof of Lemma 6.2 below.

1.13 Let $\Omega_1, \Omega_2 \subset \mathbb{R}^n$ be bounded, and suppose that for the bijective continuously differentiable mapping $F : \Omega_1 \to \Omega_2$, $\|DF(x)\|$ and $\|(DF(x))^{-1}\|$ are bounded for $x \in \Omega$. Verify that $v \in H^1(\Omega_2)$ implies $v \circ F \in H^1(\Omega_1)$.

1.14 Exhibit a function in $C[0, 1]$ which is not contained in $H^1[0, 1]$. – To illustrate that $H_0^0(\Omega) = H^0(\Omega)$, exhibit a sequence in $C_0^\infty(0, 1)$ which converges to the constant function $v = 1$ in the $L_2[0, 1]$ sense.

1.15 Let ℓ_p denote the space of infinite sequences $(x_1, x_2, \dots)$ satisfying the condition $\sum_k |x_k|^p < \infty$. It is a Banach space with the norm

$$\|x\|_p := \|x\|_{\ell_p} := \left(\sum_k |x_k|^p\right)^{1/p}, \quad 1 \le p < \infty.$$

Since $\| \cdot \|_2 \le \| \cdot \|_1$, the imbedding $\ell_1 \hookrightarrow \ell_2$ is continuous. Is it also compact?

1.16 Consider
 (a) the Fourier series $\sum_{k=-\infty}^{+\infty} c_k e^{ikx}$ on $[0, 2\pi]$,
 (b) the Fourier series $\sum_{k,\ell=-\infty}^{+\infty} c_{k\ell} e^{ikx+i\ell y}$ on $[0, 2\pi]^2$.
Express the condition $u \in H^m$ in terms of the coefficients. In particular, show the equivalence of the assertions $u \in L_2$ and $c \in \ell_2$.

Show that in case (b), $u_{xx} + u_{yy} \in L^2$ implies $u_{xy} \in L^2$.

§ 2. Variational Formulation of Elliptic Boundary-Value Problems of Second Order

A function which satisfies a given partial differential equation of second order and assumes prescribed boundary values is called a *classical solution* provided it lies in $C^2(\Omega) \cap C^0(\bar{\Omega})$ in the case of Dirichlet boundary conditions, and in $C^2(\Omega) \cap C^1(\bar{\Omega})$ in the case of Neumann boundary conditions, respectively. Classical solutions exist if the boundary of the underlying domain is sufficiently smooth, and if certain additional conditions are satisfied in the case where Neumann boundary conditions are specified on part of the boundary. In general, higher derivatives of a classical solution need not be bounded (see Example 2.1), and thus the simple convergence theory presented in Ch. I for the finite difference method may not be applicable.

In this section we discuss the variational formulation of boundary-value problems. It provides a natural approach to their numerical treatment using finite elements, and also furnishes a simple way to establish the existence of so-called *weak solutions*.
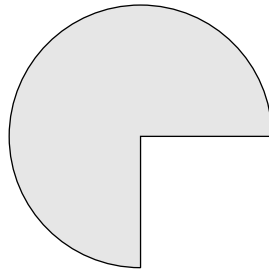
Fig. 7. Domain with reentrant corner (cf. Example 2.1)

2.1 Example. Consider the two-dimensional domain

$$\Omega = \{(x, y) \in \mathbb{R}^2; \ x^2 + y^2 < 1, \ x < 0 \text{ or } y > 0\} \tag{2.1}$$

with reentrant corner (see Fig. 7) and identify $\mathbb{R}^2$ with $\mathbb{C}$. Then $w(z) := z^{2/3}$ is analytic in Ω, and its imaginary part $u(z) := \text{Im } w(z)$ is a harmonic function solving the boundary-value problem

$$\begin{aligned}
\Delta u &= 0 && \text{in } \Omega, \\
u(e^{i\varphi}) &= \sin(\frac{2}{3}\varphi) && \text{for } 0 \le \varphi \le \frac{3\pi}{2}, \\
u &= 0 && \text{elsewhere on } \partial\Omega.
\end{aligned}$$

Since $w'(z) = \frac{2}{3} z^{-1/3}$, even the first derivatives of u are not bounded as $z \to 0$. — The singularity will be no problem when we look for a solution in the right Sobolev space.

Variational Formulation

Before formulating linear elliptic problems as variational problems, we first present the following abstract result.

2.2 Characterization Theorem. *Let V be a linear space, and suppose*

$$a : V \times V \to \mathbb{R}$$

is a symmetric positive bilinear form, i.e., $a(v, v) > 0$ for all $v \in V, v \neq 0$. In addition, let

$$\ell : V \to \mathbb{R}$$

be a linear functional. Then the quantity

$$J(v) := \frac{1}{2}a(v, v) - \langle \ell, v \rangle$$

attains its minimum over V at u if and only if

$$a(u, v) = \langle \ell, v \rangle \quad \text{for all } v \in V. \tag{2.2}$$

Moreover, there is at most one solution of (2.2).

Remark. The set of linear functionals ℓ is a linear space. Instead of $\ell(v)$, we prefer to write $\langle \ell, v \rangle$ in order to emphasize the symmetry with respect to ℓ and v.

Proof. For $u, v \in V$ and $t \in \mathbb{R}$, we have

$$
\begin{aligned}
J(u + tv) &= \frac{1}{2}a(u + tv, u + tv) - \langle \ell, u + tv \rangle \\
&= J(u) + t[a(u, v) - \langle \ell, v \rangle] + \frac{1}{2}t^2 a(v, v). \tag{2.3}
\end{aligned}
$$

If $u \in V$ satifies (2.2), then (2.3) with $t = 1$ implies

$$
\begin{aligned}
J(u + v) &= J(u) + \frac{1}{2}a(v, v) \quad \text{for all } v \in V \tag{2.4} \\
&> J(u), \quad\quad\quad\quad\quad \text{if } v \neq 0.
\end{aligned}
$$

Thus, u is a unique minimal point. Conversely, if J has a minimum at u, then for every $v \in V$, the derivative of the function $t \mapsto J(u + tv)$ must vanish at $t = 0$. By (2.3) the derivative is $a(u, v) - \langle \ell, v \rangle$, and (2.2) follows.   □

The relation (2.4) which describes the size of J at a distance v from a minimal point u will be used frequently below.

Reduction to Homogeneous Boundary Conditions

In the following, let L be a second order elliptic partial differential operator with divergence structure

$$Lu := -\sum_{i,k=1}^{n} \partial_i (a_{ik} \partial_k u) + a_0 u, \qquad (2.5)$$

where

$$a_0(x) \geq 0 \quad \text{for } x \in \Omega.$$

We begin by transforming the associated Dirichlet problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega \end{aligned} \qquad (2.6)$$

into one with homogeneous boundary conditions. To this end, we assume that there is a function u_0 which coincides with g on the boundary and for which Lu_0 exists. Then

$$\begin{aligned} Lw &= f_1 \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega, \end{aligned} \qquad (2.7)$$

where $w := u - u_0$ and $f_1 := f - Lu_0$. For simplicity, we usually assume that the boundary condition in (2.6) is already homogeneous.

We now show that the boundary-value problem (2.7) characterizes the solution of the variational problem. A similar analysis was carried out already by L. Euler, and thus the differential equation $Lu = f$ is called the *Euler equation* or the *Euler–Lagrange* equation for the variational problem.

2.3 Minimal Property. *Every classical solution of the boundary-value problem*

$$\begin{aligned} -\sum_{i,k} \partial_i (a_{ik} \partial_k u) + a_0 u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

is a solution of the variational problem

$$J(v) := \int_\Omega \left[\frac{1}{2} \sum_{i,k} a_{ik} \partial_i v \partial_k v + \frac{1}{2} a_0 v^2 - f v \right] dx \longrightarrow \ \min! \qquad (2.8)$$

among all functions in $C^2(\Omega) \cap C^0(\bar{\Omega})$ with zero boundary values.

Proof. The proof proceeds with the help of Green's formula

$$\int_\Omega v\partial_i w \, dx = -\int_\Omega w\partial_i v \, dx + \int_{\partial\Omega} vw\,\nu_i \, ds. \tag{2.9}$$

Here v and w are assumed to be C^1 functions, and ν_i is the i-th component of the outward-pointing normal ν. Inserting $w := a_{ik}\partial_k u$ in (2.9), we have

$$\int_\Omega v\partial_i (a_{ik}\partial_k u) \, dx = -\int_\Omega a_{ik}\partial_i v\partial_k u \, dx, \tag{2.10}$$

provided $v = 0$ on $\partial\Omega$. Let[2]

$$a(u, v) := \int_\Omega \Big[\sum_{i,k} a_{ik}\partial_i u \, \partial_k v + a_0 uv\Big] dx, \tag{2.11}$$

$$\langle \ell, v\rangle := \int_\Omega f v \, dx.$$

Summing (2.10) over i and k gives that for every $v \in C^1(\Omega) \cap C(\bar\Omega)$ with $v = 0$ on $\partial\Omega$,

$$a(u, v) - \langle \ell, v\rangle = \int_\Omega v\Big[-\sum_{i,k}\partial_i (a_{ik}\partial_k u) + a_0 u - f\Big] dx$$

$$= \int_\Omega v[Lu - f]\,dx \ = 0,$$

provided $Lu = f$. This is true if u is a classical solution. Now the characterization theorem implies the minimal property. $\qquad\qquad\square$

The same method of proof shows that every solution of the variational problem which lies in the space $C^2(\Omega) \cap C^0(\bar\Omega)$ is a classical solution of the boundary-value problem.

The above connection was observed by Thomson in 1847, and later by Dirichlet for the Laplace equation. Dirichlet asserted that the boundedness of $J(u)$ from below implies that J attains its minimum for some function u. This argument is now called the *Dirichlet principle*. However, in 1870 Weierstrass showed that it does not hold in general. In particular, the integral

$$J(u) = \int_0^1 u^2(t)dt \tag{2.12}$$

has infimum 0 in the set $\{v \in C^0[0, 1]; \ v(0) = v(1) = 1\}$, but the value 0 is never assumed for any function in $C[0, 1]$ with the given boundary values.

---

[2] The use of the letter a for the bilinear form and also in the expressions a_{ik} and a_0 for the coefficient functions should be no cause for confusion.

Existence of Solutions

The difficulty with the nonexistence of solutions vanishes if we solve the variational problem (2.8) in a suitable Hilbert space. This is why we don't work in the function space $C^2(\Omega)$, although to get classical solutions this would be desirable. – In Theorem 2.2 only the linear structure was used for the characterization of a solution. But, for existence, the choice of the topology is crucial.

2.4 Definition. Let H be a Hilbert space. A bilinear form $a : H \times H \to \mathbb{R}$ is called *continuous* provided there exists $C > 0$ such that

$$|a(u, v)| \leq C \|u\| \|v\| \quad \text{for all } u, v \in H.$$

A symmetric continuous bilinear form a is called *H-elliptic*, or for short *elliptic* or *coercive*, provided for some $\alpha > 0$,

$$a(v, v) \geq \alpha \|v\|^2 \quad \text{for all } v \in H. \tag{2.13}$$

Clearly, every H-elliptic bilinear form a induces a norm via

$$\|v\|_a := \sqrt{a(v, v)}. \tag{2.14}$$

This is equivalent to the norm of the Hilbert space H. The norm (2.14) is called the *energy norm*.

As usual, the space of continuous linear functionals on a normed linear space V will be denoted by V'.

2.5 The Lax–Milgram Theorem (*for Convex Sets*). *Let V be a closed convex set in a Hilbert space H, and let $a : H \times H \to \mathbb{R}$ be an elliptic bilinear form. Then, for every $\ell \in H'$, the variational problem*

$$J(v) := \frac{1}{2}a(v, v) - \langle \ell, v \rangle \longrightarrow \min !$$

has a unique solution in V.

Proof. J is bounded from below since

$$J(v) \geq \frac{1}{2}\alpha \|v\|^2 - \|\ell\| \|v\|$$
$$= \frac{1}{2\alpha}(\alpha \|v\| - \|\ell\|)^2 - \frac{\|\ell\|^2}{2\alpha} \geq -\frac{\|\ell\|^2}{2\alpha}.$$

Let $c_1 := \inf\{J(v); v \in V\}$, and let (v_n) be a minimizing sequence. Then

$$\alpha \|v_n - v_m\|^2 \le a(v_n - v_m, v_n - v_m)$$
$$= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m)$$
$$= 4J(v_n) + 4J(v_m) - 8J\left(\frac{v_m + v_n}{2}\right)$$
$$\le 4J(v_n) + 4J(v_m) - 8c_1,$$

since V is convex and thus $\frac{1}{2}(v_n + v_m) \in V$. Now $J(v_n), J(v_m) \to c_1$ implies $\|v_n - v_m\| \to 0$ for $n, m \to \infty$. Thus, (v_n) is a Cauchy sequence in H, and $u = \lim_{n \to \infty} v_n$ exists. Since V is closed, we also have $u \in V$. The continuity of J implies $J(u) = \lim_{n \to \infty} J(v_n) = \inf_{v \in V} J(v)$.

We now show that the solution is unique. Suppose u_1 and u_2 are both solutions. Clearly, $u_1, u_2, u_1, u_2, \ldots$ is a minimizing sequence. As we saw above, every minimizing sequence is a Cauchy sequence. This is only possible if $u_1 = u_2$. $\quad\square$

2.6 Remarks. (1) The above proof makes use of the following *parallelogram law*: the sum of the squares of the lengths of the diagonals in any parallelogram is equal to the sum of the squares of the lengths of the sides.

(2) In the special case $V = H$, Theorem 2.5 implies that given $\ell \in H'$, there exists an element $u \in H$ with

$$a(u, v) = \langle \ell, v \rangle \quad \text{for all } v \in H.$$

(3) If we further specialize to the case $a(u, v) := (u, v)$, where (u, v) is the defining scalar product on H, then we obtain the *Riesz representation theorem*: given $\ell \in H'$, there exists an element $u \in H$ with

$$(u, v) = \langle \ell, v \rangle \quad \text{for all } v \in H.$$

This defines a mapping $H' \to H, \ell \mapsto u$ which is called the *canonical imbedding of H' in H*.

(4) The Characterization Theorem 2.2 can be generalized to convex sets as follows. The function u is the minimal solution in a convex set V if and only if the so-called *variational inequality*

$$a(u, v - u) \ge \langle \ell, v - u \rangle \quad \text{for all } v \in V \tag{2.15}$$

holds. We leave the proof to the reader.

If the underlying space has finite dimension, i.e., is the Euclidean space $\mathbb{R}^N$, then instead of (2.13) we only need to require that

$$a(v, v) > 0 \quad \text{for all } v \in H, v \ne 0. \tag{2.16}$$

Then the compactness of the unit ball implies (2.13) for some $\alpha > 0$. The fact that (2.16) does not suffice in the infinite-dimensional case can already be seen in the example (2.12). To make this point even clearer, we consider another simple example.

2.7 Example. Let $H = \ell_2$ be the space of infinite sequences $(x_1, x_2, \ldots)$, equipped with the norm $\|x\|^2 := \sum_m x_m^2$. The form

$$a(x, y) := \sum_{m=1}^{\infty} 2^{-m} x_m y_m$$

is positive and continuous but not coercive, and $\langle \ell, x \rangle := \sum_{m=1}^{\infty} 2^{-m} x_m$ defines a continuous linear functional. However, $J(x) = \frac{1}{2} a(x, x) - \langle \ell, x \rangle$ does not attain a minimum in ℓ_2. Indeed, a necessary condition for a minimal solution in this case is that $x_m = 1$ for $m = 1, 2, \ldots$, and this contradicts $\sum_m x_m^2 < \infty$. □

With the above preparations, we can now make the concept of a solution of the boundary-value problem more precise.

2.8 Definition. A function $u \in H_0^1(\Omega)$ is called a *weak solution* of the second order elliptic boundary-value problem

$$\begin{aligned} Lu &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{2.17}$$

with homogeneous Dirichlet boundary conditions, provided that

$$a(u, v) = (f, v)_0 \quad \text{for all } v \in H_0^1(\Omega), \tag{2.18}$$

where a is the associated bilinear form defined in (2.11).

In other cases we shall also refer to a function as a weak solution of an elliptic boundary-value problem provided it is a solution of an associated variational problem. – Throughout the above, we have implicitly assumed that the coefficient functions are sufficiently smooth. For the following theorem, $a_{ij} \in L_\infty(\Omega)$ and $f \in L_2(\Omega)$ suffice.

2.9 Existence Theorem. *Let L be a second order uniformly elliptic partial differential operator. Then the Dirichlet problem (2.17) always has a weak solution in $H_0^1(\Omega)$. It is a minimum of the variational problem*

$$\frac{1}{2} a(v, v) - (f, v)_0 \longrightarrow \text{min !}$$

over $H_0^1(\Omega)$.

Proof. Let[3] $c := \sup\{|a_{ik}(x)|;\ x \in \Omega,\ 1 \le i, k \le n\}$. Then the Cauchy–Schwarz inequality implies

$$\left|\sum_{i,k} \int a_{ik} \partial_i u \partial_k v\, dx\right| \le c \sum_{i,k} \int |\partial_i u \partial_k v|\, dx$$

$$\le c \sum_{i,k} \left[\int (\partial_i u)^2 dx \int (\partial_k v)^2 dx\right]^{1/2}$$

$$\le C|u|_1\, |v|_1,$$

where $C = cn^2$. If we also assume that $C \ge \sup\{|a_0(x)|;\ x \in \Omega\}$, then we get

$$\left|\int a_0 u v\, dx\right| \le C \int |uv|\, dx \le C \cdot \|u\|_0 \cdot \|v\|_0$$

in an analogous way. Combining these, we have

$$a(u, v) \le C\|u\|_1 \|v\|_1.$$

Next, the uniform ellipticity implies the pointwise estimate

$$\sum_{i,k} a_{ik} \partial_i v \partial_k v \ge \alpha \sum_i (\partial_i v)^2,$$

for C^1 functions. Integrating both sides and using $a_0 \ge 0$ leads to

$$a(v, v) \ge \alpha \sum_i \int_\Omega (\partial_i v)^2 dx = \alpha|v|_1^2 \quad \text{for all } v \in H^1(\Omega). \tag{2.19}$$

By Friedrichs' inequality, $|\cdot|_1$ and $\|\cdot\|_1$ are equivalent norms on H_0^1. Thus, a is an H^1-elliptic bilinear form on $H_0^1(\Omega)$. By the Lax–Milgram Theorem, there exists a unique weak solution which is also a solution of the variational problem. $\square$

2.10 Example. In the model problem

$$-\Delta u = f \quad \text{in } \Omega,$$
$$u = 0 \quad \text{on } \partial\Omega,$$

the associated bilinear form is $a(u, v) = \int \nabla u \cdot \nabla v\, dx$. We will also write $(\nabla u, \nabla v)_0$ for $\int \nabla u \cdot \nabla v\, dx$. Thus, the solution is determined by

$$(\nabla u, \nabla v)_0 = (f, v)_0 \quad \text{for all } v \in H_0^1(\Omega). \tag{2.20}$$

We see that the divergence of ∇u in the sense of Definition 1.1 exists, and $-\Delta u = -\operatorname{div}\operatorname{grad} u = f$.

---

[3] $c, c_1, c_2, \ldots$ are generic constants, i.e. they can change from line to line. In general, we reserve C for the value of the norm of a in the sense of Definition 2.4.

Inhomogeneous Boundary Conditions

We now return to equation (2.6) with inhomogeneous boundary conditions. Let $u_0 \in C^2(\Omega) \cap C^0(\bar{\Omega}) \cap H^1(\Omega)$ be a function which coincides with g on the boundary of Ω. The weak formulation of (2.7) is now

Find $w \in H_0^1(\Omega)$ with

$$a(w, v) = (f - Lu_0, v)_0 \quad \text{for all } v \in H_0^1(\Omega).$$

Since $(Lu_0, v) = a(u_0, v)$, this can now be written in the following form:

Find $u \in H^1(\Omega)$ with

$$
\begin{aligned}
a(u, v) &= (f, v)_0 \quad \text{for all } v \in H_0^1(\Omega), \\
u - u_0 &\in H_0^1(\Omega).
\end{aligned}
\tag{2.21}
$$

The second part of (2.21) can be considered as a weak formulation of the boundary condition.

It follows from density considerations that it suffices to assume that $u_0 \in H^1(\Omega)$. On the other hand, it is not always possible to satisfy this requirement. In fact, it is not even satisfied in some cases for which a classical solution is known.

Example (Hadamard [1932]). Let r and φ be the polar coordinates in the unit disk $\Omega = B_1 := \{x \in \mathbb{R}^2; \ \|x\| < 1\}$. The function $u(r, \varphi) := \sum_{k=1}^{\infty} k^{-2} r^{k!} \sin(k!\varphi)$ is harmonic in Ω. If we identify $\mathbb{R}^2$ with $\mathbb{C}$, then $u(z) = \text{Im} \sum_{k=1}^{\infty} k^{-2} z^{k!}$. This shows that $\int |\nabla u|^2 dx = \infty$, and thus $u \notin H^1$. There does not exist any function in H^1 with the same boundary value as u, since for a given boundary value, the harmonic function is always the one with the smallest value of the H^1-semi-norm.

Problems

2.11 Let Ω be bounded with $\Gamma := \partial\Omega$, and let $g : \Gamma \to \mathbb{R}$ be a given function. Find the function $u \in H^1(\Omega)$ with minimal H^1-norm which coincides with g on Γ. Under what conditions on g can this problem be handled in the framework of this section?

2.12 Consider the elliptic, but not uniformly elliptic, bilinear form

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

on the interval $[0, 1]$. Show that the problem $\frac{1}{2} a(u, u) - \int_0^1 u \, dx \to \min!$ does not have a solution in $H_0^1(0, 1)$. – What is the associated (ordinary) differential equation?

2.13 Prove that in a convex set, the solution to the variational problem is characterized by (2.15).

2.14 In connection with Example 2.7, consider the continuous linear mapping

$$L : \ell_2 \to \ell_2,$$
$$(Lx)_k = 2^{-k} x_k.$$

Show that the range of L is not closed.

Hint: The closure contains the point $y \in \ell_2$ with $y_k = 2^{-k/2}$, $k = 1, 2, \ldots$.

2.15 Show that

$$\int_\Omega \phi \operatorname{div} v \, dx = - \int_\Omega \operatorname{grad} \phi \cdot v \, dx + \int_{\partial\Omega} \phi v \cdot v \, ds \qquad (2.22)$$

for all sufficiently smooth functions v and ϕ with values in $\mathbb{R}^n$ and $\mathbb{R}$, respectively. Here

$$\operatorname{div} v := \sum_{i=1}^n \frac{\partial v}{\partial x_i} \, .$$

2.16 Which variational problem is associated to the boundary-value problem with an ordinary differential equation

$$
\begin{aligned}
u''(x) &= e^x \quad \text{in } (0, 1), \\
u(0) &= u(1) = 0 \, ?
\end{aligned}
\qquad (2.23)
$$

§ 3. The Neumann Boundary-Value Problem.
A Trace Theorem

In passing from a partial differential equation to an associated variational problem, Dirichlet boundary conditions are explicitly built into the function space. This kind of boundary condition is therefore called *essential*. In contrast, Neumann boundary conditions, which are conditions on derivatives on the boundary, are implicitly forced, and thus are called *natural boundary conditions*.

Ellipticity in H^1

Suppose L is the uniformly elliptic differential operator in (2.5), and that a is the corresponding bilinear form (2.11). We now require that $a_0(x)$ be bounded from below by a positive number. After possibly reducing the number α in (2.19), we can assume

$$a_0(x) \geq \alpha > 0 \quad \text{for all } x \in \Omega.$$

Now we get the bound

$$a(v, v) \geq \alpha |v|_1^2 + \alpha \|v\|_0^2 = \alpha \|v\|_1^2 \quad \text{for all } v \in H^1(\Omega), \tag{3.1}$$

which has one more term $\int a_0(x) v^2 dx \geq \alpha \|v\|_0^2$ than the bound in (2.19). Thus, the quadratic form $a(v, v)$ is elliptic on the entire space $H^1(\Omega)$, and not just on the subspace $H_0^1(\Omega)$. In addition, for $f \in L_2(\Omega)$ and $g \in L_2(\partial\Omega)$ we can define a linear functional by

$$\langle \ell, v \rangle := \int_\Omega f v \, dx + \int_\Gamma g v \, ds, \tag{3.2}$$

where as usual, $\Gamma := \partial\Omega$. The following theorem shows that $\langle \ell, v \rangle$ is well defined for all $v \in H^1(\Omega)$, and that ℓ is a bounded linear functional.[4]

3.1 Trace Theorem. *Let Ω be bounded, and suppose Ω has a piecewise smooth boundary. In addition, suppose Ω satisfies the cone condition. Then there exists a bounded linear mapping*

$$\gamma : H^1(\Omega) \to L_2(\Gamma), \quad \|\gamma(v)\|_{0,\Gamma} \leq c\|v\|_{1,\Omega}, \tag{3.3}$$

such that $\gamma v = v_{|\Gamma}$ for all $v \in C^1(\bar\Omega)$.

---

[4] There are sharper results for Sobolev spaces with non-integer indices.

Clearly, γv is the *trace* of v on the boundary, i.e., the restriction of v to the boundary. We know that the evaluation of an H^1 function at a single point does not always make sense. Theorem 3.1 asserts that the restriction of v to the boundary is at least an L_2 function.

We delay the proof of the trace theorem until the end this section.

Boundary-Value Problems with Natural Boundary Conditions

3.2 Theorem. *Suppose the domain Ω satisfies the hypotheses of the trace theorem. Then the variational problem*

$$J(v) := \frac{1}{2}a(v, v) - (f, v)_{0,\Omega} - (g, v)_{0,\Gamma} \longrightarrow \text{ min } !$$

has exactly one solution $u \in H^1(\Omega)$. The solution of the variational problem lies in $C^2(\Omega) \cap C^1(\bar{\Omega})$ if and only if there exists a classical solution of the boundary-value problem

$$Lu = f \quad \text{in } \Omega,$$
$$\sum_{i,k} v_i a_{ik} \partial_k u = g \quad \text{on } \Gamma, \tag{3.4}$$

in which case the two solutions are identical. Here $v := v(x)$ is the outward-pointing normal defined almost everywhere on Γ.

Proof. Since a is an H^1-elliptic bilinear form, the existence of a unique minimum $u \in H^1(\Omega)$ follows from the Lax–Milgram Theorem. In particular, u is characterized by

$$a(u, v) = (f, v)_{0,\Omega} + (g, v)_{0,\Gamma} \quad \text{for all } v \in H^1(\Omega). \tag{3.5}$$

Now suppose (3.5) is satisfied for $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$. For $v \in H_0^1(\Omega)$, $\gamma v = 0$, and we deduce from (3.5) that

$$a(u, v) = (f, v)_0 \quad \text{for all } v \in H_0^1(\Omega).$$

By (2.21), u is also a solution of the Dirichlet problem, where we define the boundary condition using u. Thus, in the interior we have

$$Lu = f \quad \text{in } \Omega. \tag{3.6}$$

For $v \in H^1(\Omega)$, Green's formula (2.9) yields

$$\int_\Omega v\partial_i(a_{ik}\partial_k u)\,dx = -\int_\Omega \partial_i v a_{ik} \partial_k u\,dx + \int_\Gamma v a_{ik}\partial_k u\,v_i\,ds.$$

Hence,

$$a(u,v) - (f,v)_0 - (g,v)_{0,\Gamma} = \int_\Omega v[Lu - f]\, dx + \int_\Gamma [\sum_{i,k} \nu_i a_{ik} \partial_k u - g] v\, ds. \quad (3.7)$$

Now it follows from (3.5) and (3.6) that the second integral on the right-hand side of (3.7) vanishes. Suppose the function $v_0 := \nu_i a_{ik} \partial_k u - g$ does not vanish. Then $\int_\Gamma v_0^2 ds > 0$. Since $C^1(\bar\Omega)$ is dense in $C^0(\bar\Omega)$, there exists $v \in C^1(\bar\Omega)$ with $\int_\Gamma v_0 \cdot v\, ds > 0$. This is a contradiction, and the boundary condition must be satisfied.

On the other hand, from (3.7) we can immediately see that every classical solution of (3.4) satisfies (3.5).   □

Neumann Boundary Conditions

For the Helmholtz equation

$$-\Delta u + a_0(x)\, u = f \quad \text{in } \Omega,$$

the natural boundary condition is

$$\frac{\partial u}{\partial \nu} := \nu \cdot \nabla u = g \quad \text{on } \partial\Omega.$$

We call it *the Neumann boundary condition*. Here $\partial u / \partial \nu$ is the normal derivative, i.e., the direction perpendicular to the tangent plane (if the boundary is smooth). [In the general case, the boundary condition in (3.4) also involves the normal direction if we define orthogonality w.r.t. the metric induced by the quadratic form with the matrix $a_{ik} = a_{ik}(x)$.]

Clearly, the Poisson equation with Neumann boundary conditions,

$$\begin{aligned}
-\Delta u &= f \quad \text{in } \Omega, \\
\frac{\partial u}{\partial \nu} &= g \quad \text{on } \partial\Omega,
\end{aligned} \quad (3.8)$$

only determines a function up to an additive constant. This suggests that in formulating the weak version of this problem we should restrict ourselves to the subspace $V := \{v \in H^1(\Omega);\ \int_\Omega v\, dx = 0\}$. The bilinear form $a(u,v) = \int_\Omega \nabla u \cdot \nabla v\, dx$ is not H^1-elliptic, but in view of the variant (1.11) of Friedrichs' inequality, it is V-elliptic.

We claim that the data of the boundary-value problem (3.8) must satisfy a certain *compatibility condition*. Indeed, with $w := \nabla u$, equation (3.8) becomes

$$- \operatorname{div} w = f \text{ in } \Omega, \quad \nu' w = g \text{ on } \Gamma.$$

By the Gauss Integral Theorem, $\int_\Omega \operatorname{div} w \, dx = \int_{\partial\Omega} w\nu \, ds$, and thus

$$\int_\Omega f \, dx + \int_\Gamma g \, ds = 0. \tag{3.9}$$

This condition is not only necessary, but also sufficient. By the Lax–Milgram Theorem, we get $u \in V$ with

$$a(u, v) = (f, v)_{0,\Omega} + (g, v)_{0,\Gamma} \tag{3.10}$$

for all $v \in V$. Because of the compatibility condition, (3.10) also holds for $v = $ const, and thus for all $v \in H^1(\Omega)$. As in Theorem 3.2, we now deduce that every classical solution of the variational problem satisfies the equation (3.8).

Another method to deal with the pure Neumann problem (3.8) will be discussed in Problem III.4.21.

Mixed Boundary Conditions

In physical problems, we often encounter Neumann or natural boundary conditions whenever the flow over the boundary is prescribed. Sometimes a Neumann condition is prescribed on only part of the boundary.

3.3 Example. Suppose we want to determine the stationary temperature distribution in an isotropic body $\Omega \subset \mathbb{R}^3$. On the part of the boundary where the body is mechanically clamped, the temperature is prescribed. We denote this part of the boundary by Γ_D. On the rest of the boundary $\Gamma_N = \Gamma \setminus \Gamma_D$, we assume that the heat flux is so small that it can be considered to be 0. If there are no heat sources in Ω, then we have to solve the elliptic boundary-value problem

$$\begin{aligned}
\Delta u &= 0 \text{ in } \Omega, \\
u &= g \text{ on } \Gamma_D, \\
\frac{\partial u}{\partial \nu} &= 0 \text{ on } \Gamma_N.
\end{aligned}$$

This problem leads in a natural way to a Hilbert space which lies between $H^1(\Omega)$ and $H_0^1(\Omega)$. Consider functions of the form

$$u \in C^\infty(\Omega) \cap H^1(\Omega), \ u \text{ vanishes in a neighborhood of } \Gamma_D.$$

Then the closure of this set w.r.t. the H^1-norm leads to the desired space. This is a subspace of $H^1(\Omega)$, and by Remark 1.6, under very general hypotheses $|\cdot|_1$ is a norm which is equivalent to $\|\cdot\|_1$.

Proof of the Trace Theorem

We now present the proof of the trace theorem. For the sake of clarity, we restrict ourselves to domains in $\mathbb{R}^2$. The generalization to domains in $\mathbb{R}^n$ is straightforward, and can be left to the reader as an exercise.

Suppose the boundary is piecewise smooth. In addition, suppose a cone condition is satisfied at the (finitely many) points where the boundary is not smooth. Then we can divide the boundary into finitely many boundary pieces $\Gamma_1, \Gamma_2, \ldots, \Gamma_m$ so that for every piece Γ_i, after a rotation of the coordinate system, we have

1. For some function $\phi = \phi_i \in C^1[y_1, y_2]$,
$$\Gamma_i = \{(x, y) \in \mathbb{R}^2; \ x = \phi(y), \ y_1 \le y \le y_2\}.$$

2. The domain $\Omega_i = \{(x, y) \subset \mathbb{R}^2; \ \phi(y) < x < \phi(y) + r, \ y_1 < y < y_2\}$ is contained in Ω, where $r > 0$.

We now apply an argument used earlier for the Poincaré–Friedrichs inequality. For $v \in C^1(\bar{\Omega})$ and $(x, y) \in \Gamma$,

$$v(\phi(y), y) = v(\phi(y) + t, y) - \int_0^t \partial_1 v(\phi(y) + s, y) ds,$$

where $0 \le t \le r$. Integrating over t from 0 to r gives

$$r v(\phi(y), y) = \int_0^r v(\phi(y) + t, y) dt - \int_0^r \partial_1 v(\phi(y) + t, y)(r - t) dt.$$

We take the square of this equation, and use *Young's inequality* $(a+b)^2 \le 2a^2 + 2b^2$. Applying the Cauchy–Schwarz inequality to the squares of the integrals gives

$$r^2 v^2(\phi(y), y) \le 2 \int_0^r 1 dt \int_0^r v^2(\phi(y) + t, y) dt$$
$$+ 2 \int_0^r t^2 dt \int_0^r |\partial_1 v(\phi(y) + t, y)|^2 dt.$$

We now insert the values $\int 1 dt = r$ and $\int t^2 dt = r^3/3$. Dividing by r^2 and integrating over y, we get

$$\int_{y_1}^{y_2} v^2(\phi(y), y) dy \le 2r^{-1} \int_{\Omega_i} v^2 dx dy + r \int_{\Omega_i} |\partial_1 v|^2 dx dy.$$

The arc length differential on Γ is given by $ds = \sqrt{1 + \phi'^2} dy$. Thus, we have

$$\int_{\Gamma_i} v^2 ds \le c_i [2r^{-1} \|v\|_0^2 + r|v|_1^2],$$

where $c_i = \max\{\sqrt{1 + \phi'^2}; \ y_1 \le y \le y_2\}$. Setting $c = (r + 2r^{-1}) \sum_{i=1}^m c_i$, we finally get

$$\|v\|_{0,\Gamma} \le c \|v\|_{1,\Omega}.$$

Thus, the restriction $\gamma : H^1(\Omega) \cap C^1(\bar{\Omega}) \to L_2(\Gamma)$ is a bounded mapping on a dense set. Because of the completeness of $L_2(\Gamma)$, it can be extended to all of $H^1(\Omega)$ without enlarging the bound. $\square$

Note that the cone condition excludes *cusps* in the domain. The domain

$$\Omega := \{(x, y) \in \mathbb{R}^2;\ 0 < y < x^5 < 1\}$$

has a cusp at the origin, and $H^1(\Omega)$ contains the function

$$u(x, y) = x^{-1},$$

whose trace is not square-integrable over Γ. $\square$

We would like to point out that Green's formula (2.9) also holds for functions $u, w \in H^1(\Omega)$, provided that Ω satisfies the hypotheses of the trace theorem.

The space $H^1(\Omega)$ is isomorphic to a direct sum

$$H^1(\Omega) \sim H_0^1(\Omega) \oplus \gamma(H^1(\Omega)).$$

Specifically, every $u \in H^1(\Omega)$ can be decomposed as

$$u = v + w,$$

according to the following rule. Let w be the solution of the variational problem $|w|_1^2 \to$ min ! More exactly, suppose

$$
\begin{aligned}
(\nabla w, \nabla v)_{0,\Omega} &= 0 \qquad \text{for all } v \in H_0^1(\Omega), \\
w - u &\in H_0^1(\Omega).
\end{aligned}
$$

Let $v := u - w \in H_0^1(\Omega)$. Here γ is an injective mapping on the set of functions w which appear in the decomposition.

We now consider the connection with continuous functions. As usual, the norm $\|u\|_\infty = \|u\|_{\infty,\Omega}$ is based on the essential supremum of $|u|$ over Ω. [It is not a Sobolev norm.]

3.4 Remarks. (1) Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain, or a domain with Lipschitz continuous boundary. Then $H^2(\Omega)$ is compactly imbedded in $C(\bar{\Omega})$, and

$$\|v\|_\infty \le c\|v\|_2 \quad \text{for all } v \in H^2(\Omega), \tag{3.11}$$

for some number $c = c(\Omega)$.

(2) For every open connected domain $\Omega \subset \mathbb{R}^2$, $H^2(\Omega)$ is compactly imbedded in $C(\Omega)$.

The above results are not the sharpest possible in this framework. Because of their importance, and because they follow simply from the trace theorem, we now give their proofs.

Choose an angle φ and a radius r with the following property: for every two points $x, y \in \bar{\Omega}$ with $\|x - y\| < r$, there exists a cone K with angle φ, diameter r, and tip at x such that $y \in \partial K$. We now rotate the coordinate system so that x and y differ in only the first coordinate. By the trace theorem, we deduce that

$$\|v\|_{0,\partial K} \leq c(r, \varphi) \|v\|_{1,K}.$$

Then for $v \in H^2(\Omega)$, $\partial_1 v \in H^1(\Omega)$, and thus $\|\partial_1 v\|_{0,\partial K} \leq c(r, \varphi)\|v\|_{2,\Omega}$. By Remark 1.8, $|v(x) - v(y)| \leq c(r, \varphi) \sqrt{\|x - y\|} \cdot \|v\|_{2,\Omega}$. Thus, v is Hölder continuous with exponent 1/2. Using $\|v\|_{0,K} \leq \|v\|_{2,\Omega}$, we get a bound for $|v|$ in K, and (3.11) follows. The Arzelà–Ascoli Theorem now establishes the compactness assertion.

For every $m \geq 1$, we can find a polygonal domain Ω_m which contains all of the points $x \in \Omega$ whose distance from 0 is at most m, and whose distance from $\partial\Omega$ is at least $1/m$. Since $\Omega = \bigcup_{m>0} \Omega_m$, (2) follows from (1). □

Practical Consequences of the Trace Theorem

For practical applications, it is tempting to believe that only classical solutions are of importance, and that weak solutions with singularities are nothing more than interesting mathematical objects. However, this is far from the truth, as the following example shows.

3.5 Example. Suppose we erect a *tent* over a disk with radius R such that its height at the center is 1. Find the shape of the tent which has the minimal surface area. Suppose $u(x)$ is the height of the tent at the point x. Then it is well known that the surface area is given by

$$\int_{B_R} \sqrt{1 + (\nabla u)^2} \, dx.$$

Here B_R is the disk with radius R and center at 0. Now $\sqrt{1 + (\nabla u)^2} \leq 1 + \frac{1}{2}(\nabla u)^2$, and for small gradients the difference between the two sides of the inequality is small. Thus, we are led to the variational problem

$$\frac{1}{2} \int_{B_R} (\nabla v)^2 dx \longrightarrow \quad \min ! \tag{3.12}$$

subject to the constraints

$$v(0) = 1,$$
$$v = 0 \quad \text{on } \partial B_R.$$

We now show that the constraint $v(0) = 1$ will be ignored by the solution of the variational problem. The singular function

$$w_0(x) = \log\log\frac{eR}{r} \quad \text{with } r = r(x) = \|x\|$$

has a finite Dirichlet integral (3.12). We smooth it to get

$$w_\varepsilon(x) = \begin{cases} w_0(x) & \text{for } r(x) \geq \varepsilon, \\[2mm] \log\log\dfrac{eR}{\varepsilon} & \text{for } 0 \leq r(x) < \varepsilon. \end{cases}$$

Now $|w_\varepsilon|_{1,B_R} \leq |w_0|_{1,B_R}$, and $w_\varepsilon(0)$ tends to ∞ as $\varepsilon \to 0$. Thus,

$$u_\varepsilon = \frac{w_\varepsilon(x)}{w_\varepsilon(0)}$$

for $\varepsilon = 1, 1/2, 1/3, \ldots$ provides a minimizing sequence for $J(v)$ which converges almost everywhere to the zero function. This means that the requirement $u(0) = 1$ was ignored.

The situation is different if we require that $u = 1$ on a curve segment. This would be the case if the tent were attached to a ring on the tent pole or if the tent is put over a rope so that it assumes the shape of a roof. While the evaluation of an H^1 function at a point does not make any sense, its evaluation on a line in the L_2 sense is possible. A condition on a function which is defined on a curve segment will be respected almost everywhere.
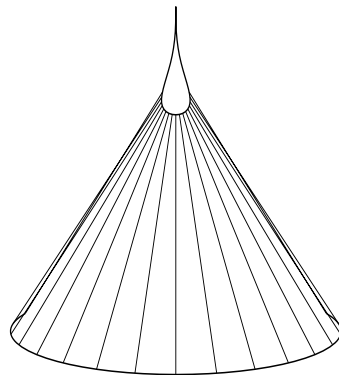

Fig. 8. Tent attached to a loop of a rope to prevent an extreme force concentration at the tip

For most larger tents, the boundary of the tent at the tip is a ring instead of a single point. Or (see Fig. 8) the tent may be attached to a loop of rope. This avoids very high forces, since the force applies to the ring or loop, rather than at a single point. The trace theorem explains why the point is to be replaced by a one-dimensional curve.

Problems

3.6 Show that every classical solution of the equations and inequalities

$$-\Delta u + a_0 u = f \quad \text{in } \Omega,$$

$$\left.\begin{array}{c} u \geq 0, \quad \dfrac{\partial u}{\partial \nu} \geq 0, \\[2mm] u \cdot \dfrac{\partial u}{\partial \nu} = 0 \end{array}\right\} \text{ on } \Gamma,$$

is a solution of a variational problem in the convex set

$$V^+ := \{v \in H^1(\Omega); \ \gamma v \geq 0 \text{ almost everywhere on } \Gamma\}.$$

– It is known from integration theory that the subset $\{\phi \in L_2(\Gamma); \ \phi \geq 0$ almost everywhere on $\Gamma\}$ is closed in $L_2(\Gamma)$.

3.7 Suppose the domain Ω has a piecewise smooth boundary, and let $u \in H^1(\Omega) \cap C(\bar{\Omega})$. Show that $u \in H_0^1(\Omega)$ is equivalent to $u = 0$ on $\partial\Omega$.

3.8 Suppose the domain Ω is divided into two subdomains Ω_1 and Ω_2 by a piecewise smooth curve Γ_0. Let $\alpha_1 \gg \alpha_2 > 0$ and $a(x) = \alpha_i$ for $x \in \Omega_i$, $i = 1, 2$. Show that for every classical solution of the variational problem

$$\int_\Omega [\frac{1}{2} a(x)(\nabla v(x))^2 - f(x)v(x)]dx \longrightarrow \text{ min!}$$

in $H_0^1(\Omega)$, the quantity $a(x)\frac{\partial u}{\partial n}$ is continuous on the curve Γ_0. [The discontinuity of $a(x)$ now implies that $\frac{\partial u}{\partial n}$ is not continuous there.]

§ 4. The Ritz–Galerkin Method
and Some Finite Elements

There is a simple natural approach to the numerical solution of elliptic boundary-value problems. Instead of minimizing the functional J defining the corresponding variational problem over all of $H^m(\Omega)$ or $H_0^m(\Omega)$, respectively, we minimize it over some suitable finite-dimensional subspace [Ritz 1908]. The standard notation for the subspace is S_h. Here h stands for a discretization parameter, and the notation suggests that the approximate solution will converge to the true solution of the given (continuous) problem as $h \to 0$.

We first consider approximation in general subspaces, and later show how to apply it to a model problem.

The solution of the variational problem

$$J(v) := \frac{1}{2} a(v, v) - \langle \ell, v \rangle \longrightarrow \min_{S_h} ! \tag{4.1}$$

in the subspace S_h can be computed using the Characterization Theorem 2.2. In particular, u_h is a solution provided

$$a(u_h, v) = \langle \ell, v \rangle \quad \text{for all } v \in S_h. \tag{4.2}$$

Suppose $\{\psi_1, \psi_2, \ldots, \psi_N\}$ is a basis for S_h. Then (4.2) is equivalent to

$$a(u_h, \psi_i) = \langle \ell, \psi_i \rangle, \quad i = 1, 2, \ldots, N.$$

Assuming u_h has the form

$$u_h = \sum_{k=1}^{N} z_k \psi_k, \tag{4.3}$$

we are led to the system of equations

$$\sum_{k=1}^{N} a(\psi_k, \psi_i) z_k = \langle \ell, \psi_i \rangle, \quad i = 1, 2, \ldots, N, \tag{4.4}$$

which we can write in matrix-vector form as

$$Az = b, \tag{4.5}$$

where $A_{ik} := a(\psi_k, \psi_i)$ and $b_i := \langle \ell, \psi_i \rangle$. Whenever a is an H^m-elliptic bilinear form, the matrix A is positive definite:

$$
\begin{aligned}
z' A z &= \sum_{i,k} z_i A_{ik} z_k \\
&= a\Big(\sum_k z_k \psi_k, \sum_i z_i \psi_i\Big) = a(u_h, u_h) \\
&\geq \alpha \|u_h\|_m^2,
\end{aligned}
$$

and so $z' A z > 0$ for $z \neq 0$. Here we have made use of the bijective mapping $\mathbb{R}^N \longrightarrow S_h$ which is defined by (4.3). Without explicitly referring to this canonical mapping, in the sequel we will identify the function space S_h with $\mathbb{R}^N$.

In engineering sciences, and in particular if the problem comes from continuum mechanics, the matrix A is called the *stiffness matrix* or *system matrix*.

Methods. There are several related methods:

Rayleigh–Ritz Method: Here the minimum of J is sought in the space S_h. Instead of the basis-free derivation via (4.2), usually one finds u_h as in (4.3) by solving the equation $(\partial/\partial z_i) J(\sum_k z_k \psi_k) = 0$.

Galerkin Method: The weak equation (4.2) is solved for problems where the bilinear form is not necessarily symmetric. If the weak equations arise from a variational problem with a positive quadratic form, then often the term *Ritz–Galerkin Method* is used.

Petrov–Galerkin Method: Here we seek $u_h \in S_h$ with

$$
a(u_h, v) = \langle \ell, v \rangle \quad \text{for all } v \in T_h,
$$

where the two N-dimensional spaces S_h and T_h need not be the same. The choice of a space of test functions which is different from S_h is particularly useful for problems with singularities.

As we saw in §§2 and 3, the boundary conditions determine whether a problem should be formulated in $H^m(\Omega)$ or in $H_0^m(\Omega)$. For the purposes of a unified notation, in the following we always suppose $V \subset H^m(\Omega)$, and that the bilinear form a is always V-elliptic, i.e.,

$$
a(v, v) \geq \alpha \|v\|_m^2 \quad \text{and} \quad |a(u, v)| \leq C \|u\|_m \|v\|_m \quad \text{for all } u, v \in V,
$$

where $0 < \alpha \leq C$. The norm $\| \cdot \|_m$ is thus equivalent to the energy norm (2.14), which we use to get our first error bounds. – In addition, let $\ell \in V'$ with $|\langle \ell, v \rangle| \leq \|\ell\| \cdot \|v\|_m$ for $v \in V$. Here $\|\ell\|$ is the (dual) norm of ℓ.

4.1 Remark. (Stability) Independent of the choice of the subspace S_h of V, the solution of (4.2) always satisfies

$$\|u_h\|_m \leq \alpha^{-1} \|\ell\|.$$

Proof. Let u_h be a solution of (4.2). Substituting $v = u_h$, we get

$$\alpha \|u_h\|_m^2 \leq a(u_h, u_h) = \langle \ell, u_h \rangle \leq \|\ell\| \, \|u_h\|_m.$$

Dividing by $\|u_h\|_m$, we get the assertion. $\qquad\Box$

The following lemma is of fundamental importance in establishing error bounds for finite element approximations. The line of proof is typical, and we will make frequent use of variants of the technique. The relation (4.7) below is often denoted as *Galerkin orthogonality*.

4.2 Céa's Lemma. *Suppose the bilinear form a is V-elliptic with $H_0^m(\Omega) \subset V \subset H^m(\Omega)$. In addition, suppose u and u_h are the solutions of the variational problem in V and $S_h \subset V$, respectively. Then*

$$\|u - u_h\|_m \leq \frac{C}{\alpha} \inf_{v_h \in S_h} \|u - v_h\|_m. \tag{4.6}$$

Proof. By the definition of u and u_h,

$$a(u, v) = \langle \ell, v \rangle \quad \text{for all } v \in V,$$
$$a(u_h, v) = \langle \ell, v \rangle \quad \text{for all } v \in S_h.$$

Since $S_h \subset V$, it follows by subtraction that

$$a(u - u_h, v) = 0 \quad \text{for all } v \in S_h. \tag{4.7}$$

Let $v_h \in S_h$. With $v = v_h - u_h \in S_h$, it now follows immediately from (4.7) that $a(u - u_h, v_h - u_h) = 0$, and

$$
\begin{aligned}
\alpha \|u - u_h\|_m^2 &\leq a(u - u_h, u - u_h) \\
&= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\
&\leq C \|u - u_h\|_m \|u - v_h\|_m.
\end{aligned}
$$

After dividing by $\|u - u_h\|_m$, we get $\alpha \|u - u_h\|_m \leq C \|u - v_h\|_m$, and the assertion is established. $\qquad\Box$
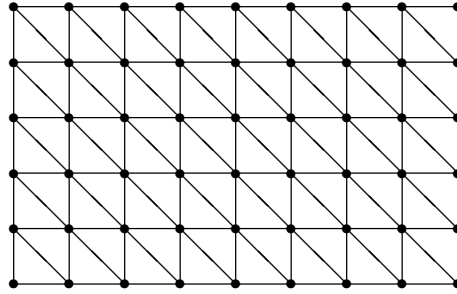
Fig. 9. A uniform triangulation of a rectangle

According to Céa's lemma, the accuracy of a numerical solution depends essentially on choosing function spaces which are capable of approximating the solution u well. For polynomials, the order of approximation is determined by the smoothness of the solution. However, for boundary-value problems, the smoothness of the solution typically decreases as we approach the boundary. Thus, it doesn't make much sense to use polynomials that are defined on the whole domain and to insist on a high accuracy by forcing the degree of the polynomials to be high. As we shall see in §§6 and 7, it makes more sense to use piecewise polynomials, and to achieve the desired accuracy by making the associated partition of Ω sufficiently fine. The so-called *h-p-methods* combine refinements of the partitions and an increase of the degree of the polynomials; see Schwab [1998].

Model Problem

4.3 Example (Courant [1943]). Suppose we want to solve the Poisson equation in the unit square (or in a general domain which can be triangulated with congruent triangles):

$$-\Delta u = f \quad \text{in } \Omega = (0,1)^2,$$
$$u = 0 \quad \text{on } \partial\Omega.$$

Suppose we partition $\bar{\Omega}$ with a uniform triangulation of mesh size h, as shown in Fig. 9. Choose

$$S_h := \{v \in C(\bar{\Omega}); v \text{ is linear in every triangle and } v = 0 \text{ on } \partial\Omega\}. \qquad (4.8)$$

In every triangle, $v \in S_h$ has the form $v(x,y) = a + bx + cy$, and is uniquely defined by its values at the three vertices of the triangle. Thus, $\dim S_h = N =$ number of interior mesh points. Globally, v is determined by its values at the N grid points (x_j, y_j). Now choose a basis $\{\psi_i\}_{i=1}^N$ with

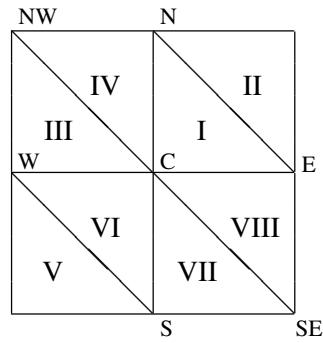$$\psi_i(x_j, y_j) = \delta_{ij}.$$

Fig. 10. Numbering of the elements in a neighborhood of the center C and the neighboring points in the compass directions: E, S, W, N, NW and SE

Table 1. Derivatives of the basis functions ψ_C shown in Fig. 10 (ψ_C has the value 1 at C and is 0 at other nodes)

	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

We compute the elements of the system matrix A_{ij}, where again we choose local indices and exploit the symmetry, (see Fig. 11 and Table 1)

$$
\begin{aligned}
a(\psi_C, \psi_C) &= \int_{I-VIII} (\nabla \psi_C)^2 dxdy \\
&= 2 \int_{I+III+IV} [(\partial_1 \psi_C)^2 + (\partial_2 \psi_C)^2] dxdy \\
&= 2 \int_{I+III} (\partial_1 \psi_C)^2 dxdy + 2 \int_{I+IV} (\partial_2 \psi_C)^2 dxdy \\
&= 2h^{-2} \int_{I+III} dxdy + 2h^{-2} \int_{I+IV} dxdy \\
&= 4,
\end{aligned}
$$

$$
\begin{aligned}
a(\psi_C, \psi_N) &= \int_{I+IV} \nabla \psi_C \cdot \nabla \psi_N dxdy \\
&= \int_{I+IV} \partial_2 \psi_C \partial_2 \psi_N dxdy = \int_{I+IV} (-h^{-1}) h^{-1} dxdy \\
&= -1.
\end{aligned}
$$

Here ψ_N is the nodal function for the point north of C. By symmetry, a similar computation gives

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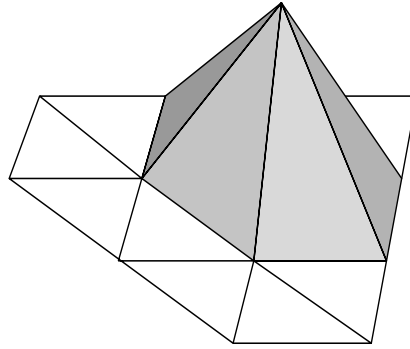$$

Fig. 11. Nodal basis function

Finally, we find that

$$a(\psi_C, \psi_{NW}) = \int_{III+IV} [\partial_1 \psi_C \partial_1 \psi_{NW} + \partial_2 \psi_C \partial_2 \psi_{NW}] dx dy = 0.$$

In evaluating $a(\psi_C, \psi_{SE})$, note that all products in the integrals vanish. Thus we get a system of linear equations with exactly the same matrix as in the finite difference method based on the standard five-point stencil

$$\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}_* . \tag{4.9}$$

We should emphasize that this connection with difference methods does not hold in general. The finite element method provides the user with a great deal of freedom, and for most other finite element approximations and other equations, there is no equivalent finite difference star. In general, the finite element approximation does not even satisfy the discrete maximum principle. – The same holds, by the way, for the method of finite volumes. Once again, we get the same matrix only in the above simple case [Hackbusch 1989].

The stiffness matrix for the model problem was determined here in a *node-oriented* way. We note that the matrices are assembled in a different way in real-life computations, i.e. *element-oriented*. First, the contribution of each triangle (element) to the stiffness matrix is determined by doing the computation only for a master triangle (reference element). Finally the contributions of all triangles are added.

Problems

4.4 As usual, let u and u_h be the functions which minimize J over V and S_h, respectively. Show that u_h is also a solution of the minimum problem

$$a(u - v, u - v) \longrightarrow \min_{v \in S_h} !$$

Because of this, the mapping

$$R_h : V \longrightarrow S_h$$
$$u \longmapsto u_h$$

is called the *Ritz projector*.

4.5 Consider the potential equation with inhomogeneous boundary conditions

$$-\Delta u = 0 \quad \text{in } \Omega = (0, 1)^2,$$
$$u = u_0 \quad \text{on } \partial\Omega,$$

and suppose we select the same regular triangulation as in Example 4.3. In addition, let u_0 be piecewise linear on the boundary. Then u_0 can be extended continuously to Ω so that u_0 is linear in every triangle and vanishes at the interior nodes. Show that in this situation, i.e. for inhomogeneous boundary conditions and with S_h as in (4.8), we get the same linear system as in Chapter I.

4.6 Suppose in Example 4.3 that on the bottom side of the square we replace the Dirichlet boundary condition by the natural boundary condition $\partial u/\partial v = 0$. Verify that this leads to the stencil

$$\begin{bmatrix} & -1 & \\ -1/2 & 2 & -1/2 \end{bmatrix}_*$$

at these boundary points.

4.7 In Example 4.3, does the part $-u_{xx}$ of the Laplace operator $-\Delta$ lead to a stencil which has only nonzero terms in one horizontal line, as is the case for the finite difference method?

4.8 Given the variational problem

$$\int_\Omega [a_1(\partial_1 v)^2 + a_2(\partial_2 v)^2 + a_3(\partial_1 v - \partial_2 v)^2 - 2fv]\,dx \longrightarrow \min!$$

with $a_1, a_2, a_3 > 0$, find the associated Euler differential equation and the difference star, using the same form of approximating function as in Example 4.3.

4.9 Consider the boundary-value problem (2.13) and apply the Galerkin method with polynomials of degree k on $[0, 1]$. Show convergence for $k \to \infty$.

§ 5. Some Standard Finite Elements

In practice, the spaces over which we solve the variational problems associated with boundary-value problems are called *finite element spaces*. We partition the given domain Ω into (finitely many) subdomains, and consider functions which reduce to a polynomial on each subdomain. The subdomains are called *elements*. For planar problems, they can be triangles or quadrilaterals. For three-dimensional problems, we can use tetrahedra, cubes, rectangular parallelepipeds, etc. For simplicity, we restrict our discussion primarily to the two-dimensional case.

Here is a list of some of the important properties characterizing different finite element spaces:

1. The kind of partition used on the domain: triangles or quadrilaterals. If all elements are congruent, we say that the partition is *regular*.

2. In two variables, we refer to

$$\mathcal{P}_t := \{u(x, y) = \sum_{\substack{i + k \leq t \\ i, k \geq 0}} c_{ik} x^i y^k\} \tag{5.1}$$

as the set of *polynomials of degree $\leq t$*. If all polynomials of degree $\leq t$ are used, we call them finite elements with *complete polynomials*.

The restrictions of the polynomials to the edges of the triangles or quadrilaterals are polynomials in one variable. Sometimes we will require that their degree be smaller than t (e.g., at most $t - 1$). Such a condition will be part of the specification of the elements.

The admissible polynomial degrees in the elements or on their edges are a local property.

3. Continuity and differentiability properties: A finite element is said to be a C^k *element* provided it is contained in $C^k(\Omega)$.[5] This property is of a global character and is often concealed in interpolation conditions.

We remark that according to this scheme, the Courant triangles in Example 4.3 would be classified as linear triangular elements in $C^0(\Omega)$.

We use the terminology *conforming finite element* if the functions lie in the Sobolev space in which the variational problem is posed. Nonconforming elements will be studied in Chapter III.

---

[5] The use of the terminology *element* may be somewhat confusing. We decompose the domain into *elements* which are geometric objects, while the *finite elements* are actually functions. However, we will deviate from this convention when discussing, e.g., C^k elements or linear elements, where the meaning is clear from the context.

The formal definition of finite elements will be given in the definitions 5.8 and 5.12. The reader will recognize the three properties above in the triple (T, Π, Σ) after the significance of the differentiability conditions will be clear.

Requirements on the Meshes

For simplicity, in the following let Ω be a polygonal domain which can be partitioned into triangles or quadrilaterals. The partition is by no means required to be as regular as the one shown in the model problem in §4.

5.1 Definition. (1) A partition $\mathcal{T} = \{T_1, T_2, \ldots, T_M\}$ of Ω into triangular or quadrilateral elements is called *admissible* provided the following properties hold (see Fig. 12):

  i. $\bar{\Omega} = \bigcup_{i=1}^{M} T_i$.
 ii. If $T_i \cap T_j$ consists of exactly one point, then it is a common vertex of T_i and T_j.
iii. If for $i \neq j$, $T_i \cap T_j$ consists of more than one point, then $T_i \cap T_j$ is a common edge of T_i and T_j.

(2) We will write $\mathcal{T}_h$ instead of $\mathcal{T}$ when every element has diameter at most $2h$.

(3) A family of partitions $\{\mathcal{T}_h\}$ is called *shape regular* provided that there exists a number $\kappa > 0$ such that every T in $\mathcal{T}_h$ contains a circle of radius ρ_T with

$$\rho_T \geq h_T / \kappa,$$

where h_T is half the diameter of T.

(4) A family of partitions $\{\mathcal{T}_h\}$ is called *uniform* provided that there exists a number $\kappa > 0$ such that every element T in $\mathcal{T}_h$ contains a circle with radius $\rho_T \geq h / \kappa$. We will often use the terminology κ-*regular*.
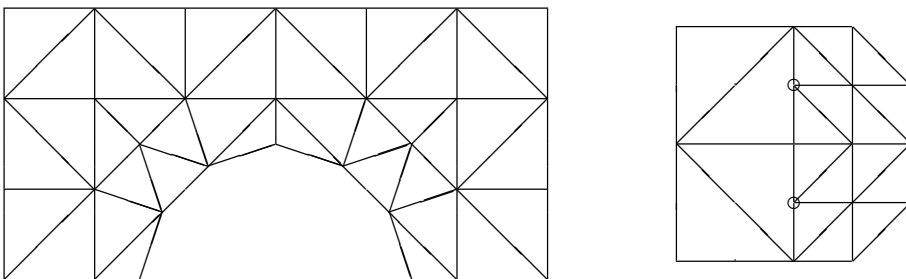

Fig. 12. An admissible triangulation (left), and one which is not because of two hanging nodes marked by ∘ (right).
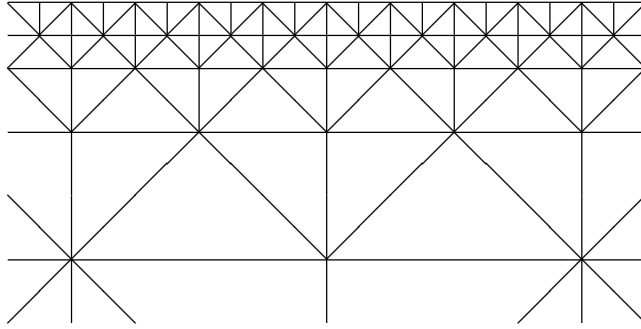
Fig. 13. A triangulation which is shape regular but not uniform

Since $h = \max_{T \in \mathcal{T}} h_T$, uniformity is a stronger requirement than shape regularity. Clearly, the triangulations shown in Figs. 13 and 14 are shape regular, independent of how many steps of the refinement in the neighborhood of the boundary or of a reentrant corner are carried out. However, if the number of steps depends on h, the partitions are no longer uniform.

In practice, we almost always use shape-regular meshes, and very frequently even uniform ones.

Significance of the Differentiability Properties

In the conforming treatment of second order elliptic problems, we choose finite elements which lie in H^1. We shall show that it is possible to use functions which are continuous but not necessarily continuously differentiable. Thus, the functions are much less smooth than required for a classical solution of the boundary-value problem.

In the following, we will always assume unless otherwise indicated that the partitions satisfy the requirements of 5.1. We say that a function u on Ω satisfies a given property *piecewise* provided that its restriction to every element has that property.

5.2 Theorem. *Let $k \geq 1$ and suppose Ω is bounded. Then a piecewise infinitely differentiable function $v : \bar{\Omega} \to \mathbb{R}$ belongs to $H^k(\Omega)$ if and only if $v \in C^{k-1}(\bar{\Omega})$.*

Proof. It suffices to give the proof for $k = 1$. For $k > 1$ the assertion then follows immediately from a consideration of the derivatives of order $k - 1$. In addition, for simplicity we restrict ourselves to domains in $\mathbb{R}^2$.

(1) Let $v \in C(\bar{\Omega})$, and suppose $\mathcal{T} = \{T_j\}_{j=1}^M$ is a partition of Ω. For $i = 1, 2$, define $w_i : \Omega \to \mathbb{R}$ piecewise by $w_i(x) := \partial_i v(x)$ for $x \in \Omega$, where on the edges we can take either of the two limiting values. Let $\phi \in C_0^\infty(\Omega)$. Green's formula
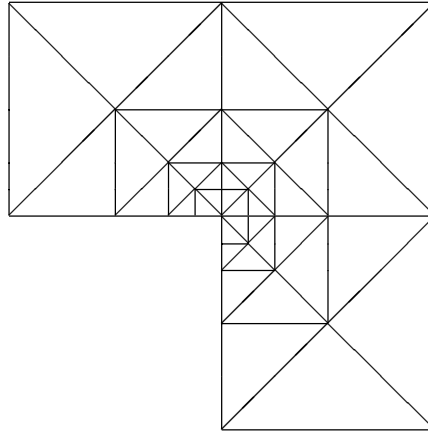
Fig. 14. Nonuniform triangulations with a reentrant vertex

can be applied in every element T_j to give

$$\int_\Omega \phi w_i \, dxdy = \sum_j \int_{T_j} \phi \partial_i v \, dxdy$$

$$= \sum_j \left\{ -\int_{T_j} \partial_i \phi v \, dxdy + \int_{\partial T_j} \phi v \, \nu_i \, ds \right\}. \qquad (5.2)$$

Since v was assumed to be continuous, the integrals over the interior edges cancel. Moreover, ϕ vanishes on $\partial\Omega$, and we are left with the integral over the domain

$$-\int_\Omega \partial_i \phi v \, dxdy.$$

By Definition 1.1, w_i is the weak derivative of v.

(2) Let $v \in H^1(\Omega)$. We do not establish the continuity of v by working backwards through the formulas (although this would be possible), but instead employ an approximation-theoretical argument. Consider v in the neighborhood of an edge, and rotate the edge so that it lies on the y-axis. Suppose the edge becomes the interval $[y_1 - \delta, y_2 + \delta]$ on the y-axis with $y_1 < y_2$ and $\delta > 0$. We now investigate the auxiliary function

$$\psi(x) := \int_{y_1}^{y_2} v(x, y) dy.$$

First, suppose $v \in C^\infty(\Omega)$. It now follows from the Cauchy–Schwarz inequality that

$$|\psi(x_2) - \psi(x_1)|^2 = \left| \int_{x_1}^{x_2} \int_{y_1}^{y_2} \partial_1 v \, dxdy \right|^2$$

$$\leq \left| \int_{x_1}^{x_2} \int_{y_1}^{y_2} 1 \, dxdy \right| \cdot |v|_{1,\Omega}^2$$

$$\leq |x_2 - x_1| \cdot |y_2 - y_1| \cdot |v|_{1,\Omega}^2.$$

Because of the density of $C^\infty(\Omega)$ in $H^1(\Omega)$, this assertion also holds for $v \in H^1(\Omega)$. Thus the function $x \mapsto \psi(x)$ is continuous, and in particular at $x = 0$. Since y_1 and y_2 are arbitrary except for $y_1 < y_2$, this can only happen if the piecewise continuous function v is continuous on the edge. $\qquad\square$

If no other additional conditions are required, continuous finite elements are easily constructed. In view of Theorem 5.2, this is of great advantage for the solution of second order boundary-value problems using conforming finite elements. The construction of C^1 elements, which according to Theorem 5.2 are required for the conforming treatment of problems of fourth order, is more difficult.
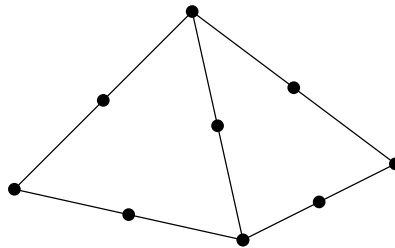

Fig. 15. Piecewise quadratic polynomials that interpolate at the points ($\bullet$) are continuous at the interface

Triangular Elements with Complete Polynomials

The simplest triangular elements to construct are C^0 elements made up of complete polynomials.

5.3 Remark. Let u be a polynomial of degree t. If we apply an affine linear transformation and express u in the new coordinates, we again get a polynomial of degree t. Thus, the set of polynomials $\mathcal{P}_t$ is invariant under affine linear transformations.

5.4 Remark. Let $t \geq 0$. Given a triangle T, suppose $z_1, z_2, \ldots, z_s$ are the $s = 1 + 2 + \cdots + (t+1)$ points in T which lie on $t + 1$ lines, as in Fig. 16. Then for every $f \in C(T)$, there is a unique polynomial p of degree $\leq t$ satisfying the interpolation conditions

$$p(z_i) = f(z_i), \quad i = 1, 2, \ldots, s. \tag{5.3}$$

Proof. The result is trivial for $t = 0$. We now assume it has been established for $t-1$, and prove it for t. In view of the invariance under affine transformations, we can assume that one of the edges of T lies on the x-axis. Suppose it is the one containing the points $z_1, z_2, \ldots, z_{t+1}$. There exists a univariate polynomial $p_0 = p_0(x)$ with

$$p_0(z_i) = f(z_i), \quad i = 1, 2 \ldots, t+1.$$

By the induction hypothesis, there also exists a polynomial $q = q(x, y)$ of degree $t - 1$ with

$$q(z_i) = \frac{1}{y_i}[f(z_i) - p_0(z_i)], \quad i = t + 2, \ldots, s.$$

Clearly, $p(x, y) = p_0(x) + yq(x, y)$ satisfies (5.3). □

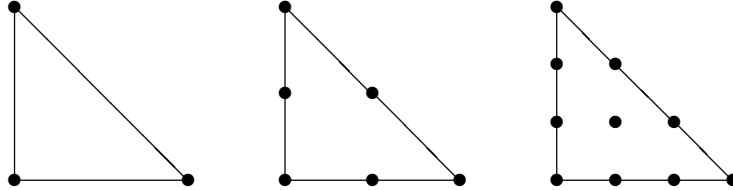**Fig. 16.** Nodes of the nodal basis for linear, quadratic, and cubic triangular elements $\mathcal{M}_0^1$, $\mathcal{M}_0^2$, and $\mathcal{M}_0^3$

5.5 Definition. Suppose that for a given finite element space, there is a set of points which uniquely determines any function in the space by its values at the points. Then the set of functions in the space which take on a nonzero value at precisely one of the points form a basis for the space, called the *nodal basis*.

The following construction, which assures continuity by using sufficiently many points on the edges of the triangle, is typical for the construction of C^0 elements.

5.6 A Nodal Basis for C^0 Elements. Let $t \geq 1$, and suppose we are given a triangulation of Ω. In each triangle, we place $s := (t + 1)(t + 2)/2$ points as indicated in Fig. 16, so that there are $t + 1$ points on each edge. By Remark 5.4, in each triangle a polynomial of degree $\leq t$ is determined by choosing values at these points. The restriction of any such polynomial to an edge is a polynomial of degree $\leq t$ in one variable. Now given an edge, the two polynomials on either side interpolate the same values at the $t + 1$ points on that edge, and thus must reduce to the same one-dimensional polynomial. This ensures that our elements are globally continuous.
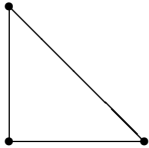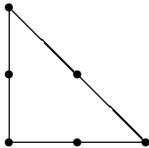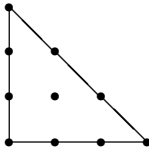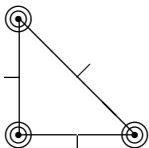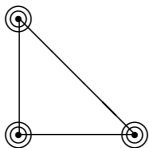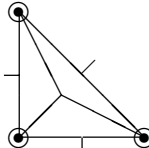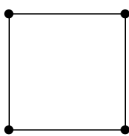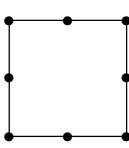
In dealing with finite elements with complete polynomials, we make use of the following notation from the literature:

$$\begin{aligned}
\mathcal{M}^k &:= \mathcal{M}_k(\mathcal{T}) := \{v \in L_2(\Omega); \ v|_T \in \mathcal{P}_k \text{ for every } T \in \mathcal{T}\}, \\
\mathcal{M}_0^k &:= \mathcal{M}^k \cap C^0(\Omega) = \mathcal{M}^k \cap H^1(\Omega), \\
\mathcal{M}_{0,0}^k &:= \mathcal{M}^k \cap H_0^1(\Omega).
\end{aligned} \tag{5.4}$$

$\mathcal{M}_0^1$ is also called the *conforming P_1 element* or *Courant triangle*.

Table 2. Interpolation with some standard finite elements

- $\bullet$ Function value prescribed
- $\odot$ Function value and 1st derivative prescribed
- $\circledcirc$ Function value and 1st and 2nd derivatives prescribed
- $\perp$ Normal derivative prescribed

Linear triangular element $\mathcal{M}_0^1$
$u \in C^0(\Omega)$
$\Pi_{\text{ref}} = \mathcal{P}_1, \quad \dim \Pi_{\text{ref}} = 3$

Quadratic triangular element $\mathcal{M}_0^2$
$u \in C^0(\Omega)$
$\Pi_{\text{ref}} = \mathcal{P}_2, \quad \dim \Pi_{\text{ref}} = 6$

Cubic triangular element $\mathcal{M}_0^3$
$u \in C^0(\Omega)$
$\Pi_{\text{ref}} = \mathcal{P}_3, \quad \dim \Pi_{\text{ref}} = 10$

Argyris triangle
$u \in C^1(\Omega)$
$\Pi_{\text{ref}} = \mathcal{P}_5, \quad \dim \Pi_{\text{ref}} = 21$

Bell triangle
$u \in C^1(\Omega)$
$\Pi_{\text{ref}} \subset \mathcal{P}_5, \ \partial_\nu u|_{\partial T_i} \in \mathcal{P}_3, \quad \dim \Pi_{\text{ref}} = 18$

Hsieh–Clough–Tocher element
$u \in C^1(\Omega)$
$T = \bigcup_{i=1}^3 K_i, \quad u|_{K_i} \in \mathcal{P}_3, \quad \dim \Pi_{\text{ref}} = 12$

Bilinear quadrilateral element Q_1
$u \in C^0(\Omega)$
$\Pi_{\text{ref}} \subset \mathcal{P}_2, \ u|_{\partial T_i} \in \mathcal{P}_1, \quad \dim \Pi_{\text{ref}} = 4$

Serendipity element
$u \in C^0(\Omega)$
$\Pi_{\text{ref}} \subset \mathcal{P}_3, \ u|_{\partial T_i} \in \mathcal{P}_2, \quad \dim \Pi_{\text{ref}} = 8$

Remarks on C^1 Elements

The construction of C^1 elements is considerably more difficult. There are two well-known constructions of triangular elements based on polynomials of degree 5. Recall that dim $\mathcal{P}_5 = 21$. We now assume that we are given values for derivatives up to order 2 at each of the vertices of the triangle. This uses $3 \times 6 = 18$ degrees of freedom.

To construct the *Argyris element*, we use the remaining three degrees of freedom to specify the values of the normal derivatives at the midpoint of each side of the triangle. The following argument shows that this leads to a global C^1 function.

Consider the univariate polynomials which are the restrictions of two neighboring polynomials to a common edge of the triangle. At the ends of the edge, these two polynomials must both interpolate the values of the given derivatives up to order 2. Since this interpolation problem has a unique solution, we get the desired continuity of the function and its tangential derivative. The normal derivatives along the edge are polynomials of degree 4. They both interpolate the given derivatives up to order 1 at the ends of the edge, along with the given value at the center of the edge. Since these five pieces of data uniquely determine a polynomial of degree 4, we have established the continuity of the normal derivative.

To construct the *triangular element of Bell*, we use the same data at the vertices as for the Argyris element. But now we restrict ourselves to the class of polynomials of degree 5 whose normal derivatives on the sides of the triangle are polynomials of degree 3 rather than 4. Again the normal derivative along an edge is uniquely determined by the derivative information at the vertices, and we get a continuous derivative. The number of degrees of freedom for this element is 3 less than for the Argyris element (see Table 2).

The *Hsieh–Clough–Tocher element* is constructed by a completely different process. First we subdivide the triangle T into three subtriangles by connecting its vertices to its center of gravity. We now build a C^1 function consisting piecewise of cubic polynomials. At each of the vertices of the original triangle, we specify the function value and the first derivatives. In addition, we specify the normal derivative at the midpoint of each of the three sides of T. It can be shown that the three cubic polynomials join together to form a C^1 function on T. The fact that two adjoining macro-elements join with C^1 continuity can be established in the same way as for the other C^1 elements. This element has exactly 12 degrees of freedom.

The *reduced Hsieh–Clough–Tocher element* is constructed in a similar way, except that now we insist that the normal derivatives along the edges of T be linear rather than quadratic. The analysis now proceeds as in our construction of the Bell element from the Argyris element (cf. Problem 6.15).

Because it involves a subpartition, the Hsieh–Clough–Tocher element is called

a *macro-element*. Another macro-element is the Powell–Sabin element; see Powell and Sabin [1977].

It should be noted that the continuity of derivatives along the element boundaries is easy to handle in terms of the Bernstein–Bézier representation of polynomials.

Bilinear Elements

The polynomial families $\mathcal{P}_t$ are not used on rectangular partitions of a domain. We can see why by looking at the simplest example, the bilinear element. Instead of using $\mathcal{P}_t$ as we did for triangles, on rectangular elements we use the polynomial family which contains *tensor products*:

$$\mathcal{Q}_t := \{u(x, y) = \sum_{0 \le i,k \le t} c_{ik} x^i y^k\}. \tag{5.5}$$

If more general quadrilateral elements are involved, we can use appropriately transformed families.

We consider first a rectangular grid whose grid lines run parallel to the coordinate axes. On each rectangle we use

$$u(x, y) = a + bx + cy + dxy, \tag{5.6}$$

where the four parameters are uniquely determined by the values of u at the four vertices of the rectangle. Although u is a polynomial of degree 2, its restriction to each edge is a linear function. Because of this, we automatically get global continuity of the elements since neighboring bilinear pieces share the same node information.

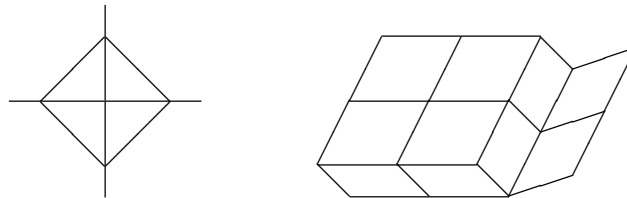**Fig. 17.** A rectangle rotated by $45°$, and a parallelogram element

The polynomial form (5.6) is not usable on a grid which has been rotated by $45°$ as in Fig. 17. Indeed, the term dxy in (5.6) vanishes at all of the vertices of the rotated square.

We can get the correct polynomial form for general parallelograms (and thus for the rotated elements shown in Fig. 17) by means of a linear transformation.

However, the treatment of general quadrilaterals requires the more general class of so-called *isoparametric* mappings, which are discussed in Chapter III. – It is possible to combine parallelograms with triangles, in order to make the partition of Ω more flexible.

Suppose the edges of a parallelogram element lie on lines of the form

$$
\begin{aligned}
\alpha_1 x + \beta_1 y &= \gamma_1, \\
\alpha_2 x + \beta_2 y &= \gamma_2
\end{aligned}
\tag{5.7}
$$

(where the coefficients vary from element to element). Then with the transformation

$$
\begin{aligned}
\xi &= \alpha_1 x + \beta_1 y, \\
\eta &= \alpha_2 x + \beta_2 y,
\end{aligned}
$$

we get the *bilinear* function

$$
u(x, y) = a + b\xi + c\eta + d\xi\eta
\tag{5.8}
$$

which is linear along the edges of the parallelogram.

5.7 Remarks. (1) We can also characterize the Q_1 elements obtained by the above construction in a coordinate-free way:

$$
S = \{v \in C^0(\bar{\Omega}); \text{ for every element } T, v|_T \in \mathcal{P}_2,
$$
$$
\text{and the restriction to each edge belongs to } \mathcal{P}_1\}.
$$

(2) If every edge of an element contains $t_k + 1$ nodes of a nodal basis, and the restrictions to the edges are all polynomials of degree t_k at most, then we automatically get globally continuous elements. – Note that the maximum degree of a polynomial restricted to an edge does not increase under a linear transformation.

Quadratic Rectangular Elements

One of the most popular elements on rectangles (or on more general parallelograms) consists of piecewise polynomials of degree 3 whose restrictions to the edges are quadratic polynomials. Using the coordinates shown in Fig. 18, we can write such a polynomial in the form

$$
\begin{aligned}
u(x, y) = {}& a + bx + cy + dxy \\
& + e(x^2 - 1)(y - 1) + f(x^2 - 1)(y + 1) \\
& + g(x - 1)(y^2 - 1) + h(x + 1)(y^2 - 1)
\end{aligned}
$$

(cf. Problem 5.16). There are eight degrees of freedom. The first four are determined by the values at the vertices. The remaining parameters e, f, g and h can be computed directly from the values at the midpoints of the sides. This element is called the *eight node element* or the *serendipity element*. If we add the term

$$k(x^2 - 1)(y^2 - 1),$$

we get one more degree of freedom, and can then interpolate a value at the center of the rectangle. By dropping some degrees of freedom, we can also get useful six node elements (with $e = f = 0$ or $g = h = 0$, respectively), as shown in Fig. 18.
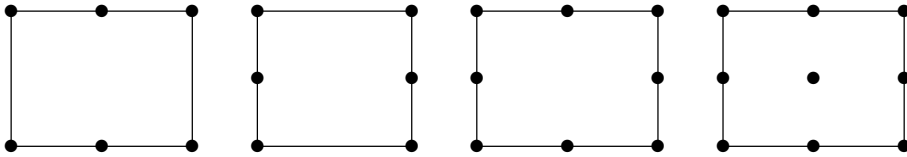

Fig. 18. Rectangular elements with 6, 8, or 9 nodes for a rectangle with edges on the lines $|x| = 1$, $|y| = 1$.

Affine Families

In the above discussion of special finite element spaces, we have implicitly made use of the following formal construction; cf. Ciarlet [1978].

5.8 Definition. A *finite element* is a triple (T, Π, Σ) with the following properties:

 (i) T is a polyhedron in $\mathbb{R}^d$. (The parts of the surface ∂T lie on hyperplanes and are called *faces*.)

 (ii) Π is a subspace of $C(T)$ with finite dimension s. (Functions in Π are called *shape functions* if they form a basis of Π.)

(iii) Σ is a set of s linearly independent functionals on Π. Every $p \in \Pi$ is uniquely defined by the values of the s functionals in Σ. – Since usually the functionals involve point evaluation of a function or its derivatives at points in T, we call these *(generalized) interpolation conditions*.

   In (ii) s is the *number of local degrees of freedom* or *local dimension*.

   Although generally Π consists of polynomials, it is not enough to look only at polynomial spaces, since otherwise we would exclude piecewise polynomial elements such as the Hsieh–Clough–Tocher element. In fact, there are even finite elements consisting of piecewise rational functions; see Wachspress [1971].

   As a first example consider the finite element families $\mathcal{M}_0^k$. We have

$$\mathcal{M}_0^k = (T, \mathcal{P}^k, \Sigma^k),$$

$$\Sigma^k := \{p(z_i); \ i = 1, 2, \ldots, \frac{(k+1)(k+2)}{2}\},$$

where the points z_i are defined in Remark 5.4 and depicted in Fig. 16 for $k \leq 3$.

The condition for a smooth join between elements is dealt with in (iii), although in fact, we actually need a still stronger formalization of this condition. However, for the C^1 elements presented in Table 2, the meaning is clear. Thus, e.g., for the Argyris triangle, by Table 2 we have

$$\Pi := \mathcal{P}_5, \qquad \dim \Pi = 21,$$
$$\Sigma := \{p(a_i), \partial_x p(a_i), \partial_y p(a_i), \partial_{xx} p(a_i), \partial_{xy} p(a_i), \partial_{yy} p(a_i), \ i = 1, 2, 3,$$
$$\partial_n p(a_{12}), \partial_n p(a_{13}), \partial_n p(a_{23})\},$$

where a_1, a_2, a_3 are the vertices and $a_{ij} = \frac{1}{2}(a_i + a_j)$ are the midpoints of the sides.

Another example elucidates the role of the functionals in Σ. Fig. 19 shows three different finite elements with $\Pi = \mathcal{P}_1$. Only the first one belongs to $H^1(\Omega)$. Although the local degree of freedom is 3 in each case, the dimensions of the resulting finite element spaces are quite different; cf. Problem 5.13. Similarly, $\mathcal{M}_0^3$ and the cubic Hermite triangle are different elements with cubic polynomials shown in Fig. 20.
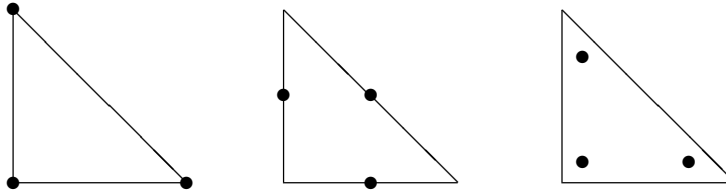

Fig. 19. The P_1 elements $\mathcal{M}_0^1$, $\mathcal{M}_*^1$, and $\mathcal{M}^1$. Here and in the other diagrams the points marked by a • refer to point evaluations from the set of functionals Σ associated to the finite element spaces in Definition 5.8(iii). The symbols for other functionals are found in Table 2 and Fig. 21.

5.9 The Cubic Hermite Triangle. The ten degrees of freedom of cubic polynomials can also be fixed in another way. In particular, we can choose the values of the polynomial and its first derivatives at the vertices a_i, $i = 1, 2, 3$, along with the value at the center $a_{123} = \frac{1}{3}(a_1 + a_2 + a_3)$. The cubic Hermite triangle is the triple $(T, \mathcal{P}_3, \Sigma_{HT})$ where

$$\Sigma_{HT} := \left\{ p(a_i), \frac{\partial}{\partial x} p(a_i), \frac{\partial}{\partial y} p(a_i), \ i = 1, 2, 3, \ \text{and} \ p(a_{123}) \right\}. \qquad (5.9)$$

The functions in Σ_{HT} are linearly independent. To verify this, we consider an edge of the triangle between the vertices a_i and a_j $(i \neq j)$. Let $q \in \mathcal{P}_3$ be the univariate polynomial which is the restriction of p to the edge. The values $q(a_i)$, $q(a_j)$ and the derivatives at these points are given by p and the directional derivative.

Since the one-dimensional Hermite interpolation problem for two points and cubic polynomials has a unique solution, we can compute q. Hence, the values at the ten nodes shown in Fig. 16 for the Lagrange interpolation are uniquely determined. Thus we have reduced the interpolation problem for the Hermite triangle to the usual Lagrange interpolation problem which is known to be solvable.

We emphasize that the derivatives are continuously joined only at the vertices [but not along the edges]. The cubic Hermite triangle is *not* a C^1 element. Nevertheless, we will see in Chapter VI that it provides appropriate nonconforming H^2 elements for the treatment of plates.
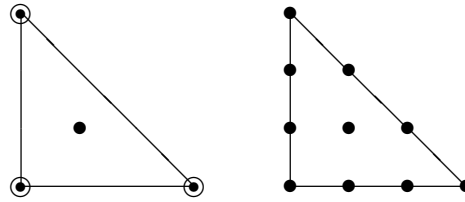

Fig. 20. Interpolation points for two elements with $\Pi = \mathcal{P}_3$, i.e. with piecewise cubic polynomials

5.10 The Bogner–Fox–Schmit rectangle. On the other hand there is a C^1-element with bicubic functions. It is called the Bogner–Fox–Schmit element and depicted in Fig. 21.

$$
\begin{aligned}
\Pi_{\text{ref}} &:= \mathcal{Q}_3, \qquad \dim \Pi_{\text{ref}} = 16, \\
\Sigma &:= \big\{ p(a_i),\, \partial_x p(a_i),\, \partial_y p(a_i),\, \partial_{xy} p(a_i),\ i = 1, 2, 3, 4. \big\}
\end{aligned}
\tag{5.10}
$$

Since the data in (5.10) refer to the tensor products of one-dimensional Hermite interpolation, the 16 functionals in Σ are linearly independent on Q_3.

To verify C^1 continuity of the Bogner–Fox–Schmit element, consider the univariate polynomial on a vertical edge of the rectangle. Its restriction to the edge is a cubic polynomial in y which is determined by p and $\partial_y p$ at the two vertices. Similarly the normal derivative $\partial_x p$ is also a cubic polynomial and determined by $\partial_x p$ and $\partial_{xy} p$ at the vertices. Thus we have continuity of p and $\partial_x p$, i.e. C^1 continuity. $\square$
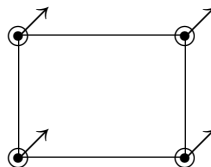

Fig. 21. C^1-element of Bogner–Fox–Schmit. The symbol $\nearrow$ refers to the mixed second derivative $\partial_{xy} p$.

5.11 Standard Elements. Strictly speaking the diagrams show the sets T and Σ from the triple (T, Π, Σ). Nevertheless, in many cases the associated family Π is considered clear and is sometimes just mentioned without a detailed specification. For example, diagrams as in Fig. 16 refer to $\Pi = \mathcal{M}_0^k$. If a point evaluation at the center of a triangle is added to $\mathcal{M}_0^1$ or $\mathcal{M}_0^2$, then the space is augmented by a bubble function as for the MINI element in Ch. III, §7 or the plate elements in Ch. VI, §6. Figures 17 and 18 show further standard elements with $\Pi = P_1$ and $\Pi = P_3$.

The functionals which are encountered with elements for scalar equations, are found in Table 2 and Fig. 21. The specification of vector valued elements often contains normal components or tangential components on the edges. Motivation is given in Problem 5.13, but applications are only contained in Chapter III and VI.

Definition 5.8 refers to a single element. The analysis of the finite element spaces can be obtained from results for a reference element, if all elements are constructed by affine transformations.

5.12 Definition. A family of finite element spaces S_h for partitions $\mathcal{T}_h$ of $\Omega \subset \mathbb{R}^d$ is called an *affine family* provided there exists a finite element $(T_{\text{ref}}, \Pi_{\text{ref}}, \Sigma)$ called the *reference element* with the following properties:

(iv) For every $T_j \in \mathcal{T}_h$, there exists an affine mapping $F_j : T_{\text{ref}} \longrightarrow T_j$ such that for every $v \in S_h$, its restriction to T_j has the form

$$v(x) = p(F_j^{-1} x) \quad \text{with } p \in \Pi_{\text{ref}}.$$

We have already encountered several examples of affine families. The families $\mathcal{M}_0^k$ and the rectangular elements considered so far are affine families. For example, $\mathcal{M}_0^k$ is defined by the triple $(\hat{T}, \mathcal{P}_k, \Sigma_k)$, where

$$\hat{T} := \{(\xi, \eta) \in \mathbb{R}^2;\ \xi \geq 0,\ \eta \geq 0,\ 1 - \xi - \eta \geq 0\} \tag{5.11}$$

is the unit triangle and $\Sigma_k := \{p(z_i);\ i = 1, 2, \ldots, s := k(k+1)/2\}$ is the set of nodal basis points z_i in Remark 5.6.

In our above discussion of the bilinear rectangular elements and the analogous biquadratic ones, it is clear how the transformation (iv) works (cf. (5.8)), but this is not the case for the complete polynomials on triangles. For rectangular elements, the unit square $[-1, +1]^2$ is the natural reference quadrilateral.

On the other hand, whenever conditions on the normal derivatives enter into the definition (e.g., in the definition of the Argyris triangle), then we do not have an affine family; see Fig. 22. This can be remedied by combining the normal derivatives with the tangential ones in the analysis. This has led to the theory of *almost-affine families*; see Ciarlet [1978].
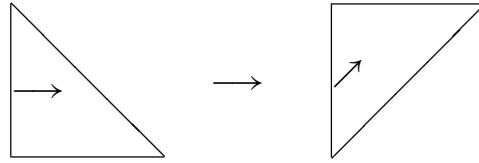
Fig. 22. Transformation of the unit triangle and one normal direction by the affine map $x \mapsto x$, $x + y \mapsto y$.

Choice of an Element

There are a large number of special elements which are useful for the treatment of systems of elliptic partial differential equations, see Chs. III and VI, or Ciarlet [1978], Bathe [1986]. It is useful to say something about how to choose an element even in the scalar case.

The choice of whether to use a triangular or rectangular partition depends primarily on the shape of the domain. Triangles are more flexible, but in solid mechanics, rectangular elements are generally preferred (cf. Ch. VI, §4).

For problems with a smooth behavior, we generally get better results using (bi-)quadratic elements than with (bi-)linear ones with the same number of free parameters. However, they do lead to linear systems with a larger bandwidth, and there is more work involved in setting up the stiffness matrices. This drawback is avoided when using standard finite element packages. Nevertheless, to save programming time and to get results as quickly as possible, linear elements are often used.

Problems

5.13 Consider the subset of all polynomials p in $\mathcal{P}_k$ for which
  a) the restriction of p to any edge lies in $\mathcal{P}_{k-1}$, or
  b) the restriction of the normal derivative $\partial_\nu p$ to any edge lies in $\mathcal{P}_{k-2}$.
Which of the two sets generates an affine family?

5.14 The completion of the space of vector-valued functions $C^\infty(\Omega)^n$ w.r.t. the norm

$$\|v\|^2 := \|v\|_{0,\Omega}^2 + \|\operatorname{div} v\|_{0,\Omega}^2$$

is denoted by $H(\operatorname{div}, \Omega)$. Obviously, $H^1(\Omega)^n \subset H(\operatorname{div}, \Omega) \subset L_2(\Omega)^n$. Show that a piecewise polynomial v is contained in $H(\operatorname{div}, \Omega)$ if and only if the components $v \cdot \nu$ in the direction of the normals are continuous on the inter-element boundaries. Hint: Apply Theorem 5.2 and use (2.22). — Similarly piecewise polynomials in the space $H(\operatorname{rot}, \Omega)$ are characterized by the continuity of the tangential components; see Problem VI.4.8.

5.15 Show that for a triangulation of a simply connected domain, *the number of triangles plus the number of nodes minus the number of edges* is always 1. Why doesn't this hold for multiply connected domains?

5.16 When considering the cubic Hermite triangle, there are three degrees of freedom per vertex and one per triangle. By the results of the previous problem, we know that the dimension must be smaller than for the standard Lagrange representation. Where are the missing degrees of freedom?

5.17 Let $f \in L_2(\Omega)$, and suppose u and u_h are the solutions of the Poisson equation $-\Delta u = f$ in $H_0^1(\Omega)$ and in a finite element space $S_h \subset C^0(\Omega)$, respectively. By construction ∇u and ∇u_h are at least L_2 functions. By the remark in Example 2.10, we know that the divergence of ∇u is an L_2 function. With the help of Problem 5.14, show that this no longer holds for the divergence of ∇u_h in general.

5.18 Show that the piecewise cubic continuous quadrilateral elements whose restrictions to the edges are quadratic polynomials, are exactly the serendipity class of eight node elements.
Hint: First consider a rectangle with sides parallel to the axes.

5.19 To construct triangular elements based on quadratic polynomials, consider the subspace of functions whose normal derivatives on the three edges are constant. Find the dimension of this space, distinguishing the cases when it is a right triangle or not.

5.20 The set of cubic polynomials whose restrictions to the edges of a triangle are quadratic form a 7-dimensional space. Give a basis for it on the unit triangle (5.9). — We will later encounter the cubic bubble function B_3. The result of this problem can be identified with $\mathcal{P}_2 \oplus B_3$.

§ 6. Approximation Properties

In this section we give error bounds for finite element approximations. By Céa's lemma, in the energy norm it suffices to know how well the solution can be approximated by elements in the corresponding finite element space $S_h = S_h(\mathcal{T}_h)$. For general methods, a suitable framework is provided by the theory of *affine families*. We do not derive results for every individual element, but instead examine a *reference element*, and use transformation formulas to carry the results over to shape-regular grids.

We intend to provide error bounds in other norms besides the energy norm.

We will concentrate primarily on affine families of triangular elements. Clearly, the error for an interpolation method provides an upper bound for the error of the *best* approximation. It turns out that we actually get the correct order of approximation in this way. – We consider C^0 elements, which according to Theorem 5.2, are not contained in $H^m(\Omega)$ for $m > 1$, and so the higher Sobolev norms are not applicable. As substitutes, we use certain *mesh-dependent norms* which are tailored to the problem at hand. We do not use the symbols $\| \cdot \|_h$ and $\| \cdot \|_{m,h}$ for fixed norms, but allow the norm to change from case to case. Often mesh-dependent norms are *broken norms* as in (6.1) or norms with weight factors h^{-m} as in Problem III.1.9.

6.1 Notation. Given a partition $\mathcal{T}_h = \{T_1, T_2, \ldots, T_M\}$ of Ω and $m \geq 1$, let

$$\|v\|_{m,h} := \sqrt{\sum_{T_j \in \mathcal{T}_h} \|v\|_{m,T_j}^2} \, . \tag{6.1}$$

Clearly, $\|v\|_{m,h} = \|v\|_{m,\Omega}$ for $v \in H^m(\Omega)$.

Let $m \geq 2$. By the Sobolev imbedding theorem (see Remark 3.4) $H^m(\Omega) \subset C^0(\Omega)$, i.e. every $v \in H^m$ has a continuous representer. For every $v \in H^m$, there exists a uniquely defined interpolant in $S_h = S_h(\mathcal{T}_h)$ associated with the points in 5.6. We denote it by $I_h v$. The goal of this section is to estimate

$$\|v - I_h v\|_{m,h} \quad \text{by} \quad \|v\|_{t,\Omega} \quad \text{for } m \leq t.$$

The Bramble–Hilbert Lemma

First we obtain an error estimate for interpolation by polynomials. We begin by establishing the result for all domains which satisfy the hypotheses of the imbedding theorem. Later we shall apply it primarily to reference elements, i.e., on convex triangles and quadrilaterals.

6.2 Lemma. *Let $\Omega \subset \mathbb{R}^2$ be a domain with Lipschitz continuous boundary which satisfies a cone condition. In addition, let $t \geq 2$, and suppose $z_1, z_2, \ldots, z_s$ are $s := t(t+1)/2$ prescribed points in $\bar{\Omega}$ such that the interpolation operator $I : H^t \to \mathcal{P}_{t-1}$ is well defined for polynomials of degree $\leq t-1$. Then there exists a constant $c = c(\Omega, z_1, \ldots, z_s)$ such that*

$$\|u - Iu\|_t \leq c|u|_t \quad \text{for all } u \in H^t(\Omega). \tag{6.2}$$

Proof. We endow $H^t(\Omega)$ with the norm

$$\|\|v\|\| := |v|_t + \sum_{i=1}^s |v(z_i)|,$$

and show that the norms $\|\|\cdot\|\|$ and $\|\cdot\|_t$ are equivalent. Then (6.2) will follow from

$$
\begin{aligned}
\|u - Iu\|_t &\leq c\|\|u - Iu\|\| \\
&= c\Big[|u - Iu|_t + \sum_{i=1}^s |(u - Iu)(z_i)|\Big] \\
&= c|u - Iu|_t = c|u|_t.
\end{aligned}
$$

Here we have made use of the fact that Iu coincides with u at the interpolation points, since $D^\alpha Iu = 0$ for all $|\alpha| = t$.

One direction of the proof of equivalence of the norms is simple. By Remarks 3.4, the imbedding $H^t \hookrightarrow H^2 \hookrightarrow C^0$ is continuous. This implies

$$|v(z_i)| \leq c\|v\|_t \quad \text{for } i = 1, 2, \ldots, s,$$

and thus $\|\|v\|\| \leq (1 + cs)\|v\|_t$.

Suppose now that the converse

$$\|v\|_t \leq c\|\|v\|\| \quad \text{for all } v \in H^t(\Omega) \tag{6.3}$$

fails for every positive number c. Then there exists a sequence (v_k) in $H^t(\Omega)$ with

$$\|v_k\|_t = 1, \quad \|\|v_k\|\| \leq \frac{1}{k}, \quad k = 1, 2, \ldots .$$

By the Rellich selection theorem (Theorem 1.9), a subsequence of (v_k) converges in $H^{t-1}(\Omega)$. Without loss of generality, we can assume that the sequence itself converges. Then (v_k) is a Cauchy sequence in $H^{t-1}(\Omega)$. From $|v_k|_t \to 0$ and $\|v_k - v_\ell\|_t^2 \le \|v_k - v_\ell\|_{t-1}^2 + (|v_k|_t + |v_\ell|_t)^2$, we conclude that (v_k) is even a Cauchy sequence in $H^t(\Omega)$. Because of the completeness of the space, this establishes convergence in the sense of H^t to an element $v^* \in H^t(\Omega)$. By continuity considerations, we have

$$\|v^*\|_t = 1 \quad \text{and} \quad \|\|v^*\|\| = 0.$$

This is a contradiction, since $|v^*|_t = 0$ implies v^* is a polynomial in $\mathcal{P}_{t-1}$, and in view of $v^*(z_i) = 0$ for $i = 1, 2, \ldots, s$, v^* can only be the null polynomial. $\square$

Using the lemma, we now immediately get the following result [Bramble and Hilbert 1970]. As usual, the kernel of a linear mapping L is denoted by $\ker L$, and $\|L\| := \sup\{\|Lv\|; \ \|v\| = 1\}$.

6.3 Bramble–Hilbert Lemma. *Let $\Omega \subset \mathbb{R}^2$ be a domain with Lipschitz continuous boundary. Suppose $t \ge 2$ and that L is a bounded linear mapping of $H^t(\Omega)$ into a normed linear space Y. If $\mathcal{P}_{t-1} \subset \ker L$, then there exists a constant $c = c(\Omega)\|L\| \ge 0$ such that*

$$\|Lv\| \le c|v|_t \quad \text{for all } v \in H^t(\Omega). \tag{6.4}$$

Proof. Let $I : H^t(\Omega) \to \mathcal{P}_{t-1}$ be an interpolation operator of the type appearing in the previous lemma. Using the lemma and the fact that $Iv \in \ker L$, we get

$$\|Lv\| = \|L(v - Iv)\| \le \|L\| \cdot \|v - Iv\|_t \le c\|L\| \cdot |v|_t,$$

where c is the constant in (6.2). $\square$

For simplicity, we have restricted ourselves to bounded two-dimensional domains in order to be able to use the Lagrange interpolation polynomials. This restriction can be removed by utilizing other interpolation procedures; cf. 6.9 and Problem 6.16.

Triangular Elements with Complete Polynomials

We turn our attention once again to C^0 elements consisting of piecewise polynomials of degree $t - 1$ on triangles. Assume $t \ge 2$. Given a triangulation $\mathcal{T}_h$ and an associated family $S_h = \mathcal{M}_0^{t-1}(\mathcal{T}_h)$, by §5 there is a well-defined interpolation operator $I_h : H^t(\Omega) \to S_h$. Moreover, by Definition 5.1(3), $\mathcal{T}_h$ is associated with a shape parameter κ. The central result is the following approximation theorem.

6.4 Theorem. *Let $t \geq 2$, and suppose $\mathcal{T}_h$ is a shape-regular triangulation of Ω. Then there exists a constant $c = c(\Omega, \kappa, t)$ such that*

$$\|u - I_h u\|_{m,h} \leq c\, h^{t-m} |u|_{t,\Omega} \quad \text{for } u \in H^t(\Omega), \quad 0 \leq m \leq t, \qquad (6.5)$$

where I_h denotes interpolation by a piecewise polynomial of degree $t - 1$.

We present the proof of this approximation theorem in full generality later. For the moment we restrict ourselves to the case of a regular grid, i.e., to the case where all triangles are congruent as in Example 4.3. Each triangle can thus be considered to be a scaled version of a reference triangle T_1.

6.5 Remark. Let $t \geq 2$, and suppose

$$T_h := h T_1 = \{x = hy; \; y \in T_1\}$$

with $h \leq 1$. Then

$$\|u - Iu\|_{m,T_h} \leq c\, h^{t-m} |u|_{t,T_h}, \qquad (6.6)$$

for $0 \leq m \leq t$, where Iu is the polynomial in $\mathcal{P}_{t-1}$ which interpolates u (at the transformed points). Here c is the constant in Lemma 6.2.

Proof of the remark. Given a function $u \in H^t(T_h)$, we define $v \in H^t(T_1)$ by

$$v(y) = u(hy).$$

Then $\partial^\alpha v = h^{|\alpha|} \partial^\alpha u$ for $|\alpha| \leq t$. Since the transformation of the area in $\mathbb{R}^2$ yields an extra factor h^{-2}, we get

$$|v|^2_{\ell,T_1} = \sum_{|\alpha|=\ell} \int_{T_1} (\partial^\alpha v)^2 dy = \sum_{|\alpha|=\ell} \int_{T_h} h^{2\ell} (\partial^\alpha u)^2 h^{-2} dx = h^{2\ell-2} |u|^2_{\ell,T_h}.$$

Assuming $h \leq 1$, after summation the smallest power dominates:

$$\|u\|^2_{m,T_h} = \sum_{\ell \leq m} |u|^2_{\ell,T_h} = \sum_{\ell \leq m} h^{-2\ell+2} |v|^2_{\ell,T_1} \leq h^{-2m+2} \|v\|^2_{m,T_1}.$$

Now inserting $u - Iu$ in place of u in this formula, we get a result for the interpolation error. Combining the last two formulas with Lemma 6.2, we get

$$\begin{aligned}
\|u - Iu\|_{m,T_h} &\leq h^{-m+1} \|v - Iv\|_{m,T_1} \leq h^{-m+1} \|v - Iv\|_{t,T_1} \\
&\leq h^{-m+1} c |v|_{t,T_1} \\
&\leq h^{t-m} c |u|_{t,T_h},
\end{aligned}$$

for all $m \leq t$, and (6.6) is proved. $\qquad \square$

For a regular grid, the assertion of Theorem 6.4 is a direct consequence of Remark 6.5, since we get $\|u - I_h u\|_{m,\Omega}$ immediately by squaring the expressions in (6.6), and summing over all triangles.

We now examine triangular elements in more detail in preparation for the proof of the general case of Theorem 6.4, which follows the same lines as for the special case of a regular grid, although the technical difficulties are much greater.

6.6 Transformation Formula. *Let Ω and $\hat{\Omega}$ be affine equivalent, i.e., there exists a bijective affine mapping*

$$F : \hat{\Omega} \to \Omega,$$
$$F\hat{x} = x_0 + B\hat{x} \tag{6.7}$$

with a nonsingular matrix B. If $v \in H^m(\Omega)$, then $\hat{v} := v \circ F \in H^m(\hat{\Omega})$, and there exists a constant $c = c(\hat{\Omega}, m)$ such that

$$|\hat{v}|_{m,\hat{\Omega}} \le c \, \|B\|^m \, |\det B|^{-1/2} |v|_{m,\Omega}. \tag{6.8}$$

Proof. Consider the derivative of order m as a multilinear form, and write the chain rule in the form

$$D^m \hat{v}(\hat{x})(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m) = D^m v(x)(B\hat{y}_1, B\hat{y}_2, \ldots, B\hat{y}_m).$$

Thus, $\|D^m \hat{v}\|_{\mathbb{R}^{nm}} \le \|B\|^m \|D^m v\|_{\mathbb{R}^{nm}}$. We apply this estimate to the partial derivatives $\partial_{i_1} \partial_{i_2} \ldots \partial_{i_m} v = D^m v(e_{i_1}, e_{i_2}, \ldots, e_{i_m})$. Taking the sum, we get

$$\sum_{|\alpha|=m} |\partial^\alpha \hat{v}|^2 \le n^m \max_{|\alpha|=m} |\partial^\alpha \hat{v}|^2 \le n^m \|D^m \hat{v}\|^2 \le n^m \|B\|^{2m} \|D^m v\|^2$$

$$\le n^{2m} \|B\|^{2m} \sum_{|\alpha|=m} |\partial^\alpha v|^2.$$

Finally we integrate, taking account of the *transformation formula* for multiple integrals:

$$\int_{\hat{\Omega}} \sum_{|\alpha|=m} |\partial^\alpha \hat{v}|^2 d\hat{x} \le n^{2m} \|B\|^{2m} \int_\Omega \sum_{|\alpha|=m} |\partial_\alpha v|^2 \cdot |\det B^{-1}| dx.$$
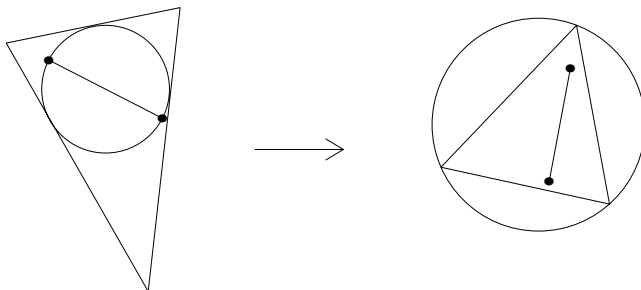
Taking the square root, we get (6.8). $\qquad\square$

Fig. 23. An affine map from a triangle T_1 onto a triangle T_2 sends a pair of points on a circle inscribed in T_1 to points in a circle which contains T_2

The fact that transformations to and from shape-regular grids do not generate extra terms with powers of h can also be seen from simple geometric considerations. Let $F : T_1 \to T_2 : \hat{x} \mapsto B\hat{x} + x_0$ be a bijective affine mapping. We write ρ_i for the radius of the largest circle inscribed in T_i, and r_i for the radius of the smallest circle containing T_i. Given $x \in \mathbb{R}^2$ with $\|x\| \leq 2\rho_1$, we find two points $y_1, z_1 \in T_1$ with $x = y_1 - z_1$, see Fig. 23. Since $F(y_1), F(z_1) \in T_2$, we have $\|Bx\| \leq 2r_2$. Thus,

$$\|B\| \leq \frac{r_2}{\rho_1}. \tag{6.9}$$

Now exchanging T_1 and T_2, we see that the inverse matrix satisfies $\|B^{-1}\| \leq r_1/\rho_2$, and thus

$$\|B\| \cdot \|B^{-1}\| \leq \frac{r_1 r_2}{\rho_1 \rho_2}. \tag{6.10}$$

Proof of Theorem 6.4. It suffices to establish the inequality

$$\|u - I_h u\|_{m, T_j} \leq ch^{t-m} |u|_{t, T_j} \quad \text{for all } u \in H^t(T_j)$$

for every triangle T_j of a shape-regular triangulation $\mathcal{T}_h$. To this end, choose a reference triangle (5.11) with $\hat{r} = 2^{-1/2}$ and $\hat{\rho} = (2 + \sqrt{2})^{-1} \geq 2/7$. Now let $F : T_{\text{ref}} \to T$ with $T = T_j \in \mathcal{T}_h$. Applying Lemma 6.2 on the reference triangle and using the transformation formula in both directions, we obtain

$$
\begin{aligned}
|u - I_h u|_{m, T} &\leq c\|B\|^{-m} |\det B|^{1/2} \, |\hat{u} - I_h \hat{u}|_{m, T_{\text{ref}}} \\
&\leq c\|B\|^{-m} |\det B|^{1/2} \cdot c|\hat{u}|_{t, T_{\text{ref}}} \\
&\leq c\|B\|^{-m} |\det B|^{1/2} \cdot c\|B\|^t \cdot |\det B|^{-1/2} \, |u|_{t, T} \\
&\leq c \, (\|B\| \cdot \|B^{-1}\|)^m \, \|B\|^{t-m} \, |u|_{t, T}.
\end{aligned}
$$

By the shape regularity, $r/\rho \leq \kappa$, and $\|B\| \cdot \|B^{-1}\| \leq (2 + \sqrt{2})\kappa$. Then (6.9) implies $\|B\| \leq h/\hat{\rho} \leq 4h$. Combining these facts, we have

$$|u - I_h u|_{\ell, T} \leq ch^{t-\ell} |u|_{t, T}.$$

Now squaring and summing over ℓ from 0 to m establishes the assertion. $\qquad \square$

Bilinear Quadrilateral Elements

For quadrilateral elements, we usually use tensor products instead of complete polynomials. Nevertheless, we can still make use of the techniques developed in the previous section to establish results on the order of approximation. The simple but important case of a bilinear element serves as a typical example.

Table 3. Error estimates for some finite elements

$\|u - I_h u\|_{m,h} \leq ch^{t-m}|u|_{t,\Omega}$	$0 \leq m \leq t$
C^0 elements	
linear triangle	$t = 2$
quadratic triangle	$2 \leq t \leq 3$
cubic triangle	$2 \leq t \leq 4$
bilinear quadrilateral	$t = 2$
serendipity element	$2 \leq t \leq 3$
9 node quadrilateral	$2 \leq t \leq 3$
C^1 elements	
Argyris element	$3 \leq t \leq 6$
Bell element	$3 \leq t \leq 5$
Hsieh–Clough–Tocher element	$3 \leq t \leq 4 \quad (m \leq 2)$
reduc. Hsieh–Clough–Tocher element	$t = 3 \quad (m \leq 2)$

6.7 Theorem. *Let $\mathcal{T}_h$ be a quasi-uniform decomposition of Ω into parallelograms. Then there exists a constant $c = c(\Omega, \kappa)$ such that*

$$\|u - I_h u\|_{m,\Omega} \leq ch^{2-m}|u|_{2,\Omega} \quad \text{for all } u \in H^2(\Omega),$$

where $I_h u$ interpolates u using bilinear elements.
Proof. For the same reasons as in the last proof, it suffices to show that for interpolation on the unit square $Q := [0, 1]^2$,

$$\|u - Iu\|_{2,Q} \leq c|u|_{2,Q} \quad \text{for all } u \in H^2(Q). \tag{6.11}$$

In view of the continuous imbedding $H^2(Q) \hookrightarrow C^0(Q)$, the function values of u at the 4 vertices are bounded by $c\|u\|_{2,Q}$. The interpolating polynomial Iu depends linearly on these 4 values, and thus $\|Iu\|_{2,Q} \leq c_1 \max_{x \in Q} |u(x)| \leq c_2\|u\|_{2,Q}$ and

$$\|u - Iu\|_2 \leq \|u\|_2 + \|Iu\|_2 \leq (c_2 + 1)\|u\|_2.$$

If u is a linear polynomial, then $Iu = u$, and $u - Iu = 0$. The Bramble–Hilbert lemma now guarantees (6.11). $\qquad\square$

Analogously, for elements in the serendipity class, we have

$$\|u - I_h u\|_{m,\Omega} \leq ch^{t-m}|u|_{t,\Omega} \quad \text{for all } u \in H^t(\Omega), \ m = 0, 1 \text{ and } t = 2, 3.$$

The approximation properties for other triangular and quadrilateral elements are listed in Table 3.

Inverse Estimates

The above approximation theorems have the form

$$\|u - Iu\|_{m,h} \leq ch^{t-m}\|u\|_t,$$

where m is *smaller* than t. For the moment we ignore the fact that on the right-hand side, the norm $\|\cdot\|_t$ may be replaced by the semi-norm $|\cdot|_t$. Thus, the approximation error is measured in a *coarser* norm than the given function. In a so-called *inverse estimate*, the reverse happens. The *finer* norm of the finite element functions will be estimated by a *coarser* one (obviously, this does not work for all functions in the Sobolev space).

6.8 Inverse Estimates. *Let (S_h) be an affine family of finite elements consisting of piecewise polynomials of degree k associated with uniform partitions. Then there exists a constant $c = c(\kappa, k, t)$ such that for all $0 \leq m \leq t$,*

$$\|v_h\|_{t,h} \leq ch^{m-t}\|v_h\|_{m,h} \quad \text{for all } v_h \in S_h.$$

Sketch of the proof. We can reduce the proof to the discussion of a reference element by using the transformation formula 6.6. It suffices to show that

$$|v|_{t,T_{\text{ref}}} \leq c|v|_{m,T_{\text{ref}}} \quad \text{for } v \in \Pi_{\text{ref}} \tag{6.12}$$

with $c = c(\Pi_{\text{ref}})$. The extension to elements of size h proceeds just as in the proof of Theorem 6.4. This leads to the factor ch^{m-t} in the estimate. Then summing the squares of the expressions over all triangles or quadrilaterals leads to the desired assertion.

To establish (6.12), we make use of the fact that the norms $\|\cdot\|_{t,T_{\text{ref}}}$ and $\|\cdot\|_{m,T_{\text{ref}}}$ are equivalent on the finite dimensional space $\Pi_{\text{ref}} \oplus \mathcal{P}_{m-1}$. Let $Iv \in \mathcal{P}_{m-1}$ be a polynomial that interpolates v at fixed points. Since $t > mi1$, we have $|Iv|_t = 0$. Combining these facts, we obtain from the Bramble–Hilbert lemma

$$\begin{aligned}
|v|_t = |v - Iv|_t &\leq \|v - Iv\|_t \\
&\leq c\|v - Iv\|_m \\
&= c'|v|_m,
\end{aligned}$$

and (6.12) is proved. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

In the Approximation Theorem 6.4 and the Inverse Estimate 6.8, the exponents in the term with h correspond to the difference between the orders of the Sobolev norms. This has been established by moving back and forth to and from the reference triangle. This technique is called a *scaling argument*.
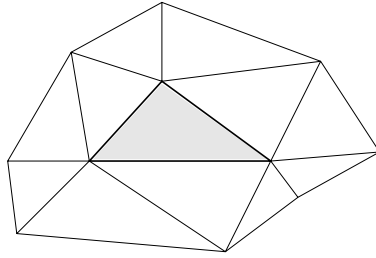
Fig. 24. The values of the Clément interpolant in the (shaded) triangle T depend on the values of the given function in its neighborhood $\tilde{\omega}_T$

Clément's Interpolation

The interpolation operator I_h in (6.5) can only be applied to H^2 functions. On the other hand, functions with less regularity can be approximated in some advanced theories. Clément [1975] has constructed an interpolation process which applies to H^1 functions. Typically this operator is used when features in H^1 and L^2 are to be combined. The crucial point is that the interpolation error depends only on the local mesh size. Thus, no power of h is lost, even if inverse estimates enter into the analysis.

The operator is defined *nearly locally*. Let $\mathcal{T}_h$ be a shape-regular triangulation of Ω. Given a node x_j, let

$$\omega_j := \omega_{x_j} := \bigcup \{T' \in \mathcal{T}_h; \, x_j \in T'\} \tag{6.13}$$

be the support of the shape function $v_j \in \mathcal{M}_0^1$. Here $v_j(x_k) = \delta_{jk}$. Furthermore, let

$$\tilde{\omega}_T := \bigcup \{\omega_j; \, x_j \in T\} \tag{6.14}$$

be a neighborhood of T. Since $\mathcal{T}_h$ is assumed to be shape regular, the area can be estimated by $\mu(\tilde{\omega}_T) \leq c(\kappa) \, h_T^2$. Moreover, the number of triangles that belong to $\tilde{\omega}_T$ is bounded.

6.9 Clément's Interpolation. *Let $\mathcal{T}_h$ be a shape-regular triangulation of Ω. Then there exists a linear mapping $I_h : H^1(\Omega) \to \mathcal{M}_0^1$ such that*

$$\begin{aligned} \|v - I_h v\|_{m,T} &\leq c h_T^{1-m} \|v\|_{1,\tilde{\omega}_T} \quad \text{for } v \in H^1(\Omega), \ m = 0, 1, \ T \in \mathcal{T}_h \\ \|v - I_h v\|_{0,e} &\leq c h_T^{1/2} \|v\|_{1,\tilde{\omega}_T} \quad \text{for } v \in H^1(\Omega), \ e \in \partial T, \ T \in \mathcal{T}_h. \end{aligned} \tag{6.15}$$

A simple construction is obtained by a combination of Clément's operator and the procedures of Scott and Zhang [1990] or Yserentant [1990]. The construction is

performed in two steps. Given a nodal point x_j, let $Q_j : L_2(\omega_j) \to \mathcal{P}_0$ be the L_2-projection onto the constant functions. It follows from the Bramble–Hilbert lemma that

$$\|v - Q_j v\|_{0,\omega_j} \le ch_j |v|_{1,\omega_j}, \tag{6.16}$$

where h_j is the diameter of ω_j. In order to cope with homogeneous Dirichlet boundary conditions on $\Gamma_D \subset \partial\Omega$ we modify the operator and set

$$\tilde{Q}_j v = \begin{cases} 0 & \text{if } x_j \in \Gamma_D, \\ Q_j v & \text{otherwise.} \end{cases} \tag{6.17}$$

Here we get an analogous estimate to (6.16) by adapting the technique of the proof of Friedrich's inequality

$$\|v - \tilde{Q}_j v\|_{0,\omega_j} = \|v\|_{0,\omega_j} \le ch_j |v|_{1,\omega_j} \quad \text{if } x_j \in \Gamma_D. \tag{6.18}$$

Next we define

$$I_h v := \sum_j (\tilde{Q}_j v) v_j \in \mathcal{M}_0^1. \tag{6.19}$$

The shape functions v_j constitute a partition of unity. Specifically, for each x, $I_h v$ contains at most three nonzero terms. For each relevant term, $v - \tilde{Q}_j v$ can be estimated by (6.16) or (6.18). resp.

$$\|v - I_h v\|_{0,T} \le \sum_j \|(v - \tilde{Q}_j v) v_j\|_{0,T} \le \sum_j \|(v - \tilde{Q}_j v)\|_{0,\omega_j} \le 3ch_T \|v\|_{1,\tilde{\omega}_T}.$$

This proves (6.15a) for $m = 0$. For the H^1-stability we refer to Corollary 7.8. $\qquad \square$

The construction is easily modified to get an analogous mapping from $H_0^1(\Omega)$ to $\mathcal{M}_0^1 \cap H_0^1(\Omega)$. If $x_j \in \partial\Omega$, then $P_j v$ may be set to zero and (6.16) follows from Friedrichs' inequality.

Appendix: On the Optimality of the Estimates

6.10 Remark. The inverse estimates show that the above approximation theorems are optimal (up to a constant). The assertions have the following structure:

Suppose the complete normed linear space X is compactly imbedded in Y. Then there exists a family (S_h) of subspaces of X satisfying the *approximation property*

$$\inf_{v_h \in S_h} \|u - v_h\|_Y \le ch^\alpha \|u\|_X \quad \text{for all } u \in X, \tag{6.20}$$

and (with $\beta = \alpha$) the *inverse estimate*

$$\|v_h\|_X \le ch^{-\beta} \|v_h\|_Y \quad \text{for all } v_h \in S_h. \tag{6.21}$$

An example of this is provided by $\| \cdot \|_X = \| \cdot \|_{2,h}$, $\| \cdot \|_Y = \| \cdot \|_{1,\Omega}$, $S_h = \mathcal{M}_0^1(\mathcal{T}_h)$, $\alpha = \beta = 1$.

A pair of inequalities of the form (6.20) and (6.21) involving an approximation property and an inverse estimate is called *optimal* provided that $\beta = \alpha$. We claim $\beta < \alpha$ is impossible. Indeed, otherwise there would exist a sequence of nested spaces $V_0 \subset V_1 \subset V_2 \subset \cdots$ with

$$\min_{v_n \in V_n} \|u - v_n\|_Y \leq 2^{-\gamma n} \|u\|_X \quad \text{for all } u \in X \qquad (6.22)$$

and

$$\|v_n\|_X \leq 2^n \|v_n\|_Y.$$

Here $1 < \gamma < 2$. Choose $m \in \mathbb{N}$ with $2^{-(\gamma+1)m} < (1 - 2^{-\gamma-1})/5$. In view of the compact imbedding of X in Y, there exists an element $u \in X$ with $\|u\|_X = 1$ and $\|u\|_Y < \varepsilon := 2^{-\gamma m}$. Suppose (6.22) holds for $v_n \in V_n$. Set

$$w_m = v_m, \quad w_n = v_n - v_{n-1} \quad \text{for } n > m.$$

Then $\|w_m\|_Y \leq \|u - w_m\|_Y + \|u\|_Y \leq 2 \cdot 2^{-\gamma m}$, and

$$\|w_n\|_Y \leq \|u - v_n\|_Y + \|u - v_{n-1}\|_Y \leq 2^{-\gamma n} + 2^{-\gamma(n-1)} \leq 5 \cdot 2^{-\gamma n}$$

for $n > m$. In view of the inverse inequality, it follows that $\|w_n\|_X \leq 5 \cdot 2^{-(\gamma-1)n}$ for $n \geq m$, and

$$\|u\|_X = \|\sum_{n=m}^{\infty} w_n\|_X \leq 5 \sum_{n=m}^{\infty} 2^{-(\gamma-1)n} < 1.$$

This is a contradiction. $\square$

6.11 Remark. The above proof also establishes that if

$$\inf_{v_h \in S_h} \|u - v_h\|_Y \leq \text{const} \cdot h^{\beta}$$

for all $u \in Y$, and the inverse estimate (6.21) also holds, then u is contained in the subspace X. This result has far-reaching consequences for the practical use of finite elements. If it is known that the solution u of a boundary-value problem does not lie in a higher-order Sobolev space, then the finite element approximation has limited accuracy.

Here we should note that pairs of inequalities of the form (6.20) and (6.21) play a major role in classical approximation theory. The most widely known results along these lines deal with the approximation of 2π-periodic functions by trigonometric

polynomials in $\mathcal{P}_{n,2\pi}$. Let $C^{k+\alpha}$ denote the space of functions whose k-th derivative is Hölder continuous with exponent α. Then by the theorems of Jackson,

$$\inf_{p \in \mathcal{P}_{n,2\pi}} \|f - p\|_{C^0} \le c n^{-k-\alpha} \|f\|_{C^{k+\alpha}},$$

while the Bernstein inequality

$$\|p\|_{C^{k+\alpha}} \le c n^{k+\alpha} \|p\|_{C^0} \quad \text{for all } p \in \mathcal{P}_{n,2\pi}$$

provides the corresponding inverse estimate.

Problems

6.12 Let $\mathcal{T}_h$ be a family of uniform partitions of Ω, and suppose S_h belong to an affine family of finite elements. Suppose the nodes of the basis are $z_1, z_2, \ldots, z_N$ with $N = N_h = \dim S_h$. Verify that for some constant c independent of h, the following inequality holds:

$$c^{-1} \|v\|_{0,\Omega}^2 \le h^2 \sum_{i=1}^N |v(z_i)|^2 \le c \|v\|_{0,\Omega}^2 \quad \text{for all } v \in S_h.$$

6.13 Under appropriate assumptions on the boundary of Ω, we showed that

$$\inf_{v \in S_h} \|u - v_h\|_{1,\Omega} \le c\,h \|u\|_{2,\Omega},$$

where for every $h > 0$, S_h is a finite-dimensional finite element space. Show that this implies the compactness of the imbedding $H^2(\Omega) \hookrightarrow H^1(\Omega)$. [Thus, the use of the compactness in the proof of the approximation theorem was not just a coincidence.]

6.14 Let $\mathcal{T}_h$ be a κ-regular partition of Ω into parallelograms, and let u_h be an associated bilinear element. Divide each parallelogram into two triangles, and let $\| \cdot \|_{m,h}$ be defined as in (6.1). Show that

$$\inf \|u_h - v_h\|_{m,\Omega} \le c(\kappa) h^{2-m} \|u_h\|_{2,\Omega}, \quad m = 0, 1,$$

where the infimum is taken over all piecewise linear functions on the triangles in $\mathcal{M}^1$.

6.15 For interpolation by piecewise linear functions, Theorem 6.4 asserts that

$$\|I_h v\|_{2,h} \le c \|v\|_{2,\Omega}.$$

Give a one-dimensional counterexample to show that

$$\|I_h v\|_{0,\Omega} \le c \|v\|_{0,\Omega}$$

is not possible with a constant c which is independent of h.

6.16 Prove the Bramble–Hilbert lemma for $t = 1$ by choosing $I v$ to be the constant function

$$I v := \frac{\int_\Omega v \, dx}{\int_\Omega dx}.$$

6.17 Consider the situation as in the construction of Cléments' operator. We modify the definition of the operator $\tilde{Q}_j : L_2(\omega_j) \to \mathcal{P}_0$ by the rule

$$\tilde{Q}_j v := v(x_j) \quad \text{if } v|_{\omega_j} \in \mathcal{M}_0^1(\mathcal{T}_\ell), \tag{6.23}$$

i.e., if the restriction of v to ω_j is a finite element function with the present grid. Show that also in this case

$$\|v - \tilde{Q}_j v\|_{0,\omega_j} \le h_j |v|_{1,\omega_j}$$

with c depending only on the shape parameter of the triangulation of ω_j.

Hint: The modification has an advantage. If the given function coincides with a piecewise linear function on a subdomain $\tilde{\Omega}$, then the projector reproduces v at the nodes in the interior of $\tilde{\Omega}$.

§ 7. Error Bounds for Elliptic Problems
of Second Order

Now we are ready to establish error estimates for finite element solutions. Usually error bounds are derived first with respect to the energy norm. The extension to the L_2-norm is performed by a duality technique that is often found in proofs of advanced results. We are looking for bounds of the form

$$\|u - u_h\| \leq c \, h^p \tag{7.1}$$

for the difference between the true solution u and the approximate solution u_h in S_h. Here p is called the *order of approximation*. In general, it depends on the regularity of the solution, the degree of the polynomials in the finite elements, and the Sobolev norm in which the error is measured.

Remarks on Regularity

7.1 Definition. Let $m \geq 1$, $H_0^m(\Omega) \subset V \subset H^m(\Omega)$, and suppose $a(\cdot, \cdot)$ is a V-elliptic bilinear form. Then the variational problem

$$a(u, v) = (f, v)_0 \quad \text{for all} \ \ v \in V$$

is called H^s-*regular* provided that there exists a constant $c = c(\Omega, a, s)$ such that for every $f \in H^{s-2m}(\Omega)$, there is a solution $u \in H^s(\Omega)$ with

$$\|u\|_s \leq c \, \|f\|_{s-2m}. \tag{7.2}$$

In this section we will make use of this definition only for $s \geq 2m$. We will drop this restriction later, in Ch. III, after norms with negative index are defined.

Regularity results for the Dirichlet problem of second order with zero boundary conditions can be found, e.g., in Gilbarg and Trudinger [1983] and Kadlec [1964]. For simplicity, we do not present the most general results; see the remarks for Example 2.10 and Problem 7.12.

7.2 Regularity Theorem. *Let a be an H_0^1-elliptic bilinear form with sufficiently smooth coefficient functions.*
(1) If Ω is convex, then the Dirichlet problem is H^2-regular.
(2) If Ω has a C^s boundary with $s \geq 2$, then the Dirichlet problem is H^s-regular.
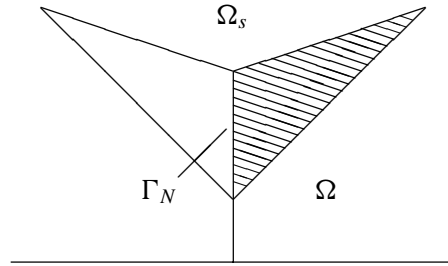
Fig. 25. Reflection of a convex domain Ω along the edge Γ_N on which a Neumann condition is given

We see from Example 2.1 with a domain with reentrant corner that the assumptions on the boundary cannot be dropped since the solution there is not in $H^2(\Omega)$.

We now give an example to show that the situation is more complicated if a Neumann condition is prescribed *on a part of the boundary*. Let Ω be the convex domain on the right-hand side of the y-axis shown in Fig. 25. Suppose the Neumann condition

$$\frac{\partial u}{\partial \nu} = 0$$

is prescribed on $\Gamma_N := \{(x, y) \in \partial\Omega;\ x = 0\}$, and that a Dirichlet boundary condition is prescribed on $\Gamma_D := \Gamma \setminus \Gamma_N$. The union of Ω with its reflection in Γ_N defines a symmetric domain Ω_s. Set

$$u(-x, y) = u(x, y) \quad \text{for } (x, y) \in \Omega_s \setminus \Omega.$$

Then its continuation is also a solution of a Dirichlet problem on Ω_s. But since Ω_s has a reentrant corner, the solution is not always in $H^2(\Omega_s)$, which means that $u \in H^2(\Omega)$ cannot hold for all problems on Ω with mixed boundary conditions.

Error Bounds in the Energy Norm

In the following, suppose that Ω is a polygonal domain. This means that it can be partitioned into triangles or quadrilaterals. In addition, in order to use Theorem 7.2, suppose Ω is convex.

7.3 Theorem. *Suppose $\mathcal{T}_h$ is a family of shape-regular triangulations of Ω. Then the finite element approximation $u_h \in S_h = \mathcal{M}_0^k$ ($k \geq 1$) satisfies*

$$\begin{aligned}
\|u - u_h\|_1 &\leq ch\|u\|_2 \\
&\leq ch\|f\|_0.
\end{aligned} \tag{7.3}$$

Proof. By the convexity of Ω, the problem is H^2-regular, and $\|u\|_2 \leq c_1\|f\|_0$. By Theorem 6.4, there exists $v_h \in S_h$ with $\|u - v_h\|_{1,\Omega} = \|u - v_h\|_{1,h} \leq c_2 h\|u\|_{2,\Omega}$. Combining these facts with Céa's Lemma gives (7.3) with $c := (1 + c_1)c_2 c_3 / \alpha$. $\square$

7.4 Remark. According to Theorem 6.4, we should get a higher-order error bound for quadratic triangular elements under the assumption of H^3-regularity. This observation is misleading, however, since – except in some special cases – smooth boundaries are required for H^3-regularity. But a domain Ω with smooth boundary cannot be decomposed into triangles, and the usual problems arise along the curved boundaries (cf. Ch. III, §1).

There is more regularity in the interior of the domain, and in most cases, the finite element approximations with quadratic or cubic triangles are so much better than with piecewise linear ones that it is worth the extra effort to use them.

The estimate (7.3) holds for any affine family of triangular elements which contains the P_1 elements as a subset. Moreover, by Theorem 6.7 analogous results hold if we use bilinear quadrilateral elements instead of linear triangles. Using the same arguments as in the proof of the previous theorem, we get

7.5 Theorem. *Suppose we are given a set of shape-regular partitions of Ω into parallelograms. Then the finite element approximation u_h by bilinear quadrilateral elements in S_h satisfies*

$$\|u - u_h\|_1 \le ch\|f\|_0. \tag{7.4}$$

L_2-**Estimates**

If the polynomial approximation error is measured in the L_2-norm (i.e., in the H^0-norm), then by Theorem 6.4 the order of approximation is better by one power of h. It is not at all obvious that this property carries over to finite element solutions. The proof uses the H^2-regularity a second time, and requires a *duality argument* which has been called *Nitsche's Trick*. We now present an abstract formulation of it; cf. Aubin [1967] and Nitsche [1968].

7.6 Aubin–Nitsche Lemma. *Let H be a Hilbert space with the norm $|\cdot|$ and the scalar product $(\cdot, \cdot)$. Let V be a subspace which is also a Hilbert space under another norm $\|\cdot\|$. In addition, let*

$$V \hookrightarrow H \quad \text{be continuous.}$$

Then the finite element solution in $S_h \subset V$ satisfies

$$|u - u_h| \le C \|u - u_h\| \sup_{g \in H} \left\{ \frac{1}{|g|} \inf_{v \in S_h} \|\varphi_g - v\| \right\} \tag{7.5}$$

where for every $g \in H$, $\varphi_g \in V$ denotes the corresponding unique (weak) solution of the equation

$$a(w, \varphi_g) = (g, w) \quad \text{for all } w \in V. \tag{7.6}$$

Proof. By duality, the norm of an element in a Hilbert space can be computed by

$$|w| = \sup_{g \in H} \frac{(g, w)}{|g|}. \tag{7.7}$$

Here and in (7.5), the supremum is taken only over those g with $g \neq 0$. We recall that u and u_h are given by

$$a(u, v) = \langle f, v \rangle \quad \text{for all } v \in V,$$
$$a(u_h, v) = \langle f, v \rangle \quad \text{for all } v \in S_h.$$

Hence, $a(u - u_h, v) = 0$ for all $v \in S_h$. Moreover, if we insert $w := u - u_h$ in (7.6), we get

$$
\begin{aligned}
(g, u - u_h) &= a(u - u_h, \varphi_g) \\
&= a(u - u_h, \varphi_g - v) \leq C \|u - u_h\| \cdot \|\varphi_g - v\|.
\end{aligned}
$$

Here we have used the continuity of the bilinear form a, i.e., the fact that $a(u, v) \leq C \|u\| \cdot \|v\|$. It follows that

$$(g, u - u_h) \leq C \|u - u_h\| \inf_{v \in S_h} \|\varphi_g - v\|.$$

Now the duality argument (7.7) leads to

$$
\begin{aligned}
|u - u_h| &= \sup_{g \in H} \frac{(g, u - u_h)}{|g|} \\
&\leq C \|u - u_h\| \sup_{g \in H} \left\{ \inf_{v \in S_h} \frac{\|\varphi_g - v\|}{|g|} \right\}. \qquad \square
\end{aligned}
$$

7.7 Corollary. *Under the hypotheses of either Theorem 7.3 or Theorem 7.5, if $u \in H^1(\Omega)$ is the solution of the associated variational problem, then*

$$\|u - u_h\|_0 \leq cCh \|u - u_h\|_1.$$

If in addition $f \in L_2(\Omega)$ so that $u \in H^2(\Omega)$, then

$$\|u - u_h\|_0 \leq cC^2 h^2 \|f\|_0.$$

Here c and C are the constants appearing in (7.3) and in (7.4)–(7.5), respectively.

Proof. Setting

$$
\begin{aligned}
H &:= H^0(\Omega), \quad |\cdot| := \|\cdot\|_0, \\
V &:= H_0^1(\Omega), \quad \|\cdot\| := \|\cdot\|_1,
\end{aligned}
$$

we see that $V \subset H$, and the continuity of the imbedding is clear from $\|\cdot\|_0 \leq \|\cdot\|_1$. The Aubin–Nitsche Lemma is now applicable. In view of Theorem 7.3 or 7.5, the quantity in the curly brackets in (7.5) is at most ch, and the lemma immediately implies the desired result. $\qquad \square$

A Simple L_∞-Estimate

The above estimates do not exclude the possibility that the error is large at certain points. To prevent this, we need to work with the L_∞-norm $\|v\|_{\infty,\Omega} := \sup_{x\in\Omega} |v(x)|$. For problems in two-dimensional domains with H^2-regularity, we have

$$\|u - u_h\|_\infty \leq c\, h^2 |\log h|^{3/2} \|D^2 u\|_\infty.$$

A proof of this fact based on weighted norms can be found, e.g., in Ciarlet [1978]. Here we restrict ourselves to proving the much weaker assertion

$$\|u - u_h\|_\infty \leq c\, h\, |u|_2. \tag{7.8}$$

Given a function $v \in H^2(T_{\text{ref}})$, let Iv be its interpolant in the polynomial space Π_{ref}. Since $H^2 \subset C^0$, by the Bramble–Hilbert Lemma we have

$$\|v - Iv\|_{\infty,T_{\text{ref}}} \leq c\, |v|_{2,T_{\text{ref}}}. \tag{7.9}$$

Let u be the solution of the variational problem, and let $I_h u$ be its interpolant in S_h. Pick an element T from the triangulation (which we assume to be uniform). Let $\hat{u}$ be the affine transformation of $u|_T$ onto the reference triangle. By (7.9) and the transformation formula (6.8), we get

$$\begin{aligned}
\|u - I_h u\|_{\infty,T} = \|\hat{u} - I\hat{u}\|_{\infty,T_{\text{ref}}} &\leq c|\hat{u}|_{2,T_{\text{ref}}} \\
&\leq ch|u|_{2,T} \leq ch|u|_{2,\Omega}.
\end{aligned} \tag{7.10}$$

Taking the maximum over all triangles, we have

$$\|u - I_h u\|_{\infty,\Omega} \leq c\, h|u|_{2,\Omega}.$$

Similarly, by an affine argument, we get the inverse estimate

$$\|v_h\|_{\infty,\Omega} \leq c\, h^{-1} \|v_h\|_{0,\Omega} \quad \text{for all } v_h \in S_h.$$

Now by Theorems 6.4 and 7.7, it follows that $\|u_h - I_h u\|_{0,\Omega} \leq c\, h^2 |u|_{2,\Omega}$ for $u_h - I_h u = (u - I_h u) - (u - u_h) \in S_h$. Using the inverse estimate, we now get

$$\begin{aligned}
\|u - u_h\|_{\infty,\Omega} &\leq \|u - I_h u\|_{\infty,\Omega} + \|u_h - I_h u\|_{\infty,\Omega} \\
&\leq \|u - I_h u\|_{\infty,\Omega} + c\, h^{-1} \|u_h - I_h u\|_{0,\Omega},
\end{aligned}$$

and the result follows from (6.5) and (7.10). $\qquad\square$

The L_2-Projector

The norm of the L_2-projector onto a finite element space is not always bounded by an h-independent number when it is considered in H^1. The boundedness is easily derived in the case in which we obtain an L_2-estimate by a duality argument. We will encounter the same technique of proof in Ch. III, §6 and Ch. VI, §6.

7.8 Corollary. *Assume that the hypothesis of Theorem 7.3 (or Theorem 7.5) are satisfied, and that $\{\mathcal{T}_h\}$ is a family of uniform triangulations of Ω. Let Q_h be the L_2-projector onto $S_h \subset H^1(\Omega)$. Then*

$$\|Q_h v\|_1 \le c\|v\|_1 \quad \text{for all } v \in H^1(\Omega) \tag{7.11}$$

holds with a constant c which is independent of h.

Proof. Given $v \in H^1(\Omega)$ let $v_h \in S_h$ be the solution of the variational problem

$$(\nabla v_h, \nabla w)_0 + (v_h, w)_0 = \langle \ell, w \rangle \quad \text{for all } w \in S_h$$

with $\langle \ell, w \rangle := (\nabla v, \nabla w)_0 + (v, w)_0$. Obviously, $\|v_h\|_1 \le \|v\|_1$. An essential ingredient is the L_2-error estimate from Corollary 7.7

$$\|v - v_h\|_0 \le c_1 h \|v - v_h\|_1 \le 2c_1 h \|v\|_1. \tag{7.12}$$

Combining this with an inverse estimate, we get

$$\begin{aligned}
\|Q_h v\|_1 &\le \|Q_h v - v_h\|_1 + \|v_h\|_1 \\
&\le c_2 h^{-1} \|Q_h(v - v_h)\|_0 + \|v_h\|_1 \le c_2 h^{-1} \|v - v_h\|_0 + \|v\|_1 \\
&\le c_2 h^{-1} 2c_1 h \|v\|_1 + \|v\|_1.
\end{aligned}$$

This proves the assertion with $c = 1 + 2c_1 c_2$. $\qquad\square$

The assumptions of the corollary are very restrictive and in some cases the stability of the L_2-projector is wanted in locally refined meshes; cf. Ch. V, §5 and Yserentant [1990].

7.9 Lemma. *Let $\mathcal{T}_h$ be a shape-regular triangulation of Ω and Q_h be the L_2-projector onto $\mathcal{M}_0^1$. Then*

$$\|Q_h v\|_1 \le c\|v\|_1 \quad \text{for all } v \in H_0^1(\Omega) \tag{7.13}$$

holds with a constant c which is independent of h.

Proof. We start with Clément's interpolation operator and obtain

$$\|v - I_h v\|_0^2 \le c \sum_T h_T^2 \|v\|_{1,T}^2 . \tag{7.14}$$

Since the triangulation is assumed to be shape-regular, we could estimate the diameters of all triangles in ω_T by $c h_T$, i.e. the diameter of T. The estimates are still local. The minimal property of the L_2-projector Q_h implies

$$\|v - Q_h v\|_0^2 \leq c \sum_T h_T^2 \|v\|_{1,T}^2 . \tag{7.15}$$

Next from the Bramble–Hilbert lemma and a standard scaling argument we know that there is a piecewise constant function $w_h \in \mathcal{M}^0(\mathcal{T}_h)$ such that

$$\|v - w_h\|_{0,T} \leq c h_T |v|_{1,T} .$$

Now we apply an inverse estimate on each triangle:

$$
\begin{aligned}
|Q_h v|_1^2 = \sum_T |Q_h v|_{1,T}^2 &= \sum_T |Q_h v - w_h|_{1,T}^2 \\
&\leq c \sum_T h_T^{-2} \|Q_h v - w_h\|_{0,T}^2 \\
&\leq c \sum_T 2 h_T^{-2} \left(\|Q_h v - v\|_{0,T}^2 + \|v - w_h\|_{0,T}^2 \right) \\
&\leq c \sum_T \|v\|_{1,T}^2 = c \|v\|_1^2 .
\end{aligned}
$$

Since $\|Q_h v\|_0 \leq \|v\|_0 \leq \|v\|_1$, the proof is complete. $\square$

Note that Problem 6.15 illustrates that one has to be careful when dealing with a projector for one norm and considering stability for another one.

Problems

7.10 Consider solving the boundary-value problem

$$
\begin{aligned}
-\Delta u &= 0 && \text{in } \Omega := (-1, +1)^2 \subset \mathbb{R}^2, \\
u(x, y) &= xy && \text{on } \partial\Omega
\end{aligned}
$$

using linear triangular elements on a regular triangular grid with $2/h \in \mathbb{N}$ as in the model problem 4.3. When the reduction to homogeneous boundary conditions as in (2.21) is performed with

$$
u_0(x, y) := \begin{cases} 1 + x - y & \text{for } x \geq y, \\ 1 + y - x & \text{for } x \leq y, \end{cases}
$$

the finite element approximation at the grid points is

$$u_h(x_i, y_i) = u(x_i, y_i) = x_i y_i. \tag{7.16}$$

Verify that the minimal value for the variational functional on S_h is

$$J(u_h) = \frac{8}{3} + \frac{4}{3} h^2,$$

and hence, $J(u_h)$ is only an approximation.

7.11 Let $\Omega = (0, 2\pi)^2$ be a square, and suppose $u \in H_0^1(\Omega)$ is a weak solution of $-\Delta u = f$ with $f \in L_2(\Omega)$. Using Problem 1.16, show that $\Delta u \in L_2(\Omega)$, and then use the Cauchy–Schwarz inequality to show that all second derivatives lie in L_2, and thus u is an H^2 function.

7.12 *(A superconvergence property)* The boundary-value problem with the ordinary differential equation

$$-u''(x) = f(x), \quad x \in (0, 1),$$
$$u(0) = u(1) = 0$$

characterizes the solution of a variational problem with the bilinear form

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

Let u_h be the solution in the set of piecewise linear functions on a partition of $(0, 1)$, and let v_h be the interpolant of u in the same set. Show that $u_h = v_h$ by verifying $a(u_h - v_h, w_h) = 0$ for all piecewise linear w_h.

§ 8. Computational Considerations

The computation of finite element approximations can be divided into two parts:
1. construction of a grid by partitioning Ω, and setting up the stiffness matrix.
2. solution of the system of equations.

The central topic of this section is the computation of the stiffness matrix. The solution of the system of equations will be treated in Chapters IV and V.

Assembling the Stiffness Matrix

For finite elements with a nodal basis, such as the linear and quadratic triangular elements, the stiffness matrix can be assembled elementwise. This can be seen from the associated quadratic form. For simplicity, we consider only the principal part:

$$a(u, v) = \int_{\Omega} \sum_{k,l} a_{kl} \partial_k u \partial_l v \, dx.$$

Then

$$A_{ij} = a(\psi_i, \psi_j) = \int_{\Omega} \sum_{k,l} a_{kl} \partial_k \psi_i \partial_l \psi_j \, dx$$

$$= \sum_{T \in \mathcal{T}_h} \int_T \sum_{k,l} a_{kl} \partial_k \psi_i \partial_l \psi_j \, dx. \tag{8.1}$$

In forming the sum, we need only take account of those triangles which overlap the support of both ψ_i and ψ_j.

Normally, we do not compute this matrix by locating the triangles involved for a given set of node indices i, j. Although we used this type of *node-oriented* approach for the model problem in §4, in practice it wastes too much time in repeated calculations.

Table 4. Shape functions (nodal basis functions) for linear (left) and quadratic (right) elements

$N_1 = 1 - \xi - \eta$ $N_2 = \xi$ $N_3 = \eta$	$N_1 = (1 - \xi - \eta)(1 - 2\xi - 2\eta)$ $N_2 = \xi(2\xi - 1)$ $N_3 = \eta(2\eta - 1)$ $N_4 = 4\xi(1 - \xi - \eta)$ $N_5 = 4\xi\eta$ $N_6 = 4\eta(1 - \xi - \eta)$

It turns out that it is much better to use an *element-oriented* approach. For every element $T \in \mathcal{T}_h$, we find the additive contribution from (8.1) to the stiffness matrix. If every element contains exactly s nodes, this requires finding an $s \times s$ submatrix. We transform the triangle T under consideration to the reference triangle T_{ref}. Let $F : T_{\text{ref}} \to T$, $\xi \mapsto B\xi + x_0$ be the corresponding linear mapping. Then the contribution of T is given by the integral

$$\frac{\mu(T)}{\mu(T_{\text{ref}})} \int_{T_{\text{ref}}} \sum_{\substack{k,l \\ k',l'}} a_{kl} (B^{-1})_{k'k} (B^{-1})_{l'l}\, \partial_{k'} N_i\, \partial_{l'} N_j\, d\xi. \qquad (8.2)$$

Here $\mu(T)$ is the area of T. After transformation, every function in the nodal basis coincides with one of the normed *shape functions* $N_1, N_2, \ldots, N_s$ on the reference triangle. These are listed in Table 4 for linear and quadratic elements.[6] For the model problem 4.3, using a right triangle T (with right angle at point number 1), we get

$$a(u,u)|_T = \frac{1}{2}(u_1 - u_2)^2 + \frac{1}{2}(u_1 - u_3)^2,$$

where u_i is the coefficient of u in the N_i expansion. This gives

$$a(\psi_i, \psi_j)|_T = \frac{1}{2} \begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & \\ -1 & & 1 \end{pmatrix}$$

for the stiffness matrix on the element level. For linear elements, it is also easy to find the so-called *mass matrix* whose elements are $(\psi_i, \psi_j)_{0,T}$. For an arbitrary triangle,

$$(\psi_i, \psi_j)_{0,T} = \frac{\mu(T)}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}. \qquad (8.3)$$

For differential equations with variable coefficients, the evaluation of the integrals (8.2) is usually accomplished using a Gaussian quadrature formula for multiple

---

[6] To avoid indices, in Table 4 we have written ξ and η instead of ξ_1 and ξ_2.

In addition, we note that for the quadratic triangular elements, the basis functions N_1, N_2 and N_3 can be replaced by the corresponding nodal basis functions of linear elements. The coefficients in the expansion $\sum_{i=1}^{6} z_i N_i$ then have a different meaning: z_1, z_2 and z_3 are still the values at the vertices, but z_4, z_5 and z_6 become the deviations at the midpoints of the sides from the linear function which interpolates at the vertices.

This basis is not a purely nodal basis, although the correspondence is very simple. However, it has two advantages: we get simpler integrands in (8.2), and the condition number of the system matrix is generally lower (cf. hierarchical bases).

integrals; cf. Table 5. For equations with constant coefficients, we are usually integrating polynomials which can be computed in closed form by making use of the following formula for the unit triangle (5.9):

$$I_{pqr} := \int\int_{\substack{\xi,\eta\geq 0 \\ 1-\xi-\eta\geq 0}} \xi^p \eta^q (1-\xi-\eta)^r \, d\xi \, d\eta = \frac{p!\,q!\,r!}{(p+q+r+2)!}. \qquad (8.4)$$

The formula (8.4) can be applied to triangles in arbitrary position by replacing ξ, η, and $1-\xi-\eta$ by the barycentric coordinates. – Note that for linear elements, the integrands in (8.2) are actually constants.

Table 5. Sample points (ξ_i, η_i) and weights w_i for Gaussian quadrature formulas for polynomials up to degree 5 over the unit triangle

i	ξ_i	η_i	w_i
1	$1/3$	$1/3$	$9/80$
2	$(6+\sqrt{15})/21$	$(6+\sqrt{15})/21$	
3	$(9-2\sqrt{15})/21$	$(6+\sqrt{15})/21$	$\left.\right\}\;(155+\sqrt{15})/2400$
4	$(6+\sqrt{15})/21$	$(9-2\sqrt{15})/21$	
5	$(6-\sqrt{15})/21$	$(6-\sqrt{15})/21$	
6	$(9+2\sqrt{15})/21$	$(6-\sqrt{15})/21$	$\left.\right\}\;(155-\sqrt{15})/2400$
7	$(6-\sqrt{15})/21$	$(9+2\sqrt{15})/21$	

Static Condensation

Although the stiffness matrix can be assembled additively from $s \times s$ submatrices, the bandwidth is much larger than s (cf. Example 4.3). On the other hand, the variables corresponding to interior nodes of elements are easily treated. Both the nine-point rectangular element and the cubic ten-point triangular element have one interior node, for example.

The elimination of a variable corresponding to an interior node changes only those matrix elements for the nodes of the same element. In particular, *no* zeros are filled in. The work required for the elimination is equivalent to the work needed by the Cholesky method for the elimination of variables in an $s \times s$ matrix, i.e., in a small matrix.

The process of elimination of the variables for all these nodes is called *static condensation*.

Complexity of Setting up the Matrix

In setting up the system matrix, we need to perform Ms^2 matrix element calculations, where M is the number of elements, and s is the number of local degrees of freedom. Thus, clearly one tries to avoid calculating with finite elements that have a large number of local degrees, if possible. Only recently computations with polynomials of high degree have impact on the design of finite element programs. They are so designed that their good approximation properties more than compensate for the increase of the computational effort; see Schwab [1998].

It is for this reason that in practice C^1 elements are not used for systems of partial differential equations. For planar C^1 elements, it is well known that we need at least 12 degrees of freedom per function. Thus, for elliptic systems with three variables, we would have to set up a

$$36 \times 36 \text{ matrix}$$

for each element.

Effect on the Choice of a Grid

Once we have selected an element type, the work required to set up the stiffness matrix is approximately proportional to the number of unknowns. However, the work required for the solution of the corresponding system of linear equations using classical methods increases faster than linearly. For large systems, this can quickly lead to memory problems.

These considerations suggest individually tailoring the grid to the problem in order to reduce the number of variables as much as possible.

With the development of newer methods for solving systems of equations, such as the ones in Chapters IV and V, this problem has become less critical, and once again assembling the matrix constitutes the main part of the work. Thus, it makes more sense to save computation time there if possible. One way to do this is to build the grid so that the elements are all translations of a few basic ones. If the coefficients of the differential equation are piecewise constant functions, the computational effort can be reduced. Dividing each triangle into four congruent parts means here that the matrix elements for the subtriangles can be obtained from those of the original triangles with just a few calculations.

Local Mesh Refinement

A triangle can easily be decomposed into four congruent subtriangles. Thus, using bisection we can easily perform a global grid refinement to halve the mesh size. This process leaves the regularity parameter κ (the maximum ratio of circumcircle radius to the radius of an inscribed circle) unchanged.

Sometimes, as in the following situation, it may be preferable to perform a refinement on only part of a domain Ω:

1. In some subdomain the derivatives (which determine the order of approximation) are much greater than in the rest of the domain. This may be clear from the nature of the problem, or from the computation of error estimators which will be dealt with in Ch. III, §7. In this case, refining this part of the grid can lead to a reduction of the error in the entire domain.

2. We would like to start with a very coarse grid, and let the final grid be determined by automatic refinements. Often it is appropriate for the given problem that the amount of refinement is different in different parts of the domain.

3. We want to compute the solution to higher accuracy in some subdomain.

The fact that in the ideal case it is even possible to carry out a refinement in the direction of an edge or of a vertex using only *similar triangles* is illustrated in Figs. 12 and 13. However, these are exceptional cases. Some care is required in order to generate finer grids from coarser ones automatically. In particular, if more than one level of refinement is used, we have to be careful to avoid thin triangles.
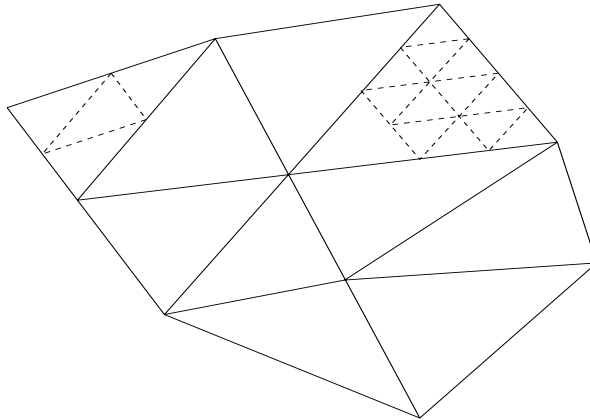

Fig. 26. Coarse grid (solid lines) and a refinement (dotted lines)

The following *refinement rule*, which can be found, e.g., in the multigrid algorithm of Bank [1990], guarantees that each of the angles in the original triangulation is bisected at most once. We may think of starting with a triangulation as in Fig. 26. This triangulation contains several *hanging nodes* (cf. Fig. 11) which must be converted to non-hanging nodes.

8.1 Refinement Rules.

(1) If an edge of a triangle T contains two or more vertices of other triangles (not counting its own vertices), then the triangle T is divided into four congruent subtriangles. This process is repeated until such triangles no longer exist.

(2) Every triangle which contains a vertex of another triangle at the midpoint of one of its edges is divided into two parts. We call the new edge a *green edge*.

(3) If a further refinement is desired, we first eliminate the green edges before proceeding.

For the triangulation in Fig. 26, we first apply rule (1) to the triangles I and VIII. This requires using the rule twice on triangle VII. Next, we construct green edges in the triangles II, V, VI, and in three subtriangles.

Despite its recursive nature, we claim that this procedure stops after a finite number of iterations. Let m denote the maximal number of levels in the desired refinement, where the maximum is to be taken over all elements (in the example, $m = 2$). Then every element will be divided at most m times. This gives an upper bound on the number of steps in (1).

Refinements of Partitions of 3-Dimensional Domain

A triangle is easily divided into four congruent subtriangles, but the situation is more involved in $\mathbb{R}^3$. A tetrahedron cannot be partitioned into eight congruent subtetrahedra; cf. Problem 8.7.

There is a partitioning that was described by Freudenthal [1942] although it is usually called *Kuhn's triangulation*. For its definition, first a cube is decomposed into $3! = 6$ tetrahedra. On the other hand the cube consists of eight subcubes which again can be partitioned into tetrahedra. The latter ones provide a decomposition of the original tetrahedra. This process also shows that six types of tetrahedra are sufficient even if the refinement procedure is repeated several times; cf. Problem 8.7. More questions concerning tetrahedral meshes were discussed by Bey [1995].

There is another technique due to Rivara [1984] that is more convenient than the implementation of Kuhn's triangulation. In the two-dimensional case, it works with splitting triangles by *halving* their longest sides.

Implementation of the Neumann Boundary-Value Problem

When the finite lement equations for the Poisson equation (3.8) with Neumann boundary conditions are assembled for the nodal basis, the stiffness matrix has a one-dimensional kernel. Fortunately, this is no handicap provided that the finite element space contains the constant function.

One can fix the value at one node, e.g., by setting it to zero. The corresponding row and column are eliminated from the matrix-vector equations. The reduced system may be solved by the well-known Cholesky decomposition. The equation that was dropped, holds as a consequence of the compatibility condition (3.9).

The constant function spans the kernel. Thus, finally a suitable constant can be subtracted from all nodal values in order to obtain a solution with mean value zero.

Problems

8.2 Set up the system matrix A_Q for solving the Poisson equation using bilinear quadrilateral elements on the unit square. Note that with a cyclic numbering, by invariance at the element level, we get the form

$$
\begin{pmatrix}
\alpha & \beta & \gamma & \beta \\
\beta & \alpha & \beta & \gamma \\
\gamma & \beta & \alpha & \beta \\
\beta & \gamma & \beta & \alpha
\end{pmatrix}
$$

with $\alpha + 2\beta + \gamma = 0$.

Clearly, the entire matrix is determined from the stiffness matrices of all elements. In particular, for a regular grid, we get the stencil

$$
\frac{8}{3}
\begin{bmatrix}
-1 & -1 & -1 \\
-1 & 8 & -1 \\
-1 & -1 & -1
\end{bmatrix}_* .
$$

Conversely, can the above matrix be computed from the stencil?

Because of the cyclic structure, the vectors

$$
(1, i^k, (-1)^k, (-i)^k), \quad k = 0, 1, 2, 3,
$$

are eigenvectors. Is it possible to find a corresponding set of real eigenvectors?
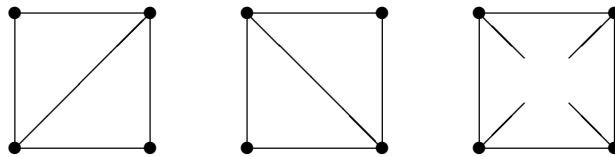

Fig. 27. a) Criss, b) Cross, and c) Criss-Cross Grids

8.3 Suppose for the model problem 4.3 that we combine two triangles in a square into a *macro-element*. Clearly, we get the same stiffness matrix as for the refinement shown in Figs. 25a and 25b. Now if we symmetrize the problem and take the function which is the average of the initial two, we get the so-called *criss-cross grid*; see Fig. 27c. Find the corresponding system matrix.

8.4 Consider the model problem, and compare the stiffness matrix A_Q in Problem 8.2 with those obtained for two standard triangles A_{2T}, and for the criss-cross grid A_{cc}. How large are the condition numbers of the matrices $A_Q^{-1} A_{2T}$, $A_Q^{-1} A_{cc}$, and $A_{cc}^{-1} A_{2T}$? In particular, show that A_{2T} is stiffer than A_Q, i.e., $A_{2T} - A_Q$ is positive semidefinite.

boundary

1/4 1/4   1/2   1
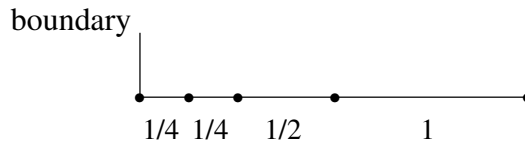
8.5 The above figure shows a line with a refinement, as could be found along a vertical grid line in Fig. 12. Extend this to a triangulation consisting of right isosceles triangles which connect to a coarse grid.

8.6 Suppose we want to solve the elliptic differential equation

$$\text{div}[a(x)\,\text{grad}\,u] = f \text{ in } \Omega$$

with suitable boundary conditions using linear triangular elements from $\mathcal{M}_0^1$. Show that we get the same solution if $a(x)$ is replaced by a function which is constant on each triangle. How can we find the right constants?

8.7 In $\mathbb{R}^2$ we can obviously decompose every triangle into four congruent subtriangles. With the help of a sketch, verify that the analogous assertion for a tetrahedron in $\mathbb{R}^3$ does not hold.

8.8 Let $\lambda_1, \lambda_2, \lambda_3$ be the barycentric coordinates of a triangle T with vertices z_1, z_2, z_3. Show that

$$p(z_i) = \frac{3}{\mu(T)} \int_T (3\lambda_i - \lambda_j - \lambda_k) p \, dx \quad \text{for } p \in \mathcal{P}_1,$$

if i, j, k is a permutation of 1, 2, 3.

8.9 The implementation of the Neumann boundary-value problem was elucidated for the case that the finite element space contains the kernel of the differential operator. What happens if that conditions is not satisfied?