Chapter III
Nonconforming and Other Methods

In the theory of conforming finite elements it is assumed that the finite element spaces lie in the function space in which the variational problem is posed. Moreover, we also require that the given bilinear form $a(\cdot, \cdot)$ can be computed exactly on the finite element spaces. However, these conditions are too restrictive for many real-life problems.

1. In general, homogeneous boundary conditions cannot be satisfied exactly for curved boundaries.
2. When we have variable coefficients or curved boundaries, we can only compute approximations to the integrals needed to assemble the stiffness matrix.
3. For plate problems and in general for fourth order elliptic differential equations, conforming methods require C^1 elements, and this leads to very large systems of equations.
4. We may want to enforce constraints only in the weak sense. A typical example is the Stokes problem, where the variational problem is posed in the space of divergence-free flows,
$$\{v \in H_0^1(\Omega)^n;\ (\operatorname{div} v, \lambda)_{0,\Omega} = 0 \quad \text{for all } \lambda \in L_2(\Omega)\}.$$
The constraint leads to saddle point problems, and we can only take into account finitely many of the infinitely many constraints.

In this chapter we show that these types of deviations from the theory of conforming elements are admissible and do not spoil convergence. In admitting them, we are committing what are called *variational crimes*.

In §1 we establish generalizations of Céa's lemma, and examine its use by looking at two applications. Then we give a short description of isoparametric elements. §§3 and 4 contain deep functional analytic methods which are of particular importance for the mixed methods of mechanics. We illustrate them in §§6 and 7 for the Stokes problem. §5 prepares the reader for nonstandard applications of saddle point problems.

§§8 and 9 will be concerned with a posteriori error estimates for finite element solutions. Here arguments from the theory of nonconforming elements and mixed methods enter even if we deal with conforming elements.

We should mention that the theory described in §3 has also recently been used to establish the convergence of difference methods and finite volume methods.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

§ 1. Abstract Lemmas and a Simple Boundary Approximation

If the finite element spaces being used to solve an H^m-elliptic problem do not lie in the Sobolev space $H^m(\Omega)$, we refer to them as *nonconforming elements*. In this case, convergence is by no means obvious. Moreover, in addition to the approximation error, there is now an error called the *consistency error*. To analyze the situation, we need certain generalizations of Céa's lemma. We shall apply these to a simple nonconforming element. We also show how they can be used when the conformity fails in a completely different way and the boundary conditions are relaxed.

Generalizations of Céa's Lemma

As usual, let $H_0^m(\Omega) \subset V \subset H^m(\Omega)$. We replace the given variational problem

$$a(u, v) = \langle \ell, v \rangle \quad \text{for all } v \in V \tag{1.1}$$

by a sequence of finite-dimensional problems: *Find $u_h \in S_h$ with*

$$a_h(u_h, v) = \langle \ell_h, v \rangle \quad \textit{for all } v \in S_h. \tag{1.2}$$

Here the bilinear forms a_h are assumed to be uniformly elliptic, i.e., there exists a constant $\alpha > 0$ independent of h such that

$$a_h(v, v) \ge \alpha \|v\|_{m,\Omega}^2 \quad \text{for all } v \in S_h. \tag{1.3}$$

Our error estimates for nonconforming methods are based on the following generalizations of Céa's lemma. For the first generalization, we do not require that a_h be defined for all functions in V. In particular, we permit the evaluation of a_h using quadrature formulas involving point evaluation functionals which are not defined for H^1 functions. However, we still require that $S_h \subset V$.

1.1 First Lemma of Strang. *Under the above hypotheses, there exists a constant c independent of h such that*

$$\|u - u_h\| \le c\Big(\inf_{v_h \in S_h} \Big\{ \|u - v_h\| + \sup_{w_h \in S_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|} \Big\} + \sup_{w_h \in S_h} \Big\{ \frac{\langle \ell, w_h \rangle - \langle \ell_h, w_h \rangle}{\|w_h\|} \Big\}\Big).$$

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

Proof. Let $v_h \in S_h$. For convenience, set $u_h - v_h = w_h$. Then by the uniform continuity and (1.2)–(1.3), we have

$$\begin{aligned}\alpha\|u_h - v_h\|^2 &\le a_h(u_h - v_h, u_h - v_h) = a_h(u_h - v_h, w_h)\\ &= a(u - v_h, w_h) + [a(v_h, w_h) - a_h(v_h, w_h)] + [a_h(u_h, w_h) - a(u, w_h)]\\ &= a(u - v_h, w_h) + [a(v_h, w_h) - a_h(v_h, w_h)]\\ &\quad - [\langle \ell, w_h\rangle - \langle \ell_h, w_h\rangle].\end{aligned}$$

Dividing through by $\|u_h - v_h\| = \|w_h\|$ and using the continuity of a, we get

$$\|u_h - v_h\| \le C\Big(\|u - v_h\| + \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|} + \frac{|\langle \ell_h, w_h\rangle - \langle \ell, w_h\rangle|}{\|w_h\|}\Big).$$

Since v_h is an arbitrary element in S_h, the assertion follows from the triangle inequality

$$\|u - u_h\| \le \|u - v_h\| + \|u_h - v_h\|.$$

□

Dropping the conformity condition $S_h \subset V$ has several consequences. In particular, the H^m-norm might not be defined for all elements in S_h, and we have to use mesh-dependent norms $\|\cdot\|_h$ as discussed, e.g., in II.6.1.

We assume that the bilinear forms a_h are defined for functions in V and in S_h, and that we have ellipticity and continuity:

$$\begin{aligned}a_h(v, v) &\ge \alpha\|v\|_h^2 && \text{for all } v \in S_h,\\ |a_h(u, v)| &\le C\|u\|_h\|v\|_h && \text{for all } u \in V + S_h,\ v \in S_h,\end{aligned} \tag{1.4}$$

with some positive constants α and C independent of h.

The following lemma is often denoted as the *second lemma of Strang.*

1.2 Lemma of Berger, Scott, and Strang. *Under the above hypotheses there exists a constant c independent of h such that*

$$\|u - u_h\|_h \le c\Big(\inf_{v_h \in S_h}\|u - v_h\|_h + \sup_{w_h \in S_h}\frac{|a_h(u, w_h) - \langle \ell_h, w_h\rangle|}{\|w_h\|_h}\Big).$$

Remark. The first term is called the *approximation error*, and the second one is called the *consistency error.*

Proof. Let $v_h \in S_h$. From (1.4) we see that

$$\begin{aligned}\alpha\|u_h - v_h\|_h^2 &\le a_h(u_h - v_h, u_h - v_h)\\ &= a_h(u - v_h, u_h - v_h) + [\langle \ell_h, u_h - v_h\rangle - a_h(u, u_h - v_h)].\end{aligned}$$

Dividing by $\|u_h - v_h\|_h$ and replacing $u_h - v_h$ by w_h, we have

$$\|u_h - v_h\|_h \le \alpha^{-1}\Big(C\|u - v_h\|_h + \frac{|a_h(u, w_h) - \langle \ell_h, w_h\rangle|}{\|w_h\|_h}\Big).$$

The assertion now follows from the triangle inequality as in the proof of the first lemma. □

1.3 Remark. Using a variant of the second lemma of Strang, we no longer need the requirement that the bilinear form a_h be defined on V. The formal extension of S_h to $S_h + V$ contains pitfalls, but according to the proof, it suffices to estimate the linear form

$$a_h(v_h, w_h) - \langle \ell_h, w_h \rangle \quad \text{for all } w_h \in S_h \tag{1.5}$$

for elements $v_h \in S_h$ whose distance from u is small. Indeed, in view of (1.2), this form coincides with $a_h(v_h - u_h, w_h)$. To evaluate (1.5), we can insert a term which can be interpreted as $a_h(u, w_h)$. The advantage is that this can be done with an individually chosen function.

Duality Methods

In using duality methods in the context of nonconforming elements, we get two additional terms as compared with the Aubin–Nitsche lemma.

1.4 Lemma. *Suppose that the Hilbert spaces V and H satisfy the hypotheses of the Aubin–Nitsche lemma. In addition, suppose $S_h \subset H$ and that the bilinear form a_h is defined on $V \cup S_h$ so that it coincides with a on V. Then the finite element solution u_h of (1.2) satisfies*

$$\begin{aligned} |u - u_h| \le \sup_{g \in H} \frac{1}{|g|} \Big\{ & c\|u - u_h\|_h \|\varphi_g - \varphi_{g,h}\|_h \\ & + |a_h(u - u_h, \varphi_g) - (u - u_h, g)| \\ & + |a_h(u, \varphi_g - \varphi_{g,h}) - \langle \ell, \varphi_g - \varphi_{g,h} \rangle| \Big\}. \end{aligned} \tag{1.6}$$

Here $\varphi_g \in V$ and $\varphi_{g,h} \in S_h$ are the weak solutions of $a_h(w, \varphi) = (w, g)$ for given $g \in H$.

Proof. By the definition of u_h, φ_g and $\varphi_{g,h}$, for every $g \in H$ we have

$$\begin{aligned} (u - u_h, g) &= a_h(u, \varphi_g) - a_h(u_h, \varphi_{g,h}) \\ &= a_h(u - u_h, \varphi_g - \varphi_{g,h}) \\ &\quad + a_h(u_h, \varphi_g - \varphi_{g,h}) + a_h(u - u_h, \varphi_{g,h}) \\ &= a_h(u - u_h, \varphi_g - \varphi_{g,h}) \\ &\quad - [a_h(u - u_h, \varphi_g) - (u - u_h, g)] \\ &\quad - [a_h(u, \varphi_g - \varphi_{g,h}) - \langle \ell, \varphi_g - \varphi_{g,h} \rangle]. \end{aligned}$$

The last equality is most easily verified by replacing the linear functionals in the square brackets by terms involving the bilinear form a_h, and then comparing terms. The assertion now follows from (II.7.7) and the continuity of a_h. □

The extra terms in (1.6) are basically of the same form as those in the second lemma of Strang. We shall see that in applications, the main effort is to verify that the hypotheses of the lemma hold.

The Crouzeix–Raviart Element

The Crouzeix–Raviart element is the simplest nonconforming element for the discretization of second order elliptic boundary-value problems. It is also called the *nonconforming P_1 element.*

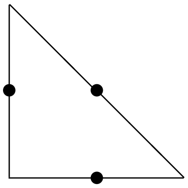


Fig. 28. The Crouzeix–Raviart element or nonconforming P_1-element

$$\begin{aligned}\mathcal{M}^1_* &:= \{v \in L_2(\Omega);\ v|_T \text{ is linear for every } T \in \mathcal{T}_h,\\ &\qquad\quad v \text{ is continuous at the midpoints of the triangle edges}\}, \\ \mathcal{M}^1_{*,0} &:= \{v \in \mathcal{M}^1_*;\ v = 0 \text{ at the midpoints of the edges on } \partial\Omega\}.\end{aligned} \tag{1.7}$$

To solve the Poisson equation, let

$$\begin{aligned} a_h(u, v) &:= \sum_{T\in\mathcal{T}_h} \int_T \nabla u \cdot \nabla v\, dx &&\text{for all } u, v \in H^1(\Omega) + \mathcal{M}^1_{*,0},\\ \|v\|_h &:= \sqrt{a_h(v, v)} &&\text{for all } v \in H^1(\Omega) + \mathcal{M}^1_{*,0}.\end{aligned}$$

By definition $\|v\|_h^2 := \sum_{T\in\mathcal{T}_h} |v|^2_{1,T}$, and it is called a *broken H^1 semi-norm.*

For simplicity, suppose Ω is a convex polyhedron. Then the problem is H^2-regular, and $u \in H^2(\Omega)$.

Given $v \in H^2(\Omega)$, let $Iv \in \mathcal{M}^1_{*,0} \cap C^0(\Omega)$ be the continuous piecewise linear function which interpolates v at the vertices of the triangles. We denote edges of the triangles by the letter e.

To apply Lemma 1.2, we compute

$$\begin{aligned} L_u(w_h) :&= a_h(u, w_h) - \langle \ell, w_h\rangle \\ &= \sum_{T\in\mathcal{T}_h} \int \nabla u \nabla w_h\, dx - \int_\Omega f w_h\, dx \\ &= \sum_{T\in\mathcal{T}_h} \Big(\int_{\partial T} \partial_\nu u\, w_h ds - \int_T \Delta u\, w_h dx\Big) - \int_\Omega f w_h dx \\ &= \sum_{T\in\mathcal{T}_h} \int_{\partial T} \partial_\nu u\, w_h ds,\end{aligned}$$

for $w_h \in \mathcal{M}^1_{*,0}$. Here we have used the fact that $-\Delta u = f$ holds in the weak sense; cf. Example II.2.10. In addition, note that each interior edge appears twice

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

in the sum. Thus, the values of the integrals do not change if we subtract the integral mean value $\overline{w_h(e)}$ on each edge e:

$$L_u(w_h) = \sum_T \sum_{e\subset\partial T} \int_e \partial_\nu u(w_h - \overline{w_h(e)})ds.$$

It follows from the definition of $\overline{w_h(e)}$ that $\int_e(w_h - \overline{w_h(e)})ds = 0$. The values of the integrals also do not change if we subtract an arbitrary constant function from $\partial_\nu u$ on each edge e. This can be $\partial_\nu Iu$ in particular, and we get

$$L_u(w_h) = \sum_T \sum_{e\subset\partial T} \int_e \partial_\nu(u - Iu)(w_h - \overline{w_h(e)})ds.$$

It follows from the Cauchy–Schwarz inequality that

$$|L_u(w_h)| \le \sum_T \sum_{e\subset\partial T} \Big[\int_e |\nabla(u - Iu)|^2 ds \int_e |w_h - \overline{w_h(e)}|^2 ds\Big]^{1/2}. \tag{1.8}$$

We now derive bounds for the integrals in (1.8). By the trace theorem and the Bramble–Hilbert lemma,

$$\int_{\partial T_{\rm ref}} |\nabla(v - Iv)|^2 ds \le c\|\nabla(v - Iv)\|^2_{1,T_{\rm ref}} \le c\|v - Iv\|^2_{2,T_{\rm ref}} \le c'|v|^2_{2,T_{\rm ref}},$$

for $v \in H^2(T_{\rm ref})$. Using the transformation formulas from Ch. II, §6, we see that

$$\int_{\partial T} |\nabla(v - Iv)|^2 ds \le ch|v|^2_{2,T} \tag{1.9}$$

for $T \in \mathcal{T}_h$. Similarly, for each edge e of $\partial T_{\rm ref}$,

$$\int_e |w_h - \overline{w_h(e)}|^2 ds \le c\|w_h\|^2_{1,T_{\rm ref}} \le c'|w_h|^2_{1,T_{\rm ref}} \quad \text{for all } w_h \in \mathcal{P}_1.$$

Here the Bramble–Hilbert lemma applies because the left-hand side vanishes for constant functions. For $e \subset T \in \mathcal{T}_h$, the transformation theorems yield

$$\int_e |w_h - \overline{w_h(e)}|^2 ds \le c\,h|w_h|^2_{1,T} \quad \text{for all } w_h \in \mathcal{M}^1_{*,0}. \tag{1.10}$$

We now insert the estimates (1.9) and (1.10) into (1.8), and use the Cauchy–Schwarz inequality for Euclidean scalar products:

$$\begin{aligned} |L_u(w_h)| &\le \sum_T 3\,ch|u|_{2,T}|w_h|_{1,T} \\ &\le c'h\Big[\sum_T |u|^2_{2,T} \sum_T |w_h|^2_{1,T}\Big]^{1/2} \\ &= c'h|u|_{2,\Omega}\|w_h\|_h\,. \end{aligned} \tag{1.11}$$

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

Finally, we observe that the conforming P_1 elements are contained in $\mathcal{M}^1_{*,0}$. This means that we do not need to establish a new approximation theorem for $\mathcal{M}^1_{*,0}$, and it follows that

$$\|u - u_h\|_h \le ch|u|_{2,\Omega} \le ch\|f\|_0. \tag{1.12}$$

We now apply Lemma 1.4 to the Crouzeix–Raviart element. Let $V = H^1(\Omega)$ and $H = L_2(\Omega)$. In particular, to estimate the first term, we regard $\varphi_g - \varphi_{g,h}$ as the discretization error for the problem $a(w, \varphi) = (w, g)_{0,\Omega}$. We make use of (1.12) and the regularity of the problem:

$$\|\varphi_g - \varphi_{g,h}\|_h \le ch|\varphi_g|_2 \le c''h\|g\|_0.$$

An essential observation is that the formula (1.11) holds for all $w \in \mathcal{M}^1_{*,0} + H^1_0$, as can be seen immediately by examining the derivation of the formula. It follows that the extra terms in (1.6) satisfy

$$\begin{aligned}
|a_h(u - u_h, \varphi_g) - (u - u_h, g)| &= |L_{\varphi_g}(u - u_h)| \\
&\le c'h\, |\varphi_g|_2\, \|u - u_h\|_h \\
&\le c'h\, \|g\|_0 \|u - u_h\|_h, \\
|a_h(u, \varphi_g - \varphi_{g,h}) - (f, \varphi_g - \varphi_{g,h})| &= |L_u(\varphi_g - \varphi_{g,h})| \\
&\le c'h\, |u|_2\, \|\varphi_g - \varphi_{g,h}\|_h.
\end{aligned}$$

Combining the last three estimates, we have

$$\begin{aligned}
|(u - u_h, g)| &\le c\,h(\|u - u_h\|_h + h|u|_2)\|g\|_0 \\
&\le c\,h^2|u|_2\, \|g\|_0.
\end{aligned}$$

Combining these duality calculations with (1.12), we obtain

1.5 Theorem. *Suppose Ω is convex or that it has a smooth boundary. Then using the Crouzeix–Raviart elements to discretize the Poisson equation, we have*

$$\|u - u_h\|_0 + h\|u - u_h\|_h \le c\,h^2|u|_2.$$

Remark. The result closely resembles the result in Ch. II, §7, but there is a difference. While for conforming methods the H^2-regularity was used only quantitatively, here it also enters qualitatively in the convergence proof. This corresponds with the practical observation that nonconforming elements are much more sensitive to “near singularities“ i.e., to the appearance of large H^2 norms.

A Simple Approximation to Curved Boundaries

We consider a second order differential equation on a domain Ω with smooth boundary. This means that for every point on $\Gamma = \partial\Omega$, there exist orthogonal coordinates (ξ, η) so that in a neighborhood, the boundary can be described as the graph of a C^2 function g. Suppose the domain Ω is decomposed into elements so that every element T has three vertices, at least one of which is an interior point of Ω. If two vertices of T lie on Γ, then the boundary piece of Γ with endpoints at these vertices is an edge of the element. Suppose all other edges of the elements are straight lines. We refer to these elements as *curved triangles*; see Fig. 29.

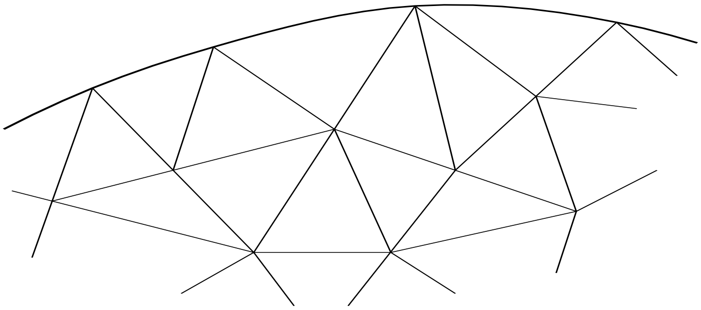


Fig. 29. Part of a triangulation of a domain with curved boundary

If we replace the boundary curves between two neighboring vertices by a line segment, we get a polygonal approximation Ω_h of Ω. The partition $\mathcal{T}_h$ of Ω induces a triangulation of Ω_h. We suppose that it is admissible. We call $\mathcal{T}_h$ *uniform* or *shape regular*, provided that the induced triangulation of Ω_h possesses the respective property.

We choose the finite elements to be the linear triangular elements, where the zero boundary conditions are enforced only at the nodes on Γ:

$$\begin{aligned} S_h := \{v \in C^0(\Omega);\ & v|_T \text{ is linear for every } T \in \mathcal{T}_h, \\ & v(z) = 0 \text{ for every node } z \in \Gamma\}. \end{aligned}$$

Thus, $S_h \not\subset H_0^1(\Omega)$. Nevertheless, since $S_h \subset H^1(\Omega)$, it is not necessary to work with a new (mesh-dependent) norm, and we can set $a_h(u, v) = a(u, v)$.

1.6 Lemma. *Let Ω be a domain with C^2 boundary, and let $\mathcal{T}_h$ be a sequence of shape-regular triangulations. Then*

$$\|v\|_{0,\Gamma} \le c\, h^{3/2}\, |v|_{1,\Omega} \quad \textit{for all } v \in S_h. \tag{1.13}$$

Proof. Let T be an element with a curved edge $\Gamma_1 := T \cap \Gamma$. We shall show that

$$\int_{\Gamma_1} v^2 ds \le c h_T^3 \int_T (|\partial_1 v|^2 + |\partial_2 v|^2) dx. \tag{1.14}$$

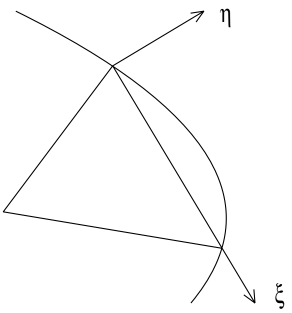


Fig. 30. Local coordinates for a curved element

Then the assertion follows after summing over all triangles of $\mathcal{T}_h$.

Suppose we choose the coordinate system so that the ξ-axis passes through the two vertices of T lying on Γ; see Fig. 30. Let $(\xi_1, 0)$, $(\xi_2, 0)$ be the coordinates of the vertices, and suppose the boundary is given by $\eta = g(\xi)$. From $g(\xi_1) = g(\xi_2) = 0$, $|\xi_1 - \xi_2| \le h_T$, and $|g''(\xi)| \le c$, it follows that

$$|g(\xi)| \le ch_T^2 \quad \text{for all } \xi_1 \le \xi \le \xi_2. \tag{1.15}$$

Since $v \in S_h$ is linear in T and vanishes at two points on the ξ-axis, v has the form

$$v(\xi, \eta) = b\eta$$

on T. The gradient is constant in T, $|\nabla u| = b$, and the area of T can be bounded from below by that of the largest inscribed circle. Its radius is at least h_T/κ, and thus

$$\int_T |\nabla v|^2 dx \ge \pi (h_T/\kappa)^2 b^2.$$

On the other hand,

$$\begin{aligned}\int_\Gamma v^2 ds &= \int_{\xi_1}^{\xi_2} [bg(\xi)]^2 \sqrt{1 + [g'(\xi)]^2}\, d\xi \\ &\le [bch_T^2]^2 \int_{\xi_1}^{\xi_2} 2\, d\xi \\ &= 2c^2 b^2 h_T^5.\end{aligned}$$

The assertion now follows by comparing the last two estimates. □

We remark that in view of (1.15), if we replace a piece of the curved boundary by a straight line, we cut off a domain $T'' := T \cap (\Omega \setminus \Omega_h)$ with an area

$$\mu(T'') \le ch\mu(T). \tag{1.16}$$

Now let u_h be the weak solution in S_h, i.e.

$$a(u_h, v) = (f, v)_{0,\Omega} \quad \text{for all } v \in S_h.$$

In addition, suppose $u \in H^2(\Omega) \cap H_0^1(\Omega)$ is the solution of the Dirichlet problem (II.2.7). Then $Lu = f$ in the sense of $L_2(\Omega)$, and integrating by parts, we have

$$\begin{aligned}(f, v)_{0,\Omega} &= (Lu, v)_{0,\Omega} \\ &= a(u, v) - \int_\Gamma \sum_{k,\ell} a_{k\ell}\, \partial_k u\, v \cdot \nu_\ell ds\end{aligned}$$

for $v \in S_h \subset H^1(\Omega)$. Applying the Cauchy–Schwarz inequality, the trace theorem, and the previous lemma, we get

$$\begin{aligned}|a(u, v) - (f, v)_{0,\Omega}| &\le c\|\nabla u\|_{0,\Gamma}\, \|v\|_{0,\Gamma} \\ &\le c\|u\|_{2,\Omega} h^{3/2} \|v\|_{1,\Omega}.\end{aligned}$$

The second lemma of Strang gives a term of order $h^{3/2}$, which is small in comparison with the usual term of order h.

1.7 Theorem. *Let Ω be a domain with C^2 boundary, and suppose we use linear triangular elements on shape-regular triangulations. Then the finite element approximation satisfies*

$$\begin{aligned}\|u - u_h\|_{1,\Omega} &\le c\, h\, \|u\|_{2,\Omega} \\ &\le c\, h\, \|f\|_{0,\Omega}.\end{aligned} \tag{1.17}$$

The estimate remains correct if we replace a by

$$a_h(u, v) := \int_{\Omega_h} \sum_{k,\ell} a_{k\ell}\, \partial_k u\, \partial_\ell v\, dx.$$

In particular, $|a_h(u, v) - a(u, v)| \le c\|u\|_{1,\Omega} \|v\|_{1,\Omega\backslash\Omega_h}$, and since ∇v is constant on every element T, (1.16) implies

$$\|v\|_{1,\Omega\backslash\Omega_h} \le ch\|v\|_{1,\Omega} \quad \text{for all } v \in S_h.$$

Modifications of the Duality Argument

The general Lemma 1.4 is not applicable here because the estimate (1.13) does not hold for all $v \in S_h + H_0^1$. For simplicity, we now apply the duality method along with the tools which we have already developed. Then even for L_2 estimates, we get an extra term of order $h^{3/2}$ which is no longer small compared with the main term. Using a more refined argument, this extra term could be avoided [Blum 1991].

In order to avoid having to work with a double sum in the boundary integrals, we restrict our attention to the Poisson equation, and note that the supremum in (II.7.7) is attained for $g = w$.

With $w := u - u_h$, let φ be the solution of equation (II.7.6); i.e., let

$$-\Delta\varphi = w \text{ in } \Omega,$$
$$\varphi = 0 \text{ on } \Gamma.$$

Since Ω is smooth, the problem is H^2-regular. Hence, $\varphi \in H^2(\Omega) \cap H_0^1(\Omega)$ and

$$\|\varphi\|_{2,\Omega} \le c\|w\|_{0,\Omega}.$$

Since $w \notin H_0^1(\Omega)$, in contrast to the calculations with conforming elements, we get boundary terms when applying Green's formula:

$$\begin{aligned}\|w\|_{0,\Omega}^2 &= (w, -\Delta\varphi)_{0,\Omega}\\ &= a(w,\varphi) - (w, \partial_\nu\varphi)_{0,\Gamma}\,.\end{aligned} \tag{1.18}$$

Let v_h be an arbitrary element in S_h. Then $a(u - u_h, -v_h) = (\partial_\nu u, -v_h)_{0,\Gamma}$, and since $\varphi \in H_0^1(\Omega)$, the last term can be replaced by $(\partial_\nu u, \varphi - v_h)_{0,\Gamma}$. By (1.18),

$$\|w\|_{0,\Omega}^2 = a(w, \varphi - v_h) - (\partial_\nu u, \varphi - v_h)_{0,\Gamma} - (w, \partial_\nu\varphi)_{0,\Gamma}\,. \tag{1.19}$$

Now we select v_h to be the interpolant of φ in S_h.

We estimate the first term in the same way as for conforming elements:

$$\begin{aligned}a(w, \varphi - v_h) &\le C\|w\|_{1,\Omega}\,\|\varphi - v_h\|_{1,\Omega}\\ &\le C\|w\|_{1,\Omega}\,ch\|\varphi\|_{2,\Omega}\\ &\le ch\|w\|_{1,\Omega}\,\|w\|_{0,\Omega}.\end{aligned}$$

To deal with the second term in (1.19), we need the estimate $\|\varphi - I\varphi\|_{0,\Gamma} \le ch^{3/2}\|\varphi\|_{2,\Omega}$, whose proof (which we do not present here) is based on a scaling argument. Now we apply the trace theorem to ∇u:

$$\begin{aligned}|(\partial_\nu u, \varphi - v_h)_{0,\Gamma}| &\le \|\nabla u\|_{0,\Gamma}\,\|\varphi - v_h\|_{0,\Gamma}\\ &\le c\|u\|_{2,\Omega}\,ch^{3/2}\|\varphi\|_{2,\Omega}\\ &\le ch^{3/2}\|u\|_{2,\Omega}\,\|w\|_{0,\Omega}.\end{aligned}$$

Next we apply Lemma 1.6 and the trace theorem to the last term to get

$$\begin{aligned}|(w, \partial_\nu\varphi)_{0,\Gamma}| &\le \|w\|_{0,\Gamma}\,\|\nabla\varphi\|_{0,\Gamma}\\ &\le \|u_h\|_{0,\Gamma}\,\|\nabla\varphi\|_{0,\Gamma}\\ &\le ch^{3/2}\|u_h\|_{1,\Omega}\,\|\varphi\|_{2,\Omega}\\ &\le ch^{3/2}(\|u\|_{1,\Omega} + \|u - u_h\|_{1,\Omega})\,\|w\|_{0,\Omega}\\ &\le ch^{3/2}\|u\|_{2,\Omega}\,\|w\|_{0,\Omega}.\end{aligned}$$

Combining the above, we have

$$\|w\|_{0,\Omega}^2 \le c\|w\|_{0,\Omega}\{h\|w\|_{1,\Omega} + h^{3/2}\|u\|_{2,\Omega}\}.$$

Recalling that $w = u - u_h$, we have

1.8 Theorem. *Under the hypotheses of Theorem 1.7,*

$$\|u - u_h\|_{0,\Omega} \le ch^{3/2}\|u\|_{2,\Omega}.$$

The error term $\mathcal{O}(h^{3/2})$ arises from the pointwise estimate of the finite-element functions $|u_h(x)| \le ch^2|\nabla u_h(x)|$ for all $x \in \Gamma$; cf. (1.15). If we approximate the boundary with quadratic (instead of linear) functions, giving a one higher power of h, the final result is improved by the same factor. This can be achieved using isoparametric elements, for example.

Problems

1.9 Let S_h be an affine family of C^0 elements. Show that in both the approximation and inverse estimates, $\|\cdot\|_{2,h}$ can be replaced by the mesh-dependent norm

$$|||v|||_h^2 := \sum_{T_j} \|v\|_{2,T_j}^2 + h^{-1} \sum_{\{e_m\}} \int_{e_m} \left\lceil \frac{\partial v}{\partial \nu} \right\rceil^2 ds.$$

Here $\{e_m\}$ is the set of inter-element boundaries, and $\lceil\cdot\rceil$ denotes the jump of a function.
Hint: In $H^2(T_{\text{ref}})$, $\|v\|_{2,T_{\text{ref}}}$ and $\left(\|v\|_{2,T_{\text{ref}}}^2 + \int_{\partial T_{\text{ref}}} |\nabla v|^2 ds\right)^{1/2}$ are equivalent norms.

1.10 The linear functional L_u appearing in the analysis of the Crouzeix–Raviart element vanishes on the subset $H_0^1(\Omega)$ by the definition of weak solutions. What is wrong with the claim that L_u vanishes for all $w \in L_2(\Omega)$ because of the density of $H_0^1(\Omega)$ in $L_2(\Omega)$?

1.11 If the stiffness matrices are computed by using numerical quadrature, then only approximations a_h of the bilinear form are obtained. This holds also for conforming elements. Estimate the influence on the error of the finite element solution, given the estimate

$$|a(v, v) - a_h(v, v)| \le \varepsilon(h)\ \|v\|_1^2 \quad \text{for all } v \in S_h.$$

1.12 The Crouzeix–Raviart element has locally the same degrees of freedom as the conforming P_1 element $\mathcal{M}_0^1$, i.e., the Courant triangle. Show that the (global) dimension of the finite element spaces differ by a factor that is close to 3 if a rectangular domain as in Fig. 9 is partitioned.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

§ 2. Isoparametric Elements

For the treatment of second order elliptic problems on domains with curved boundaries, we need to use elements with curved sides if we want to get higher accuracy. For many problems of fourth order, we even have to do a good job of approximating the boundary in the C^1-norm just to get convergence. For this reason, certain so-called *isoparametric families* of finite elements were developed. They are a generalization of the *affine* families.

For triangulations, isoparametric elements actually play a role only near the boundary. On the other hand, (simple) isoparametric quadrilaterals are often used in the interior since this allows us to generate arbitrary quadrilaterals, rather than just parallelograms.

We restrict our attention to planar domains, and consider families of elements where every $T \in \mathcal{T}_h$ is generated by a bijective mapping F:

$$\begin{aligned} T_{\text{ref}} &\longrightarrow T \\ (\xi, \eta) &\longmapsto (x, y) = F(\xi, \eta) = (p(\xi, \eta), q(\xi, \eta)). \end{aligned} \tag{2.1}$$

This framework includes the affine families when p and q are required to be linear functions. When p and q are polynomials of higher degree, we get the more general situation of isoparametric elements. More precisely, the polynomials in the parametrization are chosen from the same family Π as the shape functions of the element (T, Π, Σ).

Isoparametric Triangular Elements

The important case where p and q are quadratic polynomials is shown in Fig. 31. By Remark II.5.4, we know that six points P_i, $1 \le i \le 6$, can be prescribed. Then p and q as polynomials of degree 2 are uniquely defined by the coordinates of the points $P_1, \ldots, P_6$. In particular, if P_4, P_5, and P_6 are nodes at the midpoints of the edges of the triangle whose vertices are P_1, P_2, and P_3, then obviously we get a linear mapping.

The introduction of isoparametric elements raises the following questions:

1. Can isoparametric elements be combined with affine ones without losing the desired additional degrees of freedom?
2. How are the concepts "uniform" and "shape regular" to be understood so that the results for affine families can be carried over to isoparametric ones?

In order to keep the computational costs down, we should use elements with straight edges in the interior of Ω. This is why elements with only one curved side

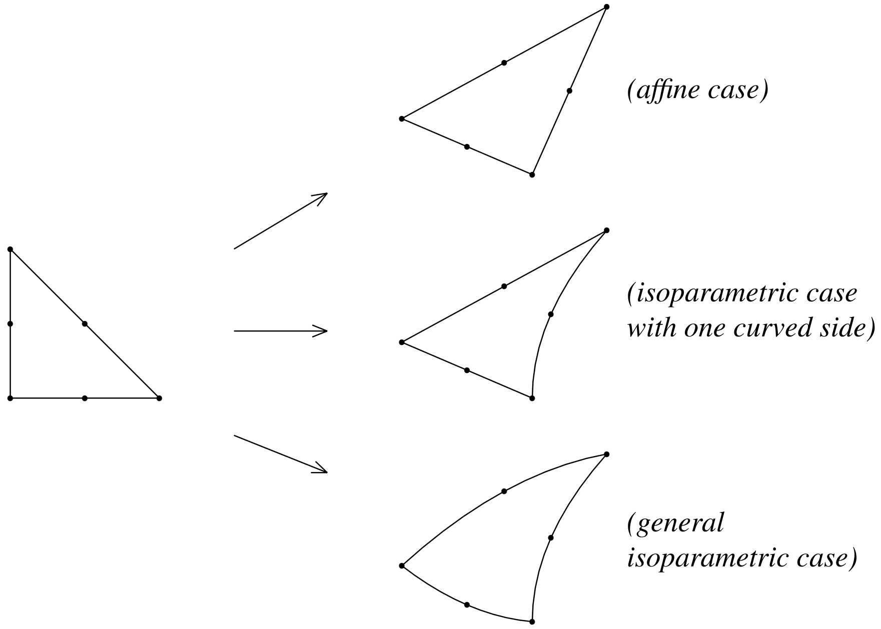

(affine case)
(isoparametric case with one curved side)
(general isoparametric case)

Fig. 31. Isoparametric elements with linear and quadratic parametrization

are of special interest. Suppose the edges of T_{ref} with $\xi = 0$ and with $\eta = 0$ are mapped to the straight edges of T. It is useful to choose the centers of these edges as the images of the corresponding center points on the edges of the reference triangle. Then in the quadratic case we have

$$\begin{aligned} p(\xi, \eta) &= a_1 + a_2\xi + a_3\eta + a_4\xi\eta, \\ q(\xi, \eta) &= b_1 + b_2\xi + b_3\eta + b_4\xi\eta. \end{aligned} \tag{2.2}$$

The restrictions of p and q to the edges $\xi = 0$ and $\eta = 0$ are linear functions, which results in a continuous match with neighboring affine elements without taking any special measures.

For affine families, the condition of shape regularity can be formulated in various ways, and a number of equivalent definitions can be found in the literature. The corresponding conditions for isoparametric elements are not completely independent, and cannot be replaced by *one* simple condition.

2.1 Definition. A family of isoparametric partitions $\mathcal{T}_h$ is called *shape regular* provided there exists a constant κ such that:

(i) For every parametrization $F : T_{\text{ref}} \longrightarrow T \in \mathcal{T}_h$,

$$\frac{\sup\{\|DF(\zeta)\cdot z\|;\ \zeta \in T_{\text{ref}},\ \|z\| = 1\}}{\inf\{\|DF(\zeta)\cdot z\|;\ \zeta \in T_{\text{ref}},\ \|z\| = 1\}} \le \kappa.$$

(ii) For every $T \in \mathcal{T}_h$, there exists an inscribed circle with radius ρ_T such that

$$diameter(T) \le \kappa\rho_T.$$

If in addition

$$diameter(T) \le 2h \quad \text{and} \quad \rho_T \ge h/\kappa,$$

then $\mathcal{T}_h$ is called *uniform.*

Isoparametric Quadrilateral Elements

Isoparametric quadrilaterals are also of use in the interior since only parallelograms can be obtained from a square with affine mappings (see Ch. II, §5).

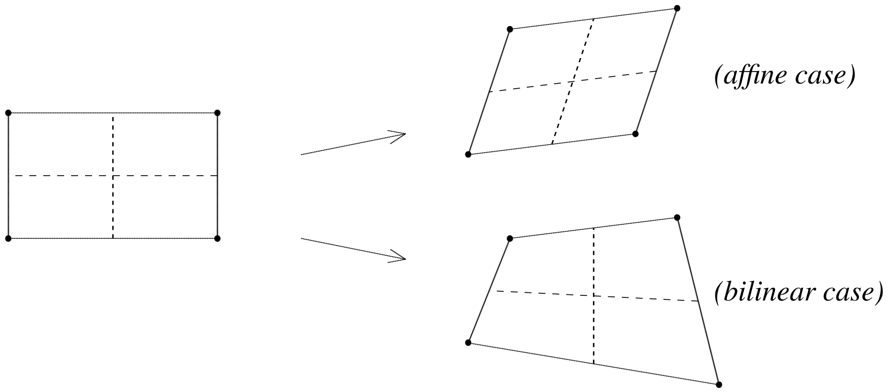


Fig. 32. Isoparametric quadrilaterals with bilinear parametrization

Let $T_{\text{ref}} = [0, 1]^2$ be the unit square. Then

$$F : \left\{ \begin{aligned} p(\xi, \eta) &= a_1 + a_2\xi + a_3\eta + a_4\xi\eta \\ q(\xi, \eta) &= b_1 + b_2\xi + b_3\eta + b_4\xi\eta \end{aligned} \right\} \tag{2.3}$$

maps T_{ref} to a general quadrilateral. From the theory of bilinear quadrilateral elements, we know that the two sets of four parameters are uniquely determined by the eight coordinates of the four corners of the image of T_{ref}.

In addition, it is clear that when ξ and η are both constant, the parametrization F is a linear function of the arc length. It follows that the image is a quadrilateral with straight edges. The vertices are numbered so that the orientation is preserved. Because of the linearity of the parametrization on the edges, connecting the element to its neighbors is no problem.

2.2 Remark. A family of partitions $\mathcal{T}_h$ involving general quadrilaterals with bilinear parametrizations is *shape regular* provided there exists a constant $\kappa > 1$ such that the following conditions are satisfied:

(i) For every quadrilateral T, the ratio of maximal to minimal edge lengths is bounded by κ.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

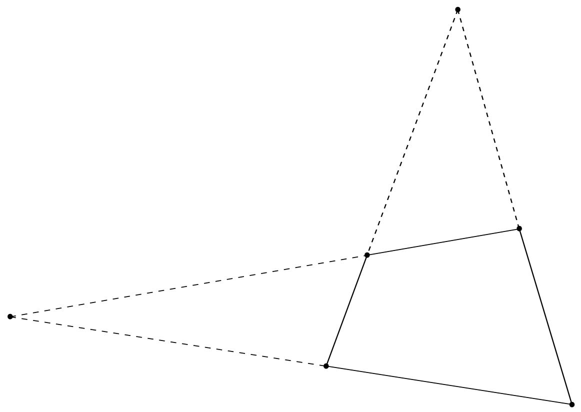


Fig. 33. General quadrilateral

(ii) Every T contains an inscribed circle with radius $\rho_T \geq h_T/\kappa$, where h_T is the diameter of T.

(iii) All angles are smaller than $\pi - \varphi_0$ with some $\varphi_0 > 0$. [We note that the second condition implies that all angles are greater than φ_0 with some $\varphi_0 > 0$.]

Moreover, we note that DF and also $\det(DF)$ depend linearly on the parameters. In particular, the determinants attain their maximum and minimum values at vertices of the quadrilateral. For the quadrilateral shown in Fig. 33, P_2 and P_4 are the extremal points since the intersections of the sides ly on their extension through P_4.

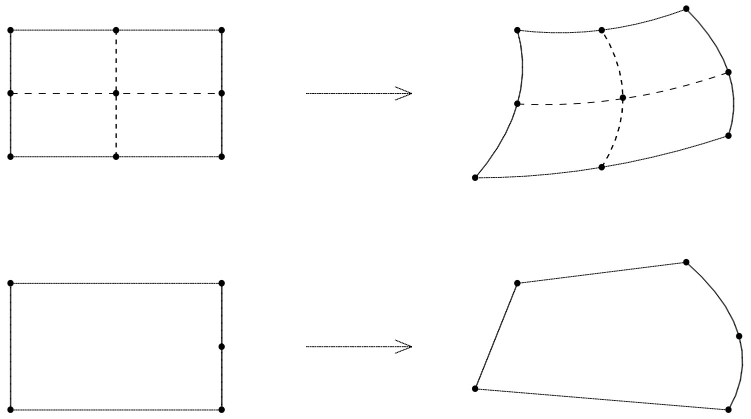


Fig. 34. Isoparametric quadrilateral elements with 9 and 5 parameters

Fig. 34 shows curved quadrilaterals which arise from biquadratic parametrizations. The parametrization for 9 prescribed points corresponds exactly with the 9 node element in Ch. II, §5. The 5 point case is of particular practical importance

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

since it can model one curved side. Here the correct parametrization is

$$\begin{aligned} x &= a_1 + a_2\xi + a_3\eta + a_4\xi\eta + a_5\xi\eta(1-\eta), \\ y &= b_1 + b_2\xi + b_3\eta + b_4\xi\eta + b_5\xi\eta(1-\eta). \end{aligned}$$

At first glance, we are tempted to replace the shape functions corresponding to the coefficients a_5 and b_5 by the simpler (quadratic) expressions $a_5\eta(1-\eta)$ and $b_5\eta(1-\eta)$. This would be possible for interpolation at the 5 points, but would not lead to a linear expression on the edge $\xi = 0$.

Problems

2.3 Suppose we have a program for generating quadrilateral elements, and now want to use it to build triangular elements. We map quadrilaterals (bilinearly) to triangles by identifying pairs of vertices in the image. What triangular elements do we get using the bilinear, 8 point, and 9 point elements, respectively?

2.4 Suppose that in setting up the system matrix we use a quadrature formula with positive weights. Show that in spite of the error in the numerical integration, the matrix is at least positive semidefinite. Describe a case in which the matrix is singular.

§ 3. Further Tools from Functional Analysis

In Ch. II the existence of solutions to the variational problem was established using the Lax–Milgram theorem. There the symmetry and the ellipticity of the bilinear form $a(\cdot, \cdot)$ were essential hypotheses. In order to treat saddle point problems, we need a more general approach which does not require that the quadratic form be positive definite.

The relation of the functional ℓ to the right-hand side f of the differential equation was discussed only briefly in Remark II.4.1. A more advanced consideration of linear functionals on Sobolev spaces is appropriate here. It leads us to the so-called *negative norms*.

Negative Norms

Let V be a Hilbert space. By the Riesz representation theorem, every continuous linear functional $\ell \in V'$ can be identified with an element from V itself. Thus, often it is not necessary to distinguish between V and V'. However, in the variational calculus, this can obscure certain important aspects of the problem. Before discussing methods of functional analysis, we first orient ourselves with a simple example.

Consider the *Helmholtz equation*

$$\begin{aligned} -\Delta u + u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{3.1}$$

with $f \in L_2(\Omega)$. The weak solution $u \in H_0^1(\Omega)$ is characterized by

$$(u, v)_1 = (f, v)_0 \quad \text{for all } v \in H_0^1(\Omega), \tag{3.2}$$

where $(\cdot, \cdot)_1$ is the scalar product on the Hilbert space $H^1(\Omega)$. The problem (3.1) can thus be formulated as follows: *Given f, find the Riesz representation for the functional*

$$\ell : H_0^1(\Omega) \longrightarrow \mathbb{R}, \quad \langle \ell, v \rangle := (f, v)_0. \tag{3.3}$$

If we identify $H_0^1(\Omega)$ with its dual, then there is nothing to do to solve the variational problem. (Note that the analogous representation for the space $H^0(\Omega) = L_2(\Omega)$ is indeed trivial.)

There is another formulation for the dual space which fits better with the form (3.3) of the given functional. The equation (3.1) is then defined for all functions f in the associated completion of $L_2(\Omega)$.

3.1 Definition. Let $m \geq 1$. Given $u \in L_2(\Omega)$, define the norm

$$\|u\|_{-m,\Omega} := \sup_{v \in H_0^m(\Omega)} \frac{(u, v)_{0,\Omega}}{\|v\|_{m,\Omega}}.$$

We define $H^{-m}(\Omega)$ to be the completion of $L_2(\Omega)$ w.r.t. $\|\cdot\|_{-m,\Omega}$.

For the Sobolev spaces built on $L^2(\Omega)$, we identify the dual space of $H_0^m(\Omega)$ with $H^{-m}(\Omega)$. Moreover, by the definition of H^{-m}, there is a dual pairing $\langle u, v\rangle$ for all $u \in H^{-m}$, $v \in H_0^m$, i.e. $\langle u, v\rangle$ is a bilinear form, and

$$\langle u, v\rangle = (u, v)_{0,\Omega}, \quad \text{whenever } u \in L_2(\Omega),\ v \in H_0^m(\Omega).$$

Clearly,

$$\ldots \supset H^{-2}(\Omega) \supset H^{-1}(\Omega) \supset L_2(\Omega) \supset H_0^1(\Omega) \supset H_0^2(\Omega) \supset \ldots$$
$$\cdots \leq \|u\|_{-2,\Omega} \leq \|u\|_{-1,\Omega} \leq \|u\|_{0,\Omega} \leq \|u\|_{1,\Omega} \leq \|u\|_{2,\Omega} \leq \cdots$$

H^{-m} was defined to be the dual space of H_0^m and not of H^m. Thus we obtain an improvement of II.2.9 only for Dirichlet problems.

3.2 Remark. Let a be an H_0^m-elliptic bilinear form. Then with the notation of the proof of the Existence Theorem II.2.9, we have

$$\|u\|_m \leq \alpha^{-1}\|f\|_{-m}. \tag{3.4}$$

Proof. By Definition 3.1, $(u, v)_0 \leq \|u\|_{-m}\|v\|_m$. Substituting $v = u$ in the weak equation gives

$$\alpha\|u\|_m^2 \leq a(u, u) = (f, u)_0 \leq \|f\|_{-m}\|u\|_m,$$

and the assertion follows after dividing by $\|u\|_m$. □

This asserts that the Dirichlet problem is H^m-regular in the sense of Definition II.7.1.

3.3 Remark. Let $V \subset U$ be Hilbert spaces, and suppose the imbedding $V \hookrightarrow U$ is continuous and dense. In addition, suppose we identify U' with U via the Riesz representation. Then

$$V \subset U \subset V'$$

is called a *Gelfand triple*. We have already encountered the following Gelfand triples:

$$H^m(\Omega) \subset L_2(\Omega) \subset H^m(\Omega)',$$

and

$$H_0^m(\Omega) \subset L_2(\Omega) \subset H^{-m}(\Omega).$$

Adjoint Operators

Let X and Y be Banach spaces whose dual spaces are X' and Y', respectively. The dual pairings will usually be written as $\langle\cdot,\cdot\rangle$ without reference to the spaces. Let $L : X \longrightarrow Y$ be a bounded linear operator. Given $y^* \in Y'$,

$$x \longmapsto \ell_{y^*}(x) := \langle y^*, Lx\rangle$$

defines a continuous linear functional on X. The associated linear mapping

$$\begin{aligned} L' : Y' &\longrightarrow X', \\ y^* &\longmapsto \ell_{y^*}, \quad \text{i.e. } \langle L'y^*, x\rangle := \langle y^*, Lx\rangle, \end{aligned}$$

is called the *adjoint* of L.

Often the adjoint operator can be used to determine the image of L. More generally, let V be a closed subspace of X. Then

$$V^0 := \{\ell \in X';\ \langle \ell, v\rangle = 0 \quad \text{for all } v \in V\}$$

is called the *polar set* of V. Since in the Hilbert space case we cannot always identify the dual space X' with X, we must distinguish between the polar set V^0 and the *orthogonal complement*

$$V^\perp = \{x \in X;\ (x, v) = 0 \quad \text{for all } v \in V\}.$$

In the sequel we shall make multiple use of the following *closed range theorem* (see, e.g., Yosida [1971]). We give a proof at the end of this section.

3.4 Theorem. *Let X and Y be Banach spaces, and let $L : X \longrightarrow Y$ be a bounded linear mapping. Then the following assertions are equivalent:*

(i) The image $L(X)$ is closed in Y,

(ii) $L(X) = (\ker L')^0$.

An Abstract Existence Theorem

Let U and V be Hilbert spaces, and suppose $a : U \times V \longrightarrow \mathbb{R}$ is a bilinear form. We define an associated linear operator $L : U \longrightarrow V'$ by

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

The variational problems discussed above had the following structure: Given $f \in V'$, find $u \in U$ so that

$$a(u, v) = \langle f, v\rangle \quad \text{for all } v \in V. \tag{3.5}$$

We can formally write $u = L^{-1} f$.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

3.5 Definition. Let U and V be normed linear spaces. A linear mapping L is an *isomorphism* if it is bijective and L and L^{-1} are continuous.

The importance of the following theorem for the finite element theory was pointed out by Babuška [1971]; see also Babuška and Aziz [1972]. It can be traced back to Nečas [1962] and Nirenberg; cf. Babuška [1971].

3.6 Theorem. *Let U and V be Hilbert spaces. Then a linear mapping $L : U \longrightarrow V'$ is an isomorphism if and only if the associated form $a : U \times V \longrightarrow \mathbb{R}$ satisfies the following conditions:*
(ii) (Continuity). There exists $C \geq 0$ such that

$$|a(u, v)| \leq C \|u\|_U \|v\|_V . \tag{3.6}$$

(ii) (Inf-sup condition). There exists $\alpha > 0$ such that

$$\sup_{v \in V} \frac{a(u, v)}{\|v\|_V} \geq \alpha \|u\|_U \quad \textit{for all } u \in U. \tag{3.7}$$

(iii) For every $v \in V$, there exists $u \in U$ with

$$a(u, v) \neq 0. \tag{3.8}$$

Supplement. If we assume only (i) and (ii), then

$$L : U \longrightarrow \{v \in V;\ a(u, v) = 0 \ \text{ for all } u \in U\}^0 \subset V' \tag{3.9}$$

is an isomorphism. Moreover, (3.7) is equivalent to

$$\|Lu\|_{V'} \geq \alpha \|u\|_U \quad \text{for all } u \in U. \tag{3.10}$$

The name for condition (3.7) comes from the equivalent formulation

$$\inf_{u \in U} \sup_{v \in V} \frac{a(u, v)}{\|u\|_U \|v\|_V} \geq \alpha > 0. \tag{3.7$'$}$$

Proof of Theorem 3.6. The equivalence of the continuity of $L : U \longrightarrow V'$ with (3.6) follows from a simple calculation.

From (3.7) we immediately deduce that L is injective. Suppose $Lu_1 = Lu_2$. Then by definition, $a(u_1, v) = a(u_2, v)$ for all $v \in V$. Thus, $\sup_v a(u_1 - u_2, v) = 0$, and (3.7) implies $u_1 - u_2 = 0$.

Given $f \in L(U)$, by the injectivity there exists a unique inverse $u = L^{-1} f$. We now apply (3.7) a second time:

$$\alpha \|u\|_U \leq \sup_{v \in V} \frac{a(u, v)}{\|v\|_V} = \sup_{v \in V} \frac{\langle f, v \rangle}{\|v\|_V} = \|f\|. \tag{3.11}$$

This is (3.10), and L^{-1} is continuous on the image of L.

The continuity of L and L^{-1} implies that $L(U)$ is closed. Now (3.9) follows from Theorem 3.4. This establishes the supplement to Theorem 3.6.

Finally, condition (iii) ensures the surjectivity of L. Indeed, by (3.9) $L(U)$ is the polar set of the null element, and so coincides with V'. Hence, the conditions (i), (ii), and (iii) are sufficient to ensure that $L : U \longrightarrow V'$ is an isomorphism.

In view of (3.11), the necessity of the conditions is immediate. □

The Lax–Milgram theorem (for linear spaces) follows as a special case. Indeed, if a is a continuous V-elliptic bilinear form, then the inf-sup condition follows from

$$\sup_v \frac{a(u,v)}{\|v\|} \ge \frac{a(u,u)}{\|u\|} \ge \alpha\|u\|.$$

In particular, the differential operator in (II.2.5) can be regarded as a bijective mapping $L : H_0^1(\Omega) \longrightarrow H^{-1}(\Omega)$. The converse follows from Problem 3.8. [However, the assertion that for H^2-regular problems, $L : H^2(\Omega) \cap H_0^1(\Omega) \longrightarrow H^0(\Omega)$ is also an isomorphism cannot be obtained in this framework.]

In the proof of Theorem 3.6 we used the closedness of the image of L. At first glance this appears to be just a technicality which allows the application of Theorem 3.4. However, the counterexample II.2.7 and Problem II.2.14 (see also Remark 6.5) show how important this point is.

An Abstract Convergence Theorem

To solve equation (3.5) numerically we are led naturally to a Galerkin method. Let $U_h \subset U$ and $V_h \subset V$ be finite-dimensional spaces. Then given $f \in V'$, we seek $u_h \in U_h$ such that

$$a(u_h, v) = \langle f, v\rangle \quad \text{for all } v \in V_h. \tag*{(3.5)$_h$}$$

In order to carry over Céa's lemma, we now require that the spaces U_h and V_h fit together.

3.7 Lemma. *Suppose the bilinear form $a : U \times V \longrightarrow \mathbb{R}$ satisfies the hypotheses of Theorem 3.6. Suppose the subspaces $U_h \subset U$ and $V_h \subset V$ are chosen so that (3.7′) and (3.8) also hold when U and V are replaced by U_h and V_h, respectively. Then*

$$\|u - u_h\| \le \left(1 + \frac{C}{\alpha}\right) \inf_{w_h \in U_h} \|u - w_h\|.$$

Remark. We say that the subspaces U_h and V_h satisfy the *Babuška condition* or an *inf-sup condition* provided (3.7′) holds for U_h and V_h as stated in Lemma 3.7.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

Proof of Lemma 3.7. By (3.5) and $(3.5)_h$,

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

Let w_h be an arbitrary element in U_h. Then

$$a(u_h-w_h,v)=a(u-w_h,v) \quad \text{for all } v\in V_h.$$

For $\langle \ell,v\rangle := a(u-w_h,v)$, we have $\|\ell\|\le C\|u-w_h\|$. By assumption, the mapping $L_h : U_h \longrightarrow V_h'$ generated by $a(u_h-w_h,\cdot)$ satisfies $\|(L_h)^{-1}\|\le 1/\alpha$. Thus,

$$\|u_h-w_h\|\le \alpha^{-1}\|\ell\|\le \alpha^{-1}C\|u-w_h\|.$$

The assertion follows from the triangle inequality $\|u-u_h\|\le\|u-w_h\|+\|u_h-w_h\|$. □

We mention that the theory described in this section has also recently been used to establish the convergence of difference methods and finite volume methods.

Proof of Theorem 3.4

For completeness we now prove Theorem 3.4.

It suffices to establish the identity

$$\overline{L(X)}=(\ker L')^0. \tag{3.12}$$

By the definition of the polar set and of the adjoint operator, we have

$$\begin{aligned}(\ker L')^0 &= \{y\in Y;\ \langle y^*,y\rangle=0 \quad \text{for } y^*\in\ker L'\}\\ &= \{y\in Y;\ \langle y^*,y\rangle=0 \quad \text{for } y^*\in Y' \text{ with } \langle L'y^*,x\rangle=0 \quad \text{for } x\in X\}\\ &= \{y\in Y;\ \langle y^*,y\rangle=0 \quad \text{for } y^*\in Y' \text{ with } \langle y^*,Lx\rangle=0 \quad \text{for } x\in X\}.\end{aligned} \tag{3.13}$$

Hence, $L(X)\subset(\ker L')^0$. Since the polar set is the intersection of closed sets, it is itself closed, and consequently so is $\overline{L(X)}\subset(\ker L')^0$.

Suppose that there exists $y_0\in(\ker L')^0$ with $y_0\notin\overline{L(X)}$. Then the distance of y_0 from $L(X)$ is positive, and there exist a small open sphere centered at the point y_0 which is disjoint from the convex set $L(X)$. By the separation theorem for convex sets, there exist a functional $y^*\in Y'$ and a number a with

$$\begin{aligned}&\langle y^*,y_0\rangle>a,\\ &\langle y^*,Lx\rangle\le a \quad \text{for all } x\in X.\end{aligned}$$

Since L is linear, this is possible only if $\langle y^*,Lx\rangle=0$ for all $x\in X$. Thus, $a>0$ and $\langle y^*,y_0\rangle\neq 0$. This would contradict (3.13), and so (3.12) must hold. □

Problems

3.8 Let $a : V \times V \to \mathbb{R}$ be a positive symmetric bilinear form satisfying the hypotheses of Theorem 3.6. Show that a is elliptic, i.e., $a(v, v) \geq \alpha_1 \|v\|_V^2$ for some $\alpha_1 > 0$.

3.9 [Nitsche, private communication] Show the following converse of Lemma 3.7: Suppose that for every $f \in V'$, the solution of (3.5) satisfies

$$\lim_{h \to 0} u_h = u := L^{-1} f.$$

Then

$$\inf_h \inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{a(u_h, v_h)}{\|u_h\|_U \|v_h\|_V} > 0.$$

Hint: Use (3.10) and apply the principle of uniform boundedness.

3.10 Show that

$$\begin{aligned} \|v\|_0^2 &\leq \|v\|_m \|v\|_{-m} && \text{for all } v \in H_0^m(\Omega), \\ \|v\|_1^2 &\leq \|v\|_0 \|v\|_2 && \text{for all } v \in H^2(\Omega) \cap H_0^1(\Omega). \end{aligned}$$

Hint: To prove the second relation, use the Helmholtz equation $-\Delta u + u = f$.

3.11 Let L be an H^1-elliptic differential operator. In which Sobolev spaces $H^s(\Omega)$ is the set

$$\{u \in H^1(\Omega);\ Lu = f \in L_2(\Omega),\ \|f\|_0 \leq 1\}$$

compact?

3.12 (Fredholm Alternative) Let H be a Hilbert space. Assume that the linear mapping $A : H \to H'$ can be decomposed in the form $A = A_0 + K$, where A_0 is H-elliptic, and K is compact. Show that either A satisfies the inf-sup condition, or there exists an element $x \in H$, $x \neq 0$, with $Ax = 0$.

3.13 Let $\Omega \subset \mathbb{R}^d$ and $p \in L_2(\Omega)$. Show that $\operatorname{grad} p \in H^{-1}(\Omega)$ and

$$\|\operatorname{grad} p\|_{-1,\Omega} \leq d\, \|p\|_{0,\Omega}\,. \tag{3.14}$$

Hint: Start with proving (3.14) for smooth functions and use Green's formula. Complete the proof by a density argument.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

§ 4. Saddle Point Problems

We now turn to variational problems with constraints. Let X and M be two Hilbert spaces, and suppose

$$a : X \times X \longrightarrow \mathbb{R}, \quad b : X \times M \longrightarrow \mathbb{R}$$

are continuous bilinear forms. Let $f \in X'$ and $g \in M'$. We denote both the dual pairing of X with X' and that of M with M' by $\langle \cdot, \cdot \rangle$. We consider the following minimum problem.

Problem (M). Find the minimum over X of

$$J(u) = \frac{1}{2} a(u, u) - \langle f, u \rangle \tag{4.1}$$

subject to the constraint

$$b(u, \mu) = \langle g, \mu \rangle \quad \text{for all } \mu \in M. \tag{4.2}$$

Saddle Points and Minima

Our starting point is the same as in the classical theory of Lagrange extremal problems. If $\lambda \in M$, then J and the *Lagrange function*

$$\mathcal{L}(u, \lambda) := J(u) + [b(u, \lambda) - \langle g, \lambda \rangle] \tag{4.3}$$

have the same values on the set of all points which satisfy the constraints. Instead of finding the minimum of J, we can seek a minimum of $\mathcal{L}(\cdot, \lambda)$ with fixed λ. This raises the question of whether $\lambda \in M$ can be selected so that the minimum of $\mathcal{L}(\cdot, \lambda)$ over the space X is assumed by an element which satisfies the given constraints. Since $\mathcal{L}(u, \lambda)$ contains only bilinear and quadratic expressions in u and λ, we are led to the following *saddle point problem*:

Problem (S). Find $(u, \lambda) \in X \times M$ with

$$\begin{aligned} a(u, v) + b(v, \lambda) &= \langle f, v \rangle \quad \text{for all } v \in X, \\ b(u, \mu) \qquad\qquad &= \langle g, \mu \rangle \quad \text{for all } \mu \in M. \end{aligned} \tag{4.4}$$

It is easy to see that every solution (u, λ) of Problem (S) must satisfy the *saddle point property*

$$\mathcal{L}(u, \mu) \leq \mathcal{L}(u, \lambda) \leq \mathcal{L}(v, \lambda) \quad \text{for all } (v, \mu) \in X \times M.$$

Here only the nonnegativity of $a(v, v) \geq 0$ is needed (cf. the Characterization Theorem II.2.2). The first component of a saddle point (u, λ) provides a solution of Problem (M).

The converse of this assertion is by no means obvious. Even if the minimum problem has a solution, we can ensure the existence of Lagrange multipliers only under additional hypotheses. We can see this already for a simple finite-dimensional example.

4.1 Example. Consider the following minimum problem in $\mathbb{R}^2$:

$$\begin{aligned} x^2 + y^2 &\longrightarrow \min! \\ x + y &= 2. \end{aligned}$$

Clearly, $x = y = 1$, $\lambda = -2$ provides a saddle point for the Lagrange function $\mathcal{L}(x, y, \lambda) = x^2 + y^2 + \lambda(x + y - 2)$, and $x = y = 1$ is a solution of Problem (M).

A formal doubling of the constraints clearly leads to a minimum problem with the same minimum:

$$\begin{aligned} x^2 + y^2 &\longrightarrow \min! \\ x + y &= 2, \\ 3x + 3y &= 6. \end{aligned}$$

However, the Lagrange multipliers for

$$\mathcal{L}(x, y, \lambda, \mu) = x^2 + y^2 + \lambda(x + y - 2) + \mu(3x + 3y - 6)$$

are no longer uniquely defined. Every combination with $\lambda + 3\mu = -2$ leads to a saddle point. Moreover, arbitrarily small perturbations of the data on the right-hand side can lead to a problem with no solution.

The inf-sup Condition

As we saw in Ch. II, §2, in infinite-dimensional spaces we have to correctly formulate the definiteness condition for the form a. The same holds for the constraints; it does not suffice to require their linear independence. An inf-sup condition provides the correct framework, similar to its appearance in Theorem 3.6. Equation (4.4) defines a linear mapping

$$\begin{aligned} L : X \times M &\longrightarrow X' \times M' \\ (u, \lambda) &\longmapsto (f, g). \end{aligned} \tag{4.5}$$

To show that L is an isomorphism we need the inf-sup condition (3.7). Brezzi [1974] has split the condition into properties of the two forms a and b. We introduce special notation for the affine space of admissible elements and for the corresponding linear spaces:

$$\begin{aligned} V(g) &:= \{v \in X;\ b(v,\mu) = \langle g, \mu\rangle \quad \text{for all } \mu \in M\}, \\ V &:= \{v \in X;\ b(v,\mu) = 0 \qquad\quad \text{for all } \mu \in M\}. \end{aligned} \tag{4.6}$$

Since b is continuous, V is a closed subspace of X.

It is often easier to handle the saddle point equation (4.4) if we reformulate it as an operator equation. To this end, we associate the mapping

$$\begin{aligned} A : X &\longrightarrow X', \\ \langle Au, v\rangle &= a(u,v) \quad \text{for all } v \in X, \end{aligned}$$

with the bilinear form a. Thus, the mapping A is defined by the action of the functional $Au \in X'$ on each $v \in X$. Similarly, we associate a mapping B and its adjoint mapping B' with the form b:

$$\begin{aligned} B : X &\longrightarrow M', & B' : M &\longrightarrow X', \\ \langle Bu, \mu\rangle &= b(u,\mu) \quad \text{for all } \mu \in M, & \langle B'\lambda, v\rangle &= b(v,\lambda) \quad \text{for all } v \in X. \end{aligned}$$

Then (4.4) is equivalent to

$$\begin{aligned} Au + B'\lambda &= f, \\ Bu \qquad &= g. \end{aligned} \tag{4.7}$$

4.2 Lemma. *The following assertions are equivalent:*

(i) There exists a constant $\beta > 0$ with

$$\inf_{\mu \in M} \sup_{v \in X} \frac{b(v,\mu)}{\|v\|\,\|\mu\|} \ge \beta. \tag{4.8}$$

(ii) The operator $B : V^\perp \longrightarrow M'$ is an isomorphism, and

$$\|Bv\| \ge \beta \|v\| \quad \textit{for all } v \in V^\perp. \tag{4.9}$$

(iii) The operator $B' : M \longrightarrow V^0 \subset X'$ is an isomorphism, and

$$\|B'\mu\| \ge \beta \|\mu\| \quad \textit{for all } \mu \in M. \tag{4.10}$$

Proof. The equivalence of (i) and (iii) is just the assertion of the supplement to Theorem 3.6.

Suppose condition (iii) is satisfied. Then for given $v \in V^\perp$, we define a functional $g \in V^0$ by $w \longmapsto (v, w)$. Since B' is an isomorphism, there exists $\lambda \in M$ with

$$b(w, \lambda) = (v, w) \quad \text{for all } w. \tag{4.11}$$

By the definition of the functional g, we have $\|g\| = \|v\|$, and (4.10) implies $\|v\| = \|g\| = \|B'\lambda\| \geq \beta\|\lambda\|$. Now substituting $w = v$ in (4.11), we get

$$\sup_{\mu \in M} \frac{b(v, \mu)}{\|\mu\|} \geq \frac{b(v, \lambda)}{\|\lambda\|} = \frac{(v, v)}{\|\lambda\|} \geq \beta\|v\|.$$

Thus $B : V^\perp \longrightarrow M'$ satisfies the three conditions of Theorem 3.6, and the mapping is an isomorphism.

Suppose condition (ii) is satisfied, i.e., $B : V^\perp \longrightarrow M'$ is an isomorphism. For given $\mu \in M$, we determine the norm via duality:

$$\begin{aligned} \|\mu\| &= \sup_{g \in M'} \frac{\langle g, \mu\rangle}{\|g\|} = \sup_{v \in V^\perp} \frac{\langle Bv, \mu\rangle}{\|Bv\|} \\ &= \sup_{v \in V^\perp} \frac{b(v, \mu)}{\|Bv\|} \leq \sup_{v \in V^\perp} \frac{b(v, \mu)}{\beta\|v\|}. \end{aligned}$$

But then condition (i) is satisfied, and everything is proved. □

Another condition which is equivalent to the inf-sup condition can be found in Problem 4.16, where we also interpret the condition 4.2(ii) as a decomposition property.

After these preparations, we are ready for the main theorem for saddle point problems [Brezzi 1974]. The condition (ii) in the theorem is often referred to as the *Brezzi condition*. The inf-sup condition is also called the *Ladyzhenskaya–Babuška–Brezzi condition* (LBB-condition) since Ladyzhenskaya provided an inequality for the divergence operator that is equivalent to the inf-sup condition for the Stokes problem. Recall that as in (4.6), the kernel of B is denoted by V.

4.3 Theorem. *(Brezzi's splitting theorem) For the saddle point problem (4.4), the mapping (4.5) defines an isomorphism $L : X \times M \longrightarrow X' \times M'$ if and only if the following two conditions are satisfied:*

(i) The bilinear form $a(\cdot, \cdot)$ is V-elliptic, i.e.,

$$a(v, v) \geq \alpha\|v\|^2 \quad \textit{for all } v \in V,$$

where $\alpha > 0$, and V is as in (4.6).

(ii) The bilinear form $b(\cdot, \cdot)$ satisfies the inf-sup condition (4.8).

Proof. Suppose the conditions on a and b are satisfied. We first show that for every pair of functionals $(f, g) \in X' \times M'$, there is exactly one solution (u, λ) of the saddle point problem satisfying

$$\begin{aligned} \|u\| &\le \alpha^{-1}\|f\| && + \beta^{-1}\Big(1+\frac{C}{\alpha}\Big)\|g\|, \\ \|\lambda\| &\le \beta^{-1}\Big(1+\frac{C}{\alpha}\Big)\|f\| && + \beta^{-1}\Big(1+\frac{C}{\alpha}\Big)\frac{C}{\beta}\|g\|. \end{aligned} \tag{4.12}$$

$V(g)$ is not empty for $g \in M'$. Indeed, by Lemma 4.2(ii), there exists $u_0 \in V^\perp$ with

$$Bu_0 = g.$$

Moreover, $\|u_0\| \le \beta^{-1}\|g\|$.

With $w := u - u_0$, (4.4) is equivalent to

$$\begin{aligned} a(w, v) + b(v, \lambda) &= \langle f, v\rangle - a(u_0, v) && \text{for all } v \in X, \\ b(w, \mu) &= 0 && \text{for all } \mu \in M. \end{aligned} \tag{4.13}$$

By the V-ellipticity of a, the function

$$\frac{1}{2}a(v, v) - \langle f, v\rangle + a(u_0, v)$$

attains its minimum for some $w \in V$ with

$$\|w\| \le \alpha^{-1}(\|f\| + C\|u_0\|) \le \alpha^{-1}(\|f\| + C\beta^{-1}\|g\|).$$

In particular, the Characterization Theorem II.2.2 implies

$$a(w, v) = \langle f, v\rangle - a(u_0, v) \quad \text{for all } v \in V. \tag{4.14}$$

The equations (4.13) will be satisfied if we can find $\lambda \in M$ such that

$$b(v, \lambda) = \langle f, v\rangle - a(u_0 + w, v) \quad \text{for all } v \in X.$$

The right-hand side defines a functional in X', which in view of (4.14) lies in V^0. By Lemma 4.2(iii), this functional can be represented as $B'\lambda$ with $\lambda \in M$, and

$$\|\lambda\| \le \beta^{-1}(\|f\| + C\|u\|).$$

This establishes the solvability. The inequalities (4.12) follow from the bounds on $\|u_0\|$, $\|w\|$, and $\|\lambda\|$ and the triangle inequality $\|u\| \le \|u_0\| + \|w\|$.

The solution is unique, as can be seen from the homogeneous equation. If we insert $f = 0$, $g = 0$, $v = u$, $\mu = -\lambda$ in (4.4) and add, we get $a(u, u) = 0$. Since $u \in V$, the V-ellipticity implies $u = 0$. Moreover,

$$\sup_v |b(v, \lambda)| = 0,$$

and $\lambda = 0$ follows from (4.8). Thus, L is injective and surjective, and (4.12) asserts that L^{-1} is continuous.

Conversely, suppose that L is an isomorphism. In particular, suppose $\|L^{-1}\| \le C$. By the Hahn–Banach theorem, every functional $f \in V'$ has an extension $\tilde{f} : X \longrightarrow \mathbb{R}$ with $\|\tilde{f}\| = \|f\|$. Set $(u, \lambda) = L^{-1}(\tilde{f}, 0)$. Then u is a minimum of $\frac{1}{2}a(v, v) - \langle f, v\rangle$ in V. The mapping $f \longmapsto u \in V$ is bounded, and thus a is V-elliptic.

Finally, for every $g \in M'$, we associate $u \in X$ with $\|u\| \le c\|g\|$ via $(u, \lambda) = L^{-1}(0, g)$. Given $u \in X$, let $u^\perp \in V^\perp$ be the projection. Since $\|u^\perp\| \le \|u\|$, the mapping $g \longmapsto u \longmapsto u^\perp$ is bounded, and $Bu^\perp = g$. Hence, $B : V^\perp \longrightarrow M'$ is an isomorphism, and by Lemma 4.2(ii), b satisfies the inf-sup condition. □

We note that coercivity of a was assumed only on the kernel of B and not on the entire space X. We will need this weak assumption in most applications. An exception will be the Stokes problem. Here coercivity is not restricted to the divergence-free functions. Note that the norm of the operator B does not enter into the a priori estimate (4.12).

If the bilinear form $a(u, v)$ is not symmetric, the assumption (i) on the ellipticity in Theorem 4.3 has to be replaced by an inf-sup condition; cf. Brezzi and Fortin [1991], p.41.

Mixed Finite Element Methods

We now discuss a natural approach to the numerical solution of saddle point problems: Choose finite-dimensional subspaces $X_h \subset X$ and $M_h \subset M$, and solve

Problem (S$_h$). Find $(u_h, \lambda_h) \in X_h \times M_h$ such that

$$\begin{aligned} a(u_h, v) + b(v, \lambda_h) &= \langle f, v\rangle \quad \text{for all } v \in X_h, \\ b(u_h, \mu) &= \langle g, \mu\rangle \quad \text{for all } \mu \in M_h. \end{aligned} \tag{4.15}$$

This approach is called a mixed method. In view of Lemma 3.7, we need to choose finite element spaces which satisfy requirements similar to those on X and M in Theorem 4.3, see Brezzi [1974] and Fortin [1977]. This is not always easy to do in practice. For fluid mechanics, the coercivity is trivial, and only the inf-sup condition is critical. For problems in elasticity theory, however, making

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

finite element spaces satisfy both conditions can often be difficult, and requires that the finite element spaces X_h and M_h fit together. Practical experience shows that enforcing these conditions is of the utmost importance for the stability of the finite element computation.

It is useful to introduce the following notation which is analogous to (4.6):

$$V_h := \{v \in X_h;\ b(v, \mu) = 0 \quad \text{for all } \mu \in M_h\}.$$

4.4 Definition. A family of finite element spaces X_h, M_h is said to satisfy the *Babuška–Brezzi condition* provided there exist constants $\alpha > 0$ and $\beta > 0$ independent of h such that

(i) The bilinear form $a(\cdot, \cdot)$ is V_h-elliptic with ellipticity constant $\alpha > 0$.

(ii)
$$\sup_{v \in X_h} \frac{b(v, \lambda_h)}{\|v\|} \geq \beta \|\lambda_h\| \quad \text{for all } \lambda_h \in M_h. \tag{4.16}$$

The terminology in the literature varies. Often the condition (ii) alone is called the *Brezzi condition*, the *Ladyzhenskaja–Babuška–Brezzi condition*, or for short the *LBB condition*. This condition is the more important of the two, and we will usually call it the *inf-sup condition*.

It is clear that – possibly after a reduction in α and β – we can take the same constants in 4.3 and 4.4.

The following result is an immediate consequence of Lemma 3.7 and Theorem 4.3.

4.5 Theorem. *Suppose the hypotheses of Theorem 4.3 hold, and suppose X_h, M_h satisfy the Babuška–Brezzi conditions. Then*

$$\|u - u_h\| + \|\lambda - \lambda_h\| \leq c\Big\{ \inf_{v_h \in X_h} \|u - v_h\| + \inf_{\mu_h \in M_h} \|\lambda - \mu_h\| \Big\}. \tag{4.17}$$

In general, $V_h \not\subset V$. We get a better result in the special case of conforming approximation where $V_h \subset V$. We note that in this case also the finite element approximation of $V(g)$ may be nonconforming for $g \neq 0$.

4.6 Definition. The spaces $X_h \subset X$ and $M_h \subset M$ satisfy condition (C) provided $V_h \subset V$, i.e., if for every $v_h \in X_h$, $b(v_h, \mu_h) = 0$ for all $\mu_h \in M_h$ implies $b(v_h, \mu) = 0$ for all $\mu \in M$.

4.7 Theorem. *Suppose the hypotheses of Theorem 4.5 are satisfied along with the condition (C). Then the solution of Problem (S_h) satisfies*

$$\|u - u_h\| \leq c \inf_{v_h \in X_h} \|u - v_h\|.$$

Proof. Let $v_h \in V_h(g)$. Then in the usual way, we have

$$\begin{aligned} a(u_h - v_h, v) &= a(u_h, v) - a(u, v) + a(u - v_h, v) \\ &= b(v, \lambda - \lambda_h) + a(u - v_h, v) \\ &\le C\|u - v_h\| \cdot \|v\| \end{aligned}$$

for all $v \in V_h$, since $b(v, \lambda - \lambda_h)$ vanishes because of condition (C). With $v := u_h - v_h$, we have $\|u_h - v_h\|^2 \le \alpha^{-1} C\|u_h - v_h\| \cdot \|u - v_h\|$, and the assertion follows after dividing by $\|u_h - v_h\|$. □

For completeness we mention that the assumption $X_h \subset X$ may be abandoned. Also mesh-dependent norms may be used. In these cases the theory above has to be combined with arguments that we encountered with Strang's lemmas as it is done, e.g., for the analysis of *mortar elements*; see Braess, Dahmen, and Wieners [2000]. The continuity of H^1 elements is replaced at some inter-element boundaries by explicit weak matching conditions. The constraints give rise to Lagrange multipliers that model the normal derivative $\partial u/\partial n$.

Fortin Interpolation

We continue with our treatment of abstract saddle point problems with a tool due to Fortin [1977] which is useful for verifying that the inf-sup condition holds.

4.8 Fortin's Criterion. *Suppose that the bilinear form* $b : X \times M \longrightarrow \mathbb{R}$ *satisfies the inf-sup condition. In addition, suppose that for the subspaces* X_h, M_h*, there exists a bounded linear projector* $\Pi_h : X \longrightarrow X_h$ *such that*

$$b(v - \Pi_h v, \mu_h) = 0 \quad \textit{for } \mu_h \in M_h. \tag{4.18}$$

If $\|\Pi_h\| \le c$ *for some constant* c *independent of* h*, then the finite element spaces* X_h, M_h *satisfy the inf-sup condition.*

$$\begin{tikzcd} X \arrow[r, "B"] \arrow[d, "\Pi_h"'] & M' \arrow[d, "J"] \\ X_h \arrow[r, "B"] & M'_h \end{tikzcd}$$

Commutative diagram property of (4.18). The symbol J refers to the injection.

Proof. By assumption, for $\mu_h \in M_h$,

$$\begin{aligned} \beta\|\mu_h\| &\le \sup_{v\in X} \frac{b(v, \mu_h)}{\|v\|} = \sup_{v\in X} \frac{b(\Pi_h v, \mu_h)}{\|v\|} \le c \sup_{v\in X} \frac{b(\Pi_h v, \mu_h)}{\|\Pi_h v\|} \\ &= c \sup_{v_h\in X_h} \frac{b(v_h, \mu_h)}{\|v_h\|}, \end{aligned}$$

since $\Pi_h v \in X_h$. □

Note that the condition in Fortin's criterion can be checked without referring explicitly to the norm of the Lagrange multipliers. This is an advantage when the space of the Lagrange multipliers is equipped with an exotic norm, and it is thus used for example when the Lagrange multipliers belong to trace spaces.

4.9 Remark. There is a converse statement to Fortin's criterion. *If the finite element spaces X_h, M_h satisfy the inf-sup condition, then there exists a bounded linear projector $\Pi_h : X \to X_h$ such that (4.18) holds.*

Indeed, given $v \in X$, define $u_h \in X_h$ as the solution of the equations

$$\begin{aligned} (u_h, w) + b(w, \lambda_h) &= (v, w) \quad \text{for all } w \in X_h, \\ b(u_h, \mu) \qquad\qquad &= b(v, \mu) \quad \text{for all } \mu \in M_h. \end{aligned} \tag{4.19}$$

Since the inner product in X is coercive by definition, the problem is stable, and from Theorem 4.3 it follows that

$$\|u_h\| \le c\|v\|.$$

Moreover, a linear mapping is defined by $v \longmapsto \Pi v := u_h$, and the proof is complete. □

The linear process defined above is called *Fortin interpolation.*

As a corollary we obtain a relationship between the approximation with the constraint induced by the bilinear form b and the approximation in the larger finite element space X_h.

4.10 Remark. If the spaces X_h and M_h satisfy the inf-sup condition, then there exists a constant c independent of h such that for every $u \in V(g)$,

$$\inf_{v_h \in V_h(g)} \|u - v_h\| \le c \inf_{w_h \in X_h} \|u - w_h\|.$$

Proof. We make use of Fortin interpolation. Obviously, $\Pi_h w_h = w_h$ for each $w_h \in X_h$. Given $u \in V(g)$ we have $\Pi_h u \in V_h(g)$ and

$$\|u - \Pi_h u\| = \|u - w_h - \Pi_h(u - w_h)\| \le (1 + c)\|u - w_h\|.$$

Since this holds for all $w_h \in X_h$, the proof is complete. □

Sometimes error estimates are wanted for some norms for which not all hypotheses of Theorem 4.3 hold. In this context we note that the norm of the bilinear form b does not enter into the a priori estimate (4.12). This fact is used for the estimate of $\|\lambda - \lambda_h\|$ when an estimate of $\|u - u_h\|$ has been established by applying other tools.

Saddle Point Problems with Penalty Term

To conclude this section, we consider a variant of Problem (S) which plays a role in elasticity theory. We want to treat so-called *problems with a small parameter t* in such a way that we get convergence as $h \to 0$ which is uniform in the parameter t. This can often be achieved by formulating a saddle point problem with penalty term. Readers who are primarily interested in the Stokes problem may want to skip the rest of this section.

Suppose that in addition to the bilinear forms a and b,

$$c : M_c \times M_c \longrightarrow \mathbb{R}, \quad c(\mu, \mu) \geq 0 \quad \text{for all } \mu \in M_c \tag{4.20}$$

is a bilinear form on a dense set $M_c \subset M$. Moreover, let t be a small real-valued parameter. Now we modify (4.4) by adding a *penalty term*:

Problem (S$_t$). Find $(u, \lambda) \in X \times M_c$ with

$$\begin{aligned} a(u, v) \ &+ b(v, \lambda) &&= \langle f, v\rangle \quad \text{for all } v \in X, \\ b(u, \mu) \ &- t^2 c(\lambda, \mu) &&= \langle g, \mu\rangle \quad \text{for all } \mu \in M_c. \end{aligned} \tag{4.21}$$

The associated bilinear form on the product space is

$$A(u, \lambda; v, \mu) := a(u, v) + b(v, \lambda) + b(u, \mu) - t^2 c(\lambda, \mu).$$

First we consider the case where c is bounded [Braess and Blömer 1990]. Then c can be extended continuously to the entire space $M \times M$, and we can suppose $M_c = M$.

4.11 Theorem. *Suppose that the hypotheses of Theorem 4.3 are satisfied and that $a(v, v) \geq 0$ for all $v \in X$. In addition, let $c : M \times M \longrightarrow \mathbb{R}$ be a continuous bilinear form with $c(\mu, \mu) \geq 0$ for all $\mu \in M$. Then (4.21) defines an isomorphism $L : X \times M \longrightarrow X' \times M'$, and L^{-1} is uniformly bounded for $0 \leq t \leq 1$.*

In Theorem 4.11 it is essential that the solution of the saddle point problem with penalty term is uniformly bounded in t for all $0 \leq t \leq 1$. We can think of the penalty term as a perturbation. It is often supposed to have a stabilizing effect. Surprisingly, this is not always true, and the norm of the form c enters into the constant in the inf-sup condition. The following example shows that this is not just an artifact of the proof, which is postponed to the end of this section.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

4.12 Example. Let $X = M := L_2(\Omega)$, $a(u, v) := 0$, $b(v, \mu) := (v, \mu)_{0,\Omega}$, and $c(\lambda, \mu) := K \cdot (\lambda, \mu)_{0,\Omega}$. Clearly, the solution of

$$\begin{aligned} b(v, \lambda) &= (f, v)_{0,\Omega}, \\ b(u, \mu) - t^2 c(\lambda, \mu) &= (g, \mu)_{0,\Omega} \end{aligned}$$

is $\lambda = f$ and $u = g + t^2 K f$. Thus, the solution grows as $K \to \infty$ and we cannot expect a bounded solution for an unbounded bilinear form c.

In plate theory we frequently encounter saddle point problems with penalty terms which represent *singular perturbations*, i.e., which stem from a differential operator of higher order. Then we introduce a semi-norm on M_c, and define the corresponding norm

$$\begin{aligned} |\mu|_c &:= \sqrt{c(\mu, \mu)}, \\ |||(v, \mu)||| &:= \|v\|_X + \|\mu\|_M + t|\mu|_c, \end{aligned} \tag{4.22}$$

on $X \times M_c$; see Huang [1990]. On the other hand, this now requires the ellipticity of a on the entire space X, rather than just on the kernel V as in Theorem 4.3. It is clear from the previous example that we indeed need some additional assumption of this kind.

4.13 Theorem. *Suppose the hypotheses of Theorem 4.3 are satisfied and that a is elliptic on X. Then the mapping L defined by the saddle point problem with penalty term satisfies the inf-sup condition*

$$\inf_{(u,\lambda)\in X\times M_c} \sup_{(v,\mu)\in X\times M_c} \frac{A(u, \lambda; v, \mu)}{|||(u, \lambda)||| \cdot |||(v, \mu)|||} \ge \gamma > 0, \tag{4.23}$$

for all $0 \le t \le 1$, *where* γ *is independent of* t.

These two theorems are consequences of the following lemma [Kirmse 1990] whose hypotheses appear to be very technical at first glance. However, by Problem 4.23, the condition (4.25) below is equivalent to the Babuška condition for the X-components,

$$\sup_{(v,\mu)} \frac{A(u, 0; v, \mu)}{|||(v, \mu)|||} \ge \alpha' \|u\|_X, \tag{4.24}$$

with suitable α'. In particular, it is therefore also necessary for stability.

4.14 Lemma. *Suppose that the hypotheses of Theorem 4.3 are satisfied, and suppose that*

$$\frac{a(u, u)}{\|u\|_X} + \sup_{\mu\in M_c} \frac{b(u, \mu)}{\|\mu\|_M + t|\mu|_c} \ge \alpha\|u\|_X \tag{4.25}$$

for all $u \in X$ *and some* $\alpha > 0$. *Then the inf-sup condition (4.23) holds.*

Proof. We consider three cases.
Case 1. Let $\|u\|_X + \|\lambda\|_M \le \delta^{-1} t|\lambda|_c$, where $\delta > 0$ will be chosen later. Then

$$\begin{aligned} A(u,\lambda; u, -\lambda) &= a(u,u) + t^2 c(\lambda,\lambda) \\ &\ge \frac{1}{2} t^2 |\lambda|_c^2 + \frac{1}{2} t^2 |\lambda|_c^2 \\ &\ge \frac{1}{2}\delta^2 \{(\|u\|_X + \|\lambda\|_M)^2 + t^2|\lambda|_c^2\} \ge \frac{1}{4}\delta^2 |||(u,\lambda)|||^2. \end{aligned}$$

Dividing through by $|||(u,\lambda)|||$, we have

$$\frac{1}{4}\delta^2 |||(u,\lambda)||| \le \frac{A(u,\lambda; u, -\lambda)}{|||(u,\lambda)|||} \le \sup_{(v,\mu)} \frac{A(u,\lambda; v,\mu)}{|||(v,\mu)|||}.$$

Case 2. Let $\|u\|_X + \|\lambda\|_M > \delta^{-1} t|\lambda|_c$ and $\|u\|_X \le \frac{\beta}{2\|a\|}\|\lambda\|_M$. By the inf-sup condition (4.8),

$$\begin{aligned} \beta\|\lambda\|_M \le \sup_v \frac{b(v,\lambda)}{\|v\|_X} &= \sup_v \frac{A(u,\lambda; v, 0) - a(u,v)}{\|v\|_X} \\ &\le \sup_{(v,\mu)} \frac{A(u,\lambda; v,\mu)}{|||(v,\mu)|||} + \|a\|\ \|u\|_X \\ &\le \sup_{(v,\mu)} \frac{A(u,\lambda; v,\mu)}{|||(v,\mu)|||} + \frac{1}{2}\beta\|\lambda\|_M. \end{aligned}$$

Now we can estimate $\|\lambda\|_M$, and in view of the case distinction $\|u\|_X$ and $t|\lambda|_c$ as well, by the first term on the right-hand side.

Case 3. Let $\|u\|_X + \|\lambda\|_M > \delta^{-1} t|\lambda|_c$ and $\|u\|_X \ge \frac{\beta}{2\|a\|}\|\lambda\|_M$. Then $\delta|||(u,\lambda)||| \le \|u\|_X$, where δ depends only on α, β and δ. By hypothesis (4.25),

$$\begin{aligned} \alpha\delta|||(u,\lambda)||| &\le \alpha\|u\|_X \\ &\le \frac{a(u,u)}{\|u\|_X} + \sup_\mu \frac{A(u,\lambda; 0,\mu) + t^2 c(\lambda,\mu)}{\|\mu\|_M + t|\mu|_c} \\ &\le \frac{A(u,\lambda; u,-\lambda)}{\|u\|_X} + \sup_\mu \frac{A(u,\lambda; 0,\mu)}{|||(0,\mu)|||} + t|\lambda|_c \\ &\le \left(\frac{1}{\delta} + 1\right) \sup_{(v,\mu)} \frac{A(u,\lambda; v,\mu)}{|||(v,\mu)|||} + t|\lambda|_c. \end{aligned}$$

With $\delta \le \frac{\alpha\beta}{4\|a\| + 2\beta}$ we have $t|\lambda|_c \le \frac{1}{2}\alpha\|u\|_X$, and the second term in the sum can be absorbed by a factor of 2.

This establishes the assertion in all cases. □

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

The previous two theorems now follow immediately. The ellipticity on the entire space X in Theorem 4.13 implies $a(u,u) \geq \alpha \|u\|_X^2$, and (4.25) is clear. On the other hand, Theorem 4.3 ensures that the Babuška condition holds for the pair $(u,0)$, and combining it with the Cauchy–Schwarz inequality gives

$$\begin{aligned}
\gamma \|u\|_X &\leq \sup_{(v,\mu)} \frac{a(u,v)+b(u,\mu)}{\|v\|_X + \|\mu\|_M} \\
&\leq \sup_v \frac{a(u,v)}{\|v\|_X} + \sup_\mu \frac{b(u,\mu)}{\|\mu\|_M} \\
&\leq [\|a\|\, a(u,u)]^{1/2} + (1+\|c\|) \sup_\mu \frac{b(u,\mu)}{\|\mu\|_M + |\mu|_c} \\
&\leq \frac{\|a\|\, a(u,u)}{\|u\|_X} + 2(1+\|c\|) \sup_\mu \frac{b(u,\mu)}{\|\mu\|_M + |\mu|_c}.
\end{aligned} \tag{4.26}$$

Here we have used the fact that the form c in Theorem 4.11 was assumed to be bounded, and have applied the same kind of argument as used in Problem 4.22. □

The uniform boundedness of the solution implies that the solution is a continuous function of the parameter.

4.15 Corollary. *Let the conditions of Theorem 4.11 prevail. Then, given $f \in X'$ and $g \in M'$, the solution (u,λ) of Problem (S_t) depends continuously on t.*

Proof. Let (u_t, λ_t) and (u_s, λ_s) be the solutions for the parameters t and s respectively. Then we have

$$\begin{aligned}
a(u_t - u_s, v) \ &+ b(v, \lambda_t - \lambda_s) &&= 0 && \text{for all } v \in X, \\
b(u_t - u_s, \mu) \ &- t^2 c(\lambda_t - \lambda_s, \mu) &&= -(t^2 - s^2) c(\lambda_s, \mu) && \text{for all } \mu \in M.
\end{aligned}$$

The stability now implies $\|u_t - u_s\|_X + \|\lambda_t - \lambda_s\|_M \leq \text{const}\, |t^2 - s^2|$, and we have continuity in the parameter. □

Typical Applications

Variational problems in saddle point form and mixed methods are used for very different purposes. We list several here.

1. Explicit constraints. When incompressible flows are investigated, there is the explicit constraint

$$\operatorname{div} u = 0.$$

In particular, the Stokes problem will be considered in §§6 and 7.

2. Splitting of a differential equation into a system. As an example, the Poisson equation is written as a system

$$\begin{aligned}
\sigma - \operatorname{grad} u &= 0, \\
\operatorname{div} \sigma &= -f;
\end{aligned}$$

see §5. Here the given differential equation of order 2 is split into a system of equations of first order. The mixed method induces a softening effect that is desired in some difficult problems of solid mechanics; cf. Ch. VI, §3. There is a different reason for a split of equations of fourth order into two equations of second order as in Problem 4.24 or with Kirchhoff plates in Ch. VI, §5. The mixed method admits the use of C^0 elements while conforming methods with the standard variational formulation require C^1 elements.

3. Modeling boundary conditions. In some cases it is more convenient to have a boundary condition $u|_\Gamma - g = 0$ as an explicit constraint than to incorporate the boundary values into the finite element functions. Here the Lagrange multiplier models $\partial u/\partial n$ or, more generally, the multiple that is encountered in natural boundary conditions. Similarly, the C^0 continuity is a handicap in domain decomposition methods and is replaced by explicit matching conditions. This holds in particular for *mortar elements*; see Bernardy, Maday, and Patera [1994] or Braess, Dahmen, and Wieners [2000].

4. Mixed elements that are equivalent to nonconforming methods. Often one finds mixed methods that are equivalent to nonconforming elements. While the nonconforming elements are more easily implemented, the mixed method may admit an easier proof of convergence; see the DKT element for Kirchhoff plates in Ch. VI, §5 and the connection between the nonconforming P_1 element and the Raviart–Thomas element described by Marini [1985].

5. Saddle point problems with penalty terms. Problems with a large parameter are often rewritten as a saddle point problem with a small penalty term. Examples are the flow of a nearly incompressible fluid and the Reissner–Mindlin plate; see Problem 4.19, Ch. VI, §§3 and 6.

6. A posteriori error estimates via saddle point problems. Nearly optimal solutions of saddle point problems provide lower estimates of the (minimal) value of variational problems and thus also a posteriori error estimates; see §9.

Problems

4.16 Show that the inf-sup condition (4.8) is equivalent to the following decomposition property: For every $u \in X$ there exists a decomposition

$$u = v + w$$

with $v \in V$ and $w \in V^\perp$ such that

$$\|w\|_X \le \beta^{-1} \|Bu\|_{M'},$$

where $\beta > 0$ is a constant independent of u.

4.17 Let X, M, and the maps a, b, f, g be as in the saddle point problem (S). Given $\rho \in M$, let $\rho^\perp := \{\mu \in M;\ (\rho, \mu) = 0\}$. We now minimize the expression (4.1) subject to the restricted set of constraints

$$b(u, \mu) = \langle g, \mu \rangle \quad \text{for all } \mu \in \rho^\perp.$$

Show that the solution is characterized by

$$\begin{aligned} a(u, v) + b(v, \lambda) \qquad\qquad &= \langle f, v \rangle \quad \text{for all } v \in X, \\ b(u, \mu) \qquad\qquad + (\sigma, \mu) &= \langle g, \mu \rangle \quad \text{for all } \mu \in M, \\ (\tau, \lambda) \qquad\qquad\qquad &= 0 \qquad\quad \text{for all } \tau \in \text{span}[\rho] \end{aligned} \tag{4.27}$$

with $u \in X$, $\lambda \in M$, $\sigma \in \text{span}[\rho]$.
Rewrite (4.27) and verify that it is a standard saddle point problem with the spaces $\tilde{X} := X \times \text{span}[\rho]$ and $\tilde{M} := M$.

4.18 Suppose that the subspaces $X_h,\ M_h$ satisfy the Babuška–Brezzi condition, and suppose we

increase or decrease X_h or M_h.

Which of the conditions in Definition 4.4 have to be rechecked?

4.19 When $M = L_2$, we can identify M with its dual space, and simply write $b(v, \mu) = (Bv, \mu)_0$. The solution of the saddle point problem does not change if $a(u, v)$ is replaced by

$$a_t(u, v) := a(u, v) + t^{-2}(Bu, Bv)_0, \quad t > 0.$$

This is called the method of the *augmented Lagrange function*; see Fortin and Glowinski [1983].

(a) Show that a_t is elliptic on the entire space X under the hypotheses of Theorem 4.3.

(b) Suppose we ignore the explicit constraints, and introduce $\lambda = t^2 Bu$ as a new variable. Show that this leads to a saddle point problem with penalty term; cf. (6.15).

4.20 As, e.g., in (5.2) and (5.5), a saddle point problem is often stable for two pairings X_1, M_1 and X_2, M_2. Now suppose $X_1 \subset X_2$ and

$$\|v\|_{X_1} \geq \|v\|_{X_2} \quad \text{on } X_1.$$

Show that, conversely,

$$\|\lambda\|_{M_1} \leq c\, \|\lambda\|_{M_2} \quad \text{on } M_1 \cap M_2,$$

where $c \geq 0$. If M_1 is also dense in M_2, then $M_1 \supset M_2$.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

4.21 The pure Neumann Problem (II.3.8)

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

is only solvable if $\int_\Omega f\,dx + \int_\Gamma g\,ds = 0$. This compatibility condition follows by applying Gauss' integral theorem to the vector field ∇u. Since $u+$ const is a solution whenever u is, we can enforce the constraint

$$\int_\Omega u dx = 0.$$

Formulate the associated saddle point problem, and use the trace theorem and the second Poincaré inequality to show that the hypotheses of Theorem 4.3 are satisfied.

4.22 Let a, b, and c be positive numbers. Show that $a \le b + c$ implies $a \le b^2/a + 2c$.

4.23 Show the equivalence of the conditions (4.24) and (4.25). For the nontrivial direction, use the same argument as in the derivation of (4.26); cf. Braess [1996].

4.24 Let u be a (classical) solution of the biharmonic equation

$$\begin{aligned} \Delta^2 u &= f \quad \text{in } \Omega, \\ u = \frac{\partial u}{\partial \nu} &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Show that $u \in H_0^1$ together with $w \in H^1$ is a solution of the saddle point problem

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

Suitable elements and analytic methods can be found, e.g., in Ciarlet [1978] and in Babuška, Osborn, and Pitkäranta [1980].

4.25 Equations of the form

$$\begin{aligned} a(u, v) + b(v, \lambda) \qquad\qquad &= \langle f, v\rangle && \text{for all } v \in X, \\ b(u, \mu) \qquad\qquad + c(\sigma, \mu) &= \langle g, \mu\rangle && \text{for all } \mu \in M, \\ c(\tau, \lambda) + d(\sigma, \tau) &= \langle h, \tau\rangle && \text{for all } \tau \in Y \end{aligned} \tag{4.28}$$

are sometimes called *double saddle point problems*. Rearrange (4.28) to obtain a standard saddle point problem.

§ 5. Mixed Methods for the Poisson Equation

The treatment of the Poisson equation by mixed methods already elucidates some characteristic features and shows that saddle point formulations are not only useful for minimization problems with given constraints as in (4.1), (4.2). For example, there are two different pairs of spaces for which the saddle point problem is stable in the sense of Babuška and Brezzi. It is interesting that different boundary conditions turn out to be natural conditions in the two cases.

The method, often called the *dual mixed method*, has been well established for a long time. On the other hand, the *primal mixed method* has recently attracted a lot of interest since it shows that mixed methods are often related to a softening of the energy functional and how elasticity problems with a small parameter can be treated in a robust way.

Moreover there are special results if X or M coincides with an L_2-space.

The Poisson Equation as a Mixed Problem

The Laplace equation or the Poisson equation $\Delta u = \operatorname{div}\operatorname{grad} u = -f$ can be written formally as the system

$$\begin{aligned} \operatorname{grad} u &= \sigma, \\ \operatorname{div} \sigma &= -f. \end{aligned} \tag{5.1}$$

Let $\Omega \subset \mathbb{R}^d$. Then (5.1) leads directly to the following saddle point problem: Find $(\sigma, u) \in L_2(\Omega)^d \times H_0^1(\Omega)$ such that

$$\begin{aligned} (\sigma, \tau)_{0,\Omega} - (\tau, \nabla u)_{0,\Omega} &= 0 && \text{for all } \tau \in L_2(\Omega)^d, \\ -(\sigma, \nabla v)_{0,\Omega} \qquad\qquad &= -(f, v)_{0,\Omega} && \text{for all } v \in H_0^1(\Omega). \end{aligned} \tag{5.2}$$

These equations can be treated in the general framework of saddle point problems with

$$\begin{aligned} X &:= L_2(\Omega)^d, & M &:= H_0^1(\Omega), \\ a(\sigma, \tau) &:= (\sigma, \tau)_{0,\Omega}, & b(\tau, v) &:= -(\tau, \nabla v)_{0,\Omega}. \end{aligned} \tag{5.3}$$

The linear forms are continuous, and a is obviously L_2-elliptic. To check the inf-sup condition, we use Friedrichs' inequality in a similar way as for the original minimum problem in Ch. II, §2. Given $v \in H_0^1(\Omega)$, consider the quotient appearing in the condition for $\tau := -\nabla v \in L_2(\Omega)^d$:

$$\frac{b(\tau, v)}{\|\tau\|_0} = \frac{-(\tau, \nabla v)_{0,\Omega}}{\|\tau\|_0} = \frac{(\nabla v, \nabla v)_{0,\Omega}}{\|\nabla v\|_0} = |v|_1 \ge \frac{1}{c}\|v\|_1.$$

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

Since c comes from Friedrichs' inequality and depends only on Ω, the saddle point problem (5.2) is stable.

It is easy to construct suitable finite elements for a triangulation $\mathcal{T}_h$. Choose $k \ge 1$, and set

$$\begin{aligned} X_h &:= (\mathcal{M}^{k-1})^d &&= \{\sigma_h \in L_2(\Omega)^d;\ \sigma_h|_T \in \mathcal{P}_{k-1} \text{ for } T \in \mathcal{T}_h\}, \\ M_h &:= \mathcal{M}^k_{0,0} &&= \{v_h \in H^1_0(\Omega);\ \ v_h|_T \in \mathcal{P}_k \quad \text{for } T \in \mathcal{T}_h\}. \end{aligned}$$

Note that only the functions in M_h are continuous. Since $\nabla M_h \subset X_h$, we can verify the inf-sup condition in the same way as for the continuous problem.

The saddle point problem with a different pairing is more important for practical computations. It refers to the space encountered in Problem II.5.14:

$$H(\text{div}, \Omega) := \{\tau \in L_2(\Omega)^d;\ \text{div}\,\tau \in L_2(\Omega)\}$$

with the graph norm of the divergence operator,

$$\|\tau\|_{H(\text{div},\Omega)} := (\|\tau\|_0^2 + \|\text{div}\,\tau\|_0^2)^{1/2}. \tag{5.4}$$

We seek $(\sigma, u) \in H(\text{div}, \Omega) \times L_2(\Omega)$ such that

$$\begin{aligned} (\sigma, \tau)_{0,\Omega} + (\text{div}\,\tau, u)_{0,\Omega} &= 0 && \text{for all } \tau \in H(\text{div}, \Omega), \\ (\text{div}\,\sigma, v)_{0,\Omega} &= -(f, v)_{0,\Omega} && \text{for all } v \in L_2(\Omega). \end{aligned} \tag{5.5}$$

To apply the general theory, we set

$$\begin{aligned} X &:= H(\text{div}, \Omega), & M &:= L_2(\Omega), \\ a(\sigma, \tau) &:= (\sigma, \tau)_{0,\Omega}, & b(\tau, v) &:= (\text{div}\,\tau, v)_{0,\Omega}. \end{aligned}$$

Clearly, the linear forms are continuous. Then since $\text{div}\,\tau = 0$ for τ in the kernel V, we have

$$a(\tau, \tau) = \|\tau\|_0^2 = \|\tau\|_0^2 + \|\text{div}\,\tau\|_0^2 = \|\tau\|^2_{H(\text{div},\Omega)}.$$

This establishes the ellipticity of a on the kernel. Moreover, for given $v \in L_2$ there exists $w \in C_0^\infty(\Omega)$ with $\|v - w\|_{0,\Omega} \le \frac{1}{2}\|v\|_{0,\Omega}$. Set $\xi := \inf\{x_1;\ x \in \Omega\}$ and

$$\begin{aligned} \tau_1(x) &= \int_\xi^{x_1} w(t, x_2, \ldots, x_n)dt, \\ \tau_i(x) &= 0 \quad \text{for } i > 1. \end{aligned}$$

Then obviously $\text{div}\,\tau = \partial\tau_1/\partial x_1 = w$, and the same argument as in the proof of Friedrichs' inequality gives $\|\tau\|_0 \le c\|w\|_0$. Hence,

$$\frac{b(\tau, v)}{\|\tau\|_{H(\text{div},\Omega)}} \ge \frac{(w, v)_{0,\Omega}}{(1 + c)\|w\|_{0,\Omega}} \ge \frac{1}{2(1 + c)}\|v\|,$$

and so the inf-sup condition is satisfied.

By Theorem 4.3, (5.5) defines a stable saddle point problem. At first glance, it appears that a solution exists only in $u \in L_2$. However, $u \in H_0^1(\Omega)$, and since $C_0^\infty(\Omega)^d \subset H(\text{div}, \Omega)$, the first equation of (5.5) says in particular that

$$\int_\Omega u \frac{\partial \tau_i}{\partial x_i} dx = -\int_\Omega \sigma_i \tau_i dx \quad \text{for } \tau_i \in C_0^\infty(\Omega).$$

Thus, in view of Definition II.1.1, u possesses a weak derivative $\frac{\partial u}{\partial x_i} = \sigma_i$, and hence $u \in H^1(\Omega)$. Now (5.5) together with Green's formula (II.2.9) and $\nabla u = \sigma$ implies

$$\begin{aligned} \int_{\partial\Omega} u \cdot \tau n \, ds &= \int_\Omega \nabla u \cdot \tau \, dx + \int_\Omega \text{div}\, \tau u \, dx \\ &= \int_\Omega \sigma \cdot \tau \, dx + \int_\Omega \text{div}\, \tau u \, dx = 0. \end{aligned} \tag{5.6}$$

Since this holds for all $\tau \in C^\infty(\Omega)^d$, we have $u = 0$ on the boundary in the generalized sense, i.e., in fact $u \in H_0^1(\Omega)$.

In the standard case the natural boundary condition is $\frac{\partial u}{\partial n} = 0$, but here the natural boundary condition is $u = 0$.

We note that the equations (5.2) characterize the solution of the variational problem

$$\begin{gathered} \frac{1}{2}(\sigma, \sigma)_0 - (f, u) \to \min! \\ \nabla u - \sigma = 0. \end{gathered} \tag*{(5.2)$_v$}$$

Here the Lagrange multiplier coincides with σ, and can be eliminated from the equations. On the other hand, (5.5) arises from the variational problem

$$\begin{gathered} -\frac{1}{2}(\sigma, \sigma)_0 \to \max! \\ \text{div}\, \sigma + f = 0. \end{gathered} \tag*{(5.5)$_v$}$$

Here the Lagrange multiplier coincides with u from (5.1).

Sometimes (5.2) with $X := L_2$ and $M := H_0^1$ is called a *primal* mixed method while (5.5) with $X := H(\text{div})$ and $M := L_2$ is called a *dual* mixed method.

When the functionals are evaluated for the optimal solutions, the values of the two variational problems $(5.2)_v$ and $(5.5)_v$ are equal. Therefore the common optimal value lies between those for arbitrary admissible functions of the variational problems. In this way, an error estimate for suboptimal solutions is obtained. It is the source of the a posteriori error estimate in Theorem 9.4. Here it is provided for the Poisson equation with mixed boundary conditions,

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= u_0 \quad \text{on } \Gamma_D, \\ \frac{\partial u}{\partial n} &= g \quad \text{on } \Gamma_N = \partial\Omega \backslash \Gamma_D. \end{aligned} \tag{5.7}$$

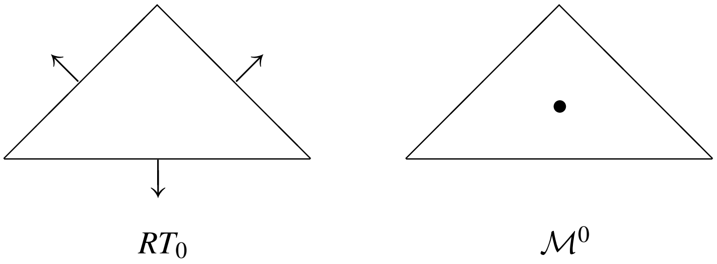


Fig. 35. Raviart–Thomas element for $k = 0$: One normal component is prescribed on each edge

5.1 Theorem of Prager and Synge. *Let* $\sigma \in H(\mathrm{div}, \Omega)$ *with* $\sigma \cdot n = g$ *on* Γ_N *and* $v \in H^1(\Omega)$ *with* $v = u_0$ *on* Γ_D. *Assume that* $\mathrm{div}\, \sigma + f = 0$. *Furthermore, let* u *be the solution of (5.7). Then,*

$$|u - v|_1^2 + \| \operatorname{grad} u - \sigma \|_0^2 = \| \operatorname{grad} v - \sigma \|_0^2.$$

Proof. By applying Green's formula and noting that $\Delta u = \mathrm{div}\, \sigma = -f$ we obtain

$$\begin{aligned}&\int_\Omega \operatorname{grad}(u - v)(\operatorname{grad} u - \sigma) dx \\ &= -\int_\Omega (u - v)(\Delta u - \mathrm{div}\, \sigma) dx + \int_{\partial\Omega} (u - v)(\frac{\partial u}{\partial n} - \sigma \cdot n) ds = 0.\end{aligned}$$

The boundary term above vanishes since $u - v = 0$ on Γ_D and $\frac{\partial u}{\partial n} - \sigma \cdot n = 0$ on Γ_N. From this orthogonality relation we conclude that

$$\| \operatorname{grad} v - \sigma \|_0^2 = \| \operatorname{grad}(v - u) \|_0^2 + \| \operatorname{grad} u - \sigma \|_0^2$$

which by the definition of the $| \cdot |_1$-semi-norm yields the desired equation. □

The Raviart–Thomas Element

The elements of Raviart and Thomas [1977] are suitable for the saddle point problem (5.5). Let $k \geq 0$, $\Omega \subset \mathbb{R}^2$, and suppose $\mathcal{T}_h$ is a shape-regular triangulation and that

$$\begin{aligned} X_h &:= RT_k \\ &:= \{\tau \in L_2(\Omega)^2;\ \tau|_T = \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T \begin{pmatrix} x \\ y \end{pmatrix},\ a_T, b_T, c_T \in \mathcal{P}_k \text{ for } T \in \mathcal{T}_h, \\ &\qquad \tau \cdot n \ \text{ is continuous on the inter-element boundaries}\}, \\ M_h &:= \mathcal{M}^k(\mathcal{T}_h) = \{v \in L_2(\Omega);\ v|_T \in \mathcal{P}_k \text{ for } T \in \mathcal{T}_h\}. \end{aligned} \tag{5.8}$$

The continuity of the normal components on the boundaries ensures the conformity $X_h \subset H(\mathrm{div}, \Omega)$; cf. Problem II.5.14.

For convenience, we consider the Raviart–Thomas element only for $k = 0$. Its construction heavily depends on the following assertion. The functions in $(\mathcal{P}_1)^2$ which have the form

$$p = \begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x \\ y \end{pmatrix}$$

are characterized by the fact that $n \cdot p$ is constant on each line $\alpha x + \beta y = \text{const}$ whenever n is orthogonal to the line. Therefore, given a triangle T, the normal component is constant and can be prescribed on each edge of T (see Fig. 35). Formally, the Raviart–Thomas element is the triple

$$(T, (\mathcal{P}_0)^2 + x \cdot \mathcal{P}_0, n_i p(z_i), i = 1, 2, 3 \text{ with } z_i \text{ being the midpoint of edge } i).$$

The solvability of the interpolation problem is easily verified. Given a vertex a_i of T, we can find a vector $r_i \in \mathbb{R}^2$ such that its projections onto the normals of the adjacent edges have the prescribed values. Now determine $p \in (\mathcal{P}_1)^2$ such that

$$p(a_i) = r_i, \quad i = 1, 2, 3.$$

It is immediate from $p \in (\mathcal{P}_1)^2$ that the normal components are linear on each edge of the triangle. They are even constant, since by construction they attain the same value at both vertices of the edge. Thus the function constructed indeed belongs to the specified subset of $(\mathcal{P}_1)^2$.

A proof of the inf-sup condition will be given below.

The Raviart–Thomas element and the similar BDM elements due to Brezzi, Douglas, and Marini [1985] are frequently used for the discretization of problems in $H(\text{div}, \Omega)$. Analogous elements for 3-dimensional problems have been described by Brezzi, Douglas, Durán, and Fortin [1987].

The finite element solution of the Raviart–Thomas element is related to the nonconforming P_1 element; see Marini [1985].

Interpolation by Raviart–Thomas elements

Due to Theorem 4.5 the error of the finite element solution for the discretization with the Raviart–Thomas element can be expressed in terms of approximation by the finite element functions. As usual the latter is estimated via interpolation. To this end an interpolation operator is defined which is based on the degrees of freedom specified in the definition of the element.

5.2 An Interpolation Operator. Let $k \geq 0$ and T be a triangle. Define

$$\rho_T : H^1(T) \to RT_k(T)$$

by

$$\int_e (q - \rho_T q) \cdot n p_k \, ds = 0 \qquad \forall p_k \in \mathcal{P}_k \text{ and each edge } e \subset \partial T, \tag{5.9a}$$

$$\int_T (q - \rho_T q) \cdot p_{k-1} \, dx = 0 \qquad \forall p_{k-1} \in \mathcal{P}_{k-1}^2 \quad (\text{if } k \geq 1). \tag{5.9b}$$

Given a triangulation $\mathcal{T}$ on Ω, define $\rho_\Omega : H^1(\Omega) \to RT_k$ locally by

$$(\rho_\Omega q)_{|T} = \rho_T(q_{|T}) \qquad \forall T \in \mathcal{T}.$$

We restrict ourselves to the case $k = 0$. We recall that the normal component of $v \in RT_0$ is constant on each edge. Equation (5.9a) states that it coincides with the mean value of the normal component of the given function. This holds for the solution of the interpolation problem.

From Gauss' integral theorem we conclude now that

$$\int_T \operatorname{div}(q - \rho_T q) dx = \sum_{e \subset \partial T} \int_e (q - \rho_T q) \cdot n ds = 0. \tag{5.10}$$

On the other hand, the Raviart–Thomas element is piecewise linear and $\alpha := \operatorname{div} \rho_T q$ is constant on T. By (5.10) α is the mean value of $\operatorname{div} v$ on T. Therefore α is the constant with the least L_2 deviation from $\operatorname{div} q$. So we have established the following property for $k = 0$. A proof for $k > 0$ can be found in Brezzi and Fortin [1990].

5.3 Minimal Property. *Given a triangulation $\mathcal{T}$ on Ω, let Π_k be the L^2-projection onto $\mathcal{M}^k$. Then we have for all $q \in H^1(\Omega)$*

$$\operatorname{div}(\rho_\Omega q) = \Pi_k \operatorname{div} q. \tag{5.11}$$

This equation is often called the *commuting diagram property*

$$\begin{tikzcd}
H_1(\Omega) \arrow[r, "\text{div}"] \arrow[d, "\rho_\Omega"'] & L_2(\Omega) \arrow[d, "\Pi_k"] \\
RT_k \arrow[r, "\text{div}"] & \mathcal{M}^k.
\end{tikzcd} \tag{5.12}$$

We note that the mapping ρ_Ω is not bounded on $H(\text{div})$; see Brezzi and Fortin [1991], p. 124–125. For a remedy, we refer to the discussion after Theorem 5.6 below.

The proof of the inf-sup condition is related to the following.

5.4 Lemma. *The mapping*

$$\operatorname{div} : RT_0 \to \mathcal{M}^0$$

is surjective.

Proof. After enlarging Ω by finitely many triangles if necessary, we may assume that Ω is convex. Given $f \in \mathcal{M}^0$, there is a $u \in H^2(\Omega) \cap H_0^1(\Omega)$ such that $\Delta u = f$. Set $q := \operatorname{grad} u$. By Gauss' integral formula we have

$$\int_{\partial T} q \cdot n ds = \int_T \operatorname{div} q dx = \int_T f dx.$$

From (5.9a) we conclude that $\int_T \operatorname{div} \rho_\Omega q dx = \int_{\partial T} (\rho_\Omega q) \cdot n ds = \int_T f dx$. Since $\operatorname{div} \rho_\Omega q$ and f are constant in T, it follows that $\operatorname{div} \rho_\Omega q = f$. □

Finally we note that the mapping $\mathcal{M}^0 \to RT_0$ in the construction above is bounded. Therefore, recalling Fortin's criterion we see that the inf-sup condition has been established simultaneously.

The error of the finite element solution will be derived from the approximation error.

5.5 Lemma. *Let $\mathcal{T}_h$ be a shape-regular triangulation of Ω. Then*

$$\|q - \rho_\Omega q\|_{H(\operatorname{div},\Omega)} \le ch\, |q|_1 + \inf_{v_h \in \mathcal{M}^0} \|\operatorname{div} q - v_h\|_0.$$

Proof. We first consider the interpolation on a triangle. By the trace theorem the functional $q \mapsto \int_e q \cdot n ds$, $e \subset \partial T$, is continuous on $H^1(T)^2$. Moreover, we have $\rho_T q = q$ for $q \in \mathcal{P}_0^2$ since $\mathcal{P}_0^2 \subset RT_0$. Therefore, the Bramble–Hilbert lemma and a scaling argument yield

$$\|q - \rho_\Omega q\|_0 \le ch\, |q|_1 \,.$$

The bound for $\operatorname{div}(q - \rho_\Omega q)$ follows from the minimal property 5.3, and the proof is complete. □

Now the error estimate of the finite element solution of (5.1)

$$\begin{aligned} &\|\sigma - \sigma_h\|_{H(\operatorname{div},\Omega)} + \|u - u_h\|_0 \\ &\qquad \le c(h\, |\sigma|_1 + h\, \|u\|_1 + \inf_{f_h \in \mathcal{M}^0} \|f - f_h\|_0) \end{aligned} \tag{5.13}$$

is a direct consequence of Theorem 4.5.

Moreover, there is a comparison with the standard finite element approximation.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

5.6 Theorem. *Let u_h be the finite element solution with the P_1 element and σ_h be the solution of the mixed method with the Raviart–Thomas element on the same mesh. Then*

$$\|\nabla u - \sigma_h\|_0 \le c\|\nabla(u - u_h)\|_0 + ch \inf_{f_h \in \mathcal{M}^0} \|f - f_h\|_0$$

with a constant c depending only on the shape regularity.

The proof is more involved and will be provided in §9 in connection with a posteriori error estimates.

The error estimate for the u-component in (5.13) is weaker than that for standard finite elements. On the other hand, the Raviart–Thomas element is more robust than the standard method for a class of problems that we will encounter in Ch. VI. Moreover the above disadvantage can be eliminated by a postprocessing procedure that will be described briefly in the next subsection.

Since the mixed method with Raviart–Thomas elements is stable, we can modify the diagram (5.12) such that the domain becomes $H(\mathrm{div})$. We restrict ourselves to $k = 0$ and define $\tilde{\rho}_\Omega$ as follows: Given $\sigma \in H(\mathrm{div})$, let $\sigma_h = \tilde{\rho}_\Omega \sigma \in RT_0$ be the solution of the mixed method

$$\begin{aligned}
(\sigma_h, \tau)_{0,\Omega} + (\mathrm{div}\,\tau, w_h)_{0,\Omega} &= (\sigma_h, \tau)_{0,\Omega} && \text{for all } \tau \in RT_0, \\
(\mathrm{div}\,\sigma, v)_{0,\Omega} &= (\mathrm{div}\,\sigma, v))_{0,\Omega} && \text{for all } v \in \mathcal{M}^0.
\end{aligned}$$

Since the divergence operator is surjective, we have exact sequences in addition to the commuting diagram property. Some larger diagrams play an important role in the construction of modern finite element spaces; see Arnold, Falk, and Winther [2006].

$$\begin{tikzcd}
H(\mathrm{div}, \Omega) \arrow[r, "\mathrm{div}"] \arrow[d, "\tilde{\rho}_\Omega"'] & L_2(\Omega) \arrow[r] \arrow[d, "\Pi_k"] & 0 \\
RT_k \arrow[r, "\mathrm{div}"] & \mathcal{M}^k \arrow[r] & 0.
\end{tikzcd}$$

Implementation and Postprocessing

In principle, the discretization leads to an indefinite system of equations. It can be turned into a positive definite system by a trick which was described by Arnold and Brezzi [1985].

Instead of initially choosing the gradients to lie in a subspace of $H(\mathrm{div}, \Omega)$, we first admit gradients in $L_2(\Omega)^2$, and later explicitly require that $\mathrm{div}\,\sigma_h \in L_2(\Omega)$. Equivalently, we require that the normal components $\sigma_h \cdot n$ do not have jumps on the edges. To achieve this, we enforce the continuity of $\sigma_h n$ on the edges as an explicit constraint. This introduces a further Lagrange multiplier.

The approximating functions for σ_h no longer involve continuity conditions, and each basis function has support on a single triangle. If we eliminate the associated variables by static condensation, the resulting equations are just as sparse as before the elimination process. In addition, we have avoided the costly construction of a basis of Raviart–Thomas elements.

A further advantage is that the Lagrange multiplier can be regarded as a finite element approximation of u on the edges. Arnold and Brezzi [1985] used them to improve the finite element solution.

Mesh-Dependent Norms for the Raviart–Thomas Element

Finite element computations with the Raviart–Thomas elements may also be analyzed in the framework of primal mixed methods, i.e., with the pairing $H^1(\Omega)$, $L_2(\Omega)$. Since the tangential components of the functions in (5.8) may have jumps on inter-element boundaries, in this context the elements are nonconforming and we need mesh-dependent norms which contain edge terms in adddition to broken norms

$$\begin{aligned}\|\tau\|_{0,h} &:= \Bigg(\|\tau\|_0^2 + h\sum_{e\subset\Gamma_h}\|\tau n\|_{0,e}^2\Bigg)^{1/2},\\ |v|_{1,h} &:= \Bigg(\sum_{T\in\mathcal{T}_h}|v|_{1,T}^2 + h^{-1}\sum_{e\subset\Gamma_h}\|J(v)\|_{0,e}^2\Bigg)^{1/2}.\end{aligned}\tag{5.14}$$

Here, $\Gamma_h := \cup_T(\partial T\cap\Omega)$ is the set of inter-element boundaries. On the edges of Γ_h the jump $J(v)$ of v and the normal component τn of τ are well defined. We note that both τn and $J(v)$ change sign if the orientation of an edge is reversed. Therefore, the product is independent of the orientation.

The continuity of the bilinear form $a(\cdot,\cdot)$ is obvious. Its coercivity follows from

$$\|\tau\|_{0,h} \le C\|\tau\|_0 \quad \text{for all } \tau\in RT_k$$

which in turn is obtained by a standard scaling argument. The bilinear form $b(\cdot,\cdot)$ is rewritten by the use of Green's formula

$$b(\tau,v) = -\sum_{T\in\mathcal{T}_\langle}\int_T \tau\cdot\operatorname{grad} v\,dx + \int\limits_{\Gamma_h} J(v)\tau n\,ds.\tag{5.15}$$

Now its continuity with respect to the norms (5.14) is immediate.

5.7 Lemma. *The inf-sup condition*

$$\sup_{\tau\in RT_k} \frac{b(\tau, v)}{\|\tau\|_{0,h}} \ge \beta |v|_{1,h} \quad \textit{for all } v \in \mathcal{M}^k$$

holds with a constant $\beta > 0$ which depends only on k and the shape regularity of the triangulation $\mathcal{T}_h$.

Proof. We restrict ourselves to the case $k = 0$. Given $v \in \mathcal{M}^0$, we note that the jump $J(v)$ is constant on each edge $e \subset \Gamma_h$. Therefore, there exists $\tau \in RT_0$ such that

$$\tau n = h^{-1} J(v) \text{ on each edge } e \subset \Gamma_h.$$

Since the area term in (5.15) vanishes on each T, it follows that

$$b(\tau, v) = h^{-1} \int_{\Gamma_h} |J(v)|^2 ds = ch^{-1} \sum_{e\subset\Gamma_h} \|J(v)\|_{0,e}^2 = |v|_{1,h}^2.$$

On the other hand we have $\|\tau\|_{0,h}^2 \le ch \sum_{e\subset\Gamma_h} \|\tau\|_{0,e}^2 = ch^{-1} \sum_{e\subset\Gamma_h} \|J(v)\|_{0,e}^2 = c|v|_{1,h}^2$. Hence $b(\tau, v) \ge c^{-1/2} |v|_{1,h} \|\tau\|_{0,h}$, and the proof of the inf-sup condition is complete. □

The Softening Behavior of Mixed Methods

The (primal) mixed method (5.2) provides a softening of the quadratic form $a(.,.)$. We will study this phenomenon since an analogous procedure has become very popular in computational mechanics during recent years.

Let $u_h \in M_h \subset H_0^1(\Omega)$ and $\sigma_h \in X_h \subset L_2(\Omega)$ be the solution of the mixed method

$$\begin{aligned} (\sigma_h, \tau)_{0,\Omega} - (\tau, \nabla u_h)_{0,\Omega} &= 0 && \text{for all } \tau \in X_h, \\ -(\sigma_h, \nabla v)_{0,\Omega} &= -(f, v)_{0,\Omega} && \text{for all } v \in M_h. \end{aligned} \tag{5.2$_h$}$$

If $E_h := \nabla M_h \subset X_h$, then the first equation implies $\sigma_h = \nabla u_h$, and $(5.2)_h$ is equivalent to the classical treatment of the Poisson equation with the finite element space M_h. This is the uninteresting case.

More interesting is the case $E_h \not\subset X_h$. Let $P_h : L_2(\Omega) \to X_h$ be the orthogonal projector onto X_h. The first equation in $(5.2)_h$ reads

$$\sigma_h = P_h(\nabla u_h)$$

and the second one

$$(P_h \nabla u_h, \nabla v)_{0,\Omega} = (f, v) \quad \text{for all } v \in M_h.$$

This is the weak equation for the relaxed minimum problem

$$\frac{1}{2}\int_\Omega [P_h \nabla v_h]^2 dx - \int_\Omega f v_h \to \min_{v_h \in M_h} . \tag{5.16}$$

Only the part of the gradient that is projected onto X_h contributes to the energy in the variational formulation. The amount of the softening is fixed by the choice of the target space of the projection.

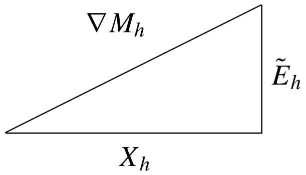


Fig. 36. Projection of the gradient onto X_h in the mixed method and the EAS method, resp.

There is another characterization. The variational equations $(5.2)_h$ can be rewritten in a form which leads to linear equations with a positive definite matrix. We may choose a subspace $\tilde{E}_h$ of the L_2-orthogonal complement of X_h such that

$$\nabla M_h \subset X_h \oplus \tilde{E}_h . \tag{5.17}$$

5.8 Remark. The mixed method $(5.2)_h$ is equivalent to the variational formulation

$$\begin{aligned} (\nabla u_h, \nabla v)_{0,\Omega} &+ (\tilde{\varepsilon}_h, \nabla v)_{0,\Omega} = (f, v)_{0,\Omega} && \text{for all } v \in M_h , \\ (\nabla u_h, \tilde{\eta})_{0,\Omega} &+ (\tilde{\varepsilon}_h, \eta)_{0,\Omega} \quad = 0 && \text{for all } \eta \in \tilde{E}_h , \end{aligned} \tag{5.18}$$

if the space $\tilde{E}_h$ of *enhanced gradients* satisfies the decomposition rule (5.17). Here the relaxation of the variational form and the projector P_h are defined by $\tilde{E}_h$, i.e., by the orthogonal complement of the target space.

The proof of the equivalence follows Yeo and Lee [1996]. Let σ_h, u_h be a solution of $(5.2)_h$. From (5.17) we have a decomposition

$$\nabla u_h = \tilde{\sigma}_h - \tilde{\varepsilon}_h \quad \text{with } \tilde{\sigma}_h \in X_h \text{ and } \tilde{\varepsilon}_h \in \tilde{E}_h .$$

From the first equation in $(5.2)_h$ we conclude that $\nabla u_h - \sigma_h$ is orthogonal to X_h, and the uniqueness of the decomposition implies $\tilde{\sigma}_h = \sigma_h$. When we insert $\sigma_h = \nabla u_h + \tilde{\varepsilon}_h$ in $(5.2)_h$, we get the first equation of the system (5.18). The second one is a reformulation of $\nabla u_h + \tilde{\varepsilon}_h \in X_h$ and $X_h \perp \tilde{E}_h$.

The converse follows from the uniqueness of the solutions. The uniqueness of the solution of (5.18) follows from ellipticity which in turn is given by Problem 5.10. □

We note that in structural mechanics an equivalent concept was derived by Simo and Rifai [1990] and called *the method of enhanced assumed strains (EAS method).*

The stability of the mixed method can be stated in terms of the enhanced elements; cf. Braess [1998]. It also shows that the stability is *not* independent of the choice of the space $\tilde{E}_h$.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

5.9 Lemma. *The spaces X_h and M_h satisfy the inf-sup condition (4.16) with a constant $\beta > 0$ if and only if a strengthened Cauchy inequality*

$$(\nabla v_h, \eta_h)_{0,\Omega} \le \sqrt{1-\beta^2}\, \|\nabla v_h\|_0\, \|\eta_h\|_0 \quad \textit{for all } v_h \in M_h, \eta_h \in \tilde{E}_h \tag{5.19}$$

holds.

Proof. Given $v_h \in M_h$, by the inf-sup condition there is a $\sigma_h \in X_h$ such that $(\nabla v_h, \sigma_h)_0 \ge \beta\|\nabla v_h\|_0$ and $\|\sigma_h\| = 1$. Now for any $\eta_h \in \tilde{E}_h$ we conclude from the orthogonality of X_h and $\tilde{E}_h$ that

$$\|\nabla v_h - \eta_h\|_0 \ge (\nabla v_h - \eta_h, \sigma_h)_0 = (\nabla v_h, \sigma_h)_0 \ge \beta\|\nabla v_h\|_0. \tag{5.20}$$

Since the strengthened Cauchy inequality is homogeneous in its arguments, it is sufficient to verify it for the case $\|\eta_h\|_0 = (1-\beta^2)^{1/2}\|\nabla v_h\|_0$,

$$\begin{aligned} 2(\nabla v_h, \eta_h)_0 &= \|\nabla v_h\|_0^2 + \|\eta_h\|_0^2 - \|\nabla v_h - \eta_h\|_0^2 \\ &\le (1-\beta^2)\|\nabla v_h\|_0^2 + \|\eta_h\|_0^2 = 2(1-\beta^2)^{1/2}\|\nabla v_h\|_0^2\, \|\eta_h\|_0^2\,, \end{aligned}$$

and the proof of (5.19) is complete.

The converse is easily proved by using the decomposition of ∇v_h. □

Problems

5.10 Given $\sigma \in H(\mathrm{div})$, find $\sigma_h \in RT_0$ such that $\|\sigma - \sigma_h\|_0$ is minimal. Characterize σ_h as the solution of a saddle point problem (mixed method).

5.11 Show that the strengthened Cauchy inequality (5.19) is equivalent to the ellipticity property

$$\int_\Omega (\nabla v_h + \eta_h)^2 dx \ge (1-\beta)(|v_h|_1^2 + \|\eta_h\|_0^2) \quad \text{for } v_h \in X_h,\ \eta_h \in \tilde{E}_h\,.$$

Further equivalent properties are presented in Problem V.5.7.

5.12 Define the vectors a_i and b_i in ℓ_2 by

$$(a_i)_j := \begin{cases} 1 & \text{if } j = 2i, \\ 0 & \text{otherwise,} \end{cases} \qquad (b_i)_j := \begin{cases} 1 & \text{if } j = 2i, \\ 2^{-i} & \text{if } j = 2i+1, \\ 0 & \text{otherwise,} \end{cases}$$

and the subspaces $A := \mathrm{span}\{a_i;\ i > 0\}$ and $B := \mathrm{span}\{b_i;\ i > 0\}$. Show that A and B are closed, but that $A + B$ is not. Is there a nontrivial strengthened Cauchy inequality between the spaces A and B?

See also Problem 9.16

§ 6. The Stokes Equation

The *Stokes equation* describes the motion of an incompressible viscous fluid in an n-dimensional domain (with $n = 2$ or 3):

$$\begin{aligned} \Delta u + \operatorname{grad} p &= -f && \text{in } \Omega, \\ \operatorname{div} u &= 0 && \text{in } \Omega, \\ u &= u_0 && \text{on } \partial\Omega. \end{aligned} \tag{6.1}$$

Here $u : \Omega \longrightarrow \mathbb{R}^n$ is the velocity field and $p : \Omega \longrightarrow \mathbb{R}$ is the pressure. Since we are assuming that the fluid is incompressible, $\operatorname{div} u = 0$ when no sources or sinks are present.

In order for a divergence-free flow to exist with given boundary values u_0, by Gauss' integral theorem we must have

$$\int_{\partial\Omega} u_0 \cdot \nu ds = \int_{\partial\Omega} u \cdot \nu ds = \int_{\Omega} \operatorname{div} u \, dx = 0. \tag{6.2}$$

This *compatibility condition* on u_0 is obviously satisfied for *homogeneous* boundary values.

By an appropriate scaling we can assume that the viscosity is 1, which we have already done in writing (6.1).

The given external force field f causes an acceleration of the flow. The pressure gradient gives rise to an additional force which prevents a change in the density. In particular, a large pressure builds up at points where otherwise a source or sink would be created. From a mathematical point of view, the pressure can be regarded as a Lagrange multiplier.

If (6.1) is satisfied for some functions $u \in [C^2(\Omega) \cap C^0(\bar{\Omega})]^n$ and $p \in C^1(\Omega)$, then we call u and p a classical solution of the Stokes problem. Note that (6.1) only determines the pressure p up to an additive constant, which is usually fixed by enforcing the normalization

$$\int_{\Omega} p dx = 0. \tag{6.3}$$

Variational Formulation

In view of the restriction $\operatorname{div} u = 0$, the weak formulation of the Stokes equation (6.1) leads to a saddle point problem. In order to make use of the general framework of §4, we set

$$X = H_0^1(\Omega)^n, \quad M = L_{2,0}(\Omega) := \{q \in L_2(\Omega);\ \textstyle\int_\Omega q\,dx = 0\},$$

$$\begin{aligned} a(u,v) &= \int_\Omega \operatorname{grad} u : \operatorname{grad} v\,dx, \\ b(v,q) &= \int_\Omega \operatorname{div} v\,q\,dx. \end{aligned} \tag{6.4}$$

Here $\operatorname{grad} u : \operatorname{grad} v := \sum_{ij} \frac{\partial u_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}$.

As usual, we restrict our attention to homogeneous boundary conditions, i.e., we assume $u_0 = 0$. Then the saddle point problem becomes: *Find* $(u,p) \in X \times M$ *such that*

$$\begin{aligned} a(u,v) + b(v,p) &= (f,v)_0 \quad && \textit{for all } v \in X, \\ b(u,q) \qquad\quad &= 0 && \textit{for all } q \in M. \end{aligned} \tag{6.5}$$

A solution (u,p) of (6.5) is called a *classical solution* provided $u \in [C^2(\Omega) \cap C^0(\bar{\Omega})]^n$ and $p \in C^1(\Omega)$.

6.1 Remark. For $v \in H_0^1$ and $q \in H^1$, Green's formula gives

$$\begin{aligned} b(v,q) = \int_\Omega \operatorname{div} v\,q\,dx &= -\int_\Omega v \cdot \operatorname{grad} q\,dx + \int_\Gamma v \cdot q\,\nu\,ds \\ &= -\int_\Omega v \cdot \operatorname{grad} q\,dx. \end{aligned} \tag{6.6}$$

Thus, we can regard div and − grad as adjoint operators. Moreover, from (6.6) we see that $b(v,q)$ does not change if we add a constant function to q. Thus, we can identify M with $L_2(\Omega)/\mathbb{R}$. In this quotient space we consider functions in L_2 to be equivalent whenever they differ only by a constant.

6.2 Remark. Every classical solution of the saddle point equation (6.5) is a solution of (6.1).

Proof. Let (u,p) be a classical solution. We split $\phi := \operatorname{div} u \in L_2$ into $\phi = q_0 + const$ with $q_0 \in M$. Since $u \in H_0^1$, combining the formula (6.6) with $v = u$ and $q = 1$ implies $\int_\Omega \operatorname{div} u\,dx = 0$. Substituting q_0 in (6.5), we get

$$\int_\Omega (\operatorname{div} u)^2 dx = b(u,q_0) + const \int_\Omega \operatorname{div} u\,dx = 0.$$

Thus, the flow is divergence-free.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

By Remark 6.1, the first equation in (6.5) can be written in the form

$$(\operatorname{grad} u, \operatorname{grad} v)_{0,\Omega} = (f - \operatorname{grad} p, v)_{0,\Omega} \quad \text{for all } v \in H_0^1(\Omega)^n.$$

Since $u \in C^2(\Omega)^n$, by the theory of scalar equations in Ch. II, §2, it follows that u is a classical solution of

$$\begin{aligned} -\Delta u &= f - \operatorname{grad} p && \text{in } \Omega, \\ u &= 0 && \text{on } \partial\Omega, \end{aligned}$$

and the proof is complete. □

The inf-sup Condition

In order to apply the general theory described in the previous section, let

$$V := \{v \in X;\ (\operatorname{div} v, q)_{0,\Omega} = 0 \quad \text{for all } q \in L_2(\Omega)\}.$$

By Friedrichs' inequality, $|u|_{1,\Omega} = \|\operatorname{grad} u\|_{0,\Omega} = a(u,u)^{1/2}$ is a norm on X. Hence, the bilinear form a is H_0^1-elliptic. Thus it is elliptic not only on the subspace V, but also on the entire space X. This means that we could get by with an even simpler theory than in §4.

In order to ensure the existence and uniqueness of a solution of the Stokes problem, it remains to verify the Brezzi condition.

By the abstract Lemma 4.2, the inf-sup condition can be expressed in terms of properties of the operators B and B'. In the concrete case of the Stokes equation with $b(v,q) = (\operatorname{div} v, q)_{0,\Omega} = -(v, \operatorname{grad} q)_{0,\Omega}$, the conditions are to be understood as properties of the operators div and grad. They are presented in the next two theorems. Their proof is beyond the scope of this book; cf. Duvaut and Lions [1976].

The following result on the divergence is attributed to Ladyšenskaya. Recall that

$$V^\perp := \{u \in X;\ (\operatorname{grad} u, \operatorname{grad} v)_{0,\Omega} = 0 \quad \text{for all } v \in V\} \tag{6.7}$$

is the H^1-orthogonal complement of V.

6.3 Theorem. *Let $\Omega \subset \mathbb{R}^n$ be a bounded connected domain with Lipschitz continuous boundary. Then the mapping*

$$\begin{aligned} \operatorname{div} : V^\perp &\longrightarrow L_{2,0}(\Omega) \\ v &\longmapsto \operatorname{div} v \end{aligned}$$

is an isomorphism. Moreover, for any $q \in L_2(\Omega)$ with $\int_\Omega q\,dx = 0$, there exists a function $v \in V^\perp \subset H_0^1(\Omega)^n$ with

$$\operatorname{div} v = q \quad \textit{and} \quad \|v\|_{1,\Omega} \le c\|q\|_{0,\Omega}, \tag{6.8}$$

where $c = c(\Omega)$ is a constant.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

The inequality (6.10) below is sometimes called *Nečas' inequality*; see Nečas [1965]. We will encounter Nečas' inequality once more in the proof of Korn's inequality in Ch. VI, §3.

6.4 Theorem. *Let $\Omega \subset \mathbb{R}^n$ be a bounded connected domain with Lipschitz continuous boundary.*

(1) The image of the linear mapping

$$\operatorname{grad} : L_2(\Omega) \longrightarrow H^{-1}(\Omega)^n \tag{6.9}$$

is closed in $H^{-1}(\Omega)^n$.

(2) Let $f \in H^{-1}(\Omega)^n$. If

$$\langle f, v\rangle = 0 \quad \textit{for all } v \in V, \tag{6.10}$$

then there exists a unique $q \in L_{2,0}(\Omega)$ with $f = \operatorname{grad} q$.

(3) There exists a constant $c = c(\Omega)$ such that

$$\|q\|_{0,\Omega} \le c(\|\operatorname{grad} q\|_{-1,\Omega} + \|q\|_{-1,\Omega}) \qquad \textit{for all } q \in L_2(\Omega), \tag{6.11}$$

$$\|q\|_{0,\Omega} \le c\, \|\operatorname{grad} q\|_{-1,\Omega} \qquad \textit{for all } q \in L_{2,0}(\Omega). \tag{6.12}$$

6.5 Remark. The inf-sup condition (4.8) for the Stokes problem (6.5) follows from Theorem 6.3 and Theorem 6.4, respectively.

Proof. (1) Given $q \in L_{2,0}$, there exists $v \in H_0^1(\Omega)^n$ that satisfies (6.8). Hence,

$$\frac{(\operatorname{div} v, q)}{\|v\|_1} = \frac{\|q\|_0^2}{\|v\|_1} \ge \frac{\|q\|_0^2}{c\,\|q\|_0} = \frac{1}{c}\|q\|_0,$$

which establishes the Brezzi condition.

(2) For $q \in L_{2,0}$, it follows from (6.12) that

$$\|\operatorname{grad} q\|_{-1} \ge c^{-1}\|q\|_0.$$

By the definition of negative norms, there exists $v \in H_0^1(\Omega)^n$ with $\|v\|_1 = 1$ and

$$(v, \operatorname{grad} q)_{0,\Omega} \ge \frac{1}{2}\|v\|_1 \|\operatorname{grad} q\|_{-1} \ge \frac{1}{2c}\|q\|_0.$$

By (6.6),

$$\frac{b(-v, q)}{\|v\|_1} = (v, \operatorname{grad} q)_{0,\Omega} \ge \frac{1}{2c}\|q\|_0.$$

which establishes the Brezzi condition. □

The properties above are also necessary for the stability of the Stokes problem; see Problem 6.7.

Nearly Incompressible Flows

Instead of directly enforcing that the flow be divergence-free, sometimes a penalty term is added to the variational functional

$$\frac{1}{2}\int [(\nabla v)^2 + t^{-2}(\operatorname{div} v)^2 - 2fv]\,dx \longrightarrow \min!$$

Here t is a parameter. The smaller is t, the more weight is placed on the restriction. In this way a nearly incompressible flow is modeled.

The solution is characterized by the equation

$$a(u, v) + t^{-2}(\operatorname{div} u, \operatorname{div} v)_{0,\Omega} = (f, v)_{0,\Omega} \quad \text{for all } v \in H_0^1(\Omega)^n. \tag{6.13}$$

In order to establish a connection with the standard formulation (6.5), we set

$$p = t^{-2} \operatorname{div} u. \tag{6.14}$$

Now (6.13) together with the weak formulation of (6.14) leads to

$$\begin{aligned} a(u, v) + (\operatorname{div} v, p)_{0,\Omega} &= (f, v)_{0,\Omega} && \text{for all } v \in H_0^1(\Omega)^n, \\ (\operatorname{div} u, q)_{0,\Omega} - t^2(p, q)_{0,\Omega} &= 0 && \text{for all } q \in L_{2,0}(\Omega)^n. \end{aligned} \tag{6.15}$$

Clearly, in comparison with (6.5), (6.15) contains a term which can be interpreted as a penalty term in the sense of §4. By the theory in §4, we know that the solution converges to the solution of the Stokes problem as $t \to 0$.

Problems

6.6 Show that among all representers of $q \in L_2(\Omega)/\mathbb{R}$, the one with the smallest L_2-norm $\|q\|_{0,\Omega} = \inf_{c\in\mathbb{R}} \|q + c\|_{0,\Omega}$ is characterized by $\int_\Omega q\,dx = 0$. [Consequently, $L_2(\Omega)/\mathbb{R}$ and $L_{2,0}(\Omega)$ are isometric.]

6.7 Find a Stokes problem with a suitable right-hand side to show that for every $q \in L_{2,0}(\Omega)$, there exists $u \in H_0^1(\Omega)$ with

$$\operatorname{div} u = q \quad \text{and} \quad \|u\|_1 \le c\|q\|_0,$$

where as usual, c is a constant independent of q.

6.8 If Ω is convex or sufficiently smooth, then one has for the Stokes problem the regularity result

$$\|u\|_2 + \|p\|_1 \le c\|f\|_0; \tag{6.16}$$

see Girault and Raviart [1986]. Show by a duality argument the L_2 error estimate

$$\|u - u_h\|_0 \le ch(\|u - u_h\|_1 + \|p - p_h\|_0). \tag{6.17}$$

§ 7. Finite Elements for the Stokes Problem

In the study of convergence for saddle point problems we assumed that the finite element spaces for velocities and pressure satisfy the inf-sup condition. This raises the question of whether this condition is only needed to get a complete mathematical theory, or whether it plays an essential role in practice.

The answer to this question is given by a well-known finite element method for which the Brezzi condition is violated. Although instabilities had been observed in computations with this element in fluid mechanics, attempts to explain its instable behavior and to overcome it in a simple way mostly proved to be unsatisfactory. The Brezzi condition turned out to be the appropriate mathematical tool for understanding and removing this instability, and it also provided the essential breakthrough in practice. There are very few areas[7] where the mathematical theory is of as great importance for the development of algorithms as in fluid mechanics.

After discussing the instable element mentioned above, we present two commonly used stable elements and another one which is easier to implement. There is also a nonconforming divergence-free element which allows the elimination of the pressure.

An Instable Element

In the Stokes equation (6.1), Δu and $\operatorname{grad} p$ are the terms with derivatives of highest order for the velocity and pressure, respectively. Thus, the orders of the differential operators differ by 1. This suggests the rule of thumb: the degree of the polynomials used to approximate the velocities should be one larger than for the approximation of the pressure. However, this "rule" is not sufficient to guarantee stability – as we shall see.

Because of its simplicity, the so-called Q_1-P_0 element has been popular for a long time. It is a rectangular element which uses bilinear functions for the velocity and piecewise constants for the pressure:

$$\begin{aligned} X_h &:= \{v \in C^0(\bar{\Omega})^2;\ v|_T \in \mathcal{Q}_1 \text{ i.e., bilinear for } T \in \mathcal{T}_h\}, \\ M_h &:= \{q \in L_{2,0}(\Omega);\ q|_T \in \mathcal{P}_0 \quad \text{for } T \in \mathcal{T}_h\}. \end{aligned}$$

[7] There are two comparable situations where purely mathematical considerations have played a major role in the development of methods for differential equations. The approximation properties of the exponential function show that to solve stiff differential equations, we need to use implicit methods. (In particular, parabolic differential equations lead to stiff systems.) For hyperbolic equations, we need to enforce the Courant–Levy condition in order to correctly model the domain of dependence in the discretization.

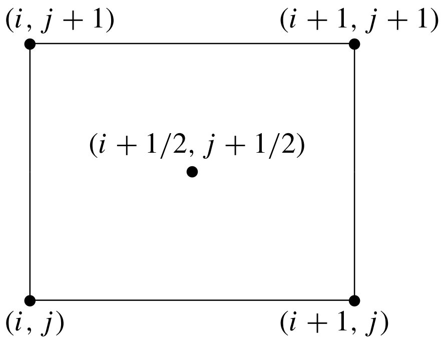


Fig. 37. Numbering of the nodes in the element T_{ij} for the Q_1-P_0 element

One indicator of the instability is the fact that the kernel of $B_h' : M_h \longrightarrow X_h'$ is nontrivial. In order to avoid unnecessary indices when showing this, we will denote the vector components of v by u and w, i.e.,

$$v = \begin{pmatrix} u \\ w \end{pmatrix}.$$

With the numbering shown in Fig. 37, the fact that q is constant and div v is linear implies

$$\begin{aligned}\int_{T_{ij}} q \operatorname{div} v\, dx &= h^2 q_{i+1/2,j+1/2} \operatorname{div} v_{i+1/2,j+1/2} \\ = h^2 q_{i+1/2,j+1/2} \frac{1}{2h} &[u_{i+1,j+1} + u_{i+1,j} - u_{i,j+1} - u_{i,j} \\ &+ w_{i+1,j+1} + w_{i,j+1} - w_{i+1,j} - w_{i,j}]. \end{aligned} \tag{7.1}$$

We now sum over the rectangles. Sorting the terms by grid points is equivalent to partial summation, and we get

$$\int_\Omega q \operatorname{div} v\, dx = h^2 \sum_{i,j} [u_{ij} (\nabla_1 q)_{ij} + w_{ij} (\nabla_2 q)_{ij}], \tag{7.2}$$

where

$$\begin{aligned}(\nabla_1 q)_{i,j} &= \frac{1}{2h}[q_{i+1/2,j+1/2} + q_{i+1/2,j-1/2} - q_{i-1/2,j+1/2} - q_{i-1/2,j-1/2}], \\ (\nabla_2 q)_{ij} &= \frac{1}{2h}[q_{i+1/2,j+1/2} + q_{i-1/2,j+1/2} - q_{i+1/2,j-1/2} - q_{i-1/2,j-1/2}]\end{aligned}$$

are the difference quotients. Since $v \in H_0^1(\Omega)^2$, the summation runs over all interior nodes. Now $q \in \ker(B_h')$ provided

$$\int_\Omega q \operatorname{div} v\ dx = 0 \quad \text{for all } v \in X_h,$$

and thus $\nabla_1 q$ and $\nabla_2 q$ vanish at all interior nodes. This happens if

$$q_{i+1/2,j+1/2} = q_{i-1/2,j-1/2}, \quad q_{i+1/2,j-1/2} = q_{i-1/2,j+1/2}.$$

These equations do not mean that q must be a constant. They only require that

$$q_{i+1/2,j+1/2} = \begin{cases} a & \text{for } i+j \text{ even}, \\ b & \text{for } i+j \text{ odd}. \end{cases}$$

Here the numbers a and b must be chosen so that (6.3) holds, and thus $q \in L_{2,0}(\Omega)$. In particular, a and b must have opposite signs, giving the *checkerboard pattern* shown in Fig. 38. In the following we use ρ to denote the corresponding pressure (up to a constant factor).

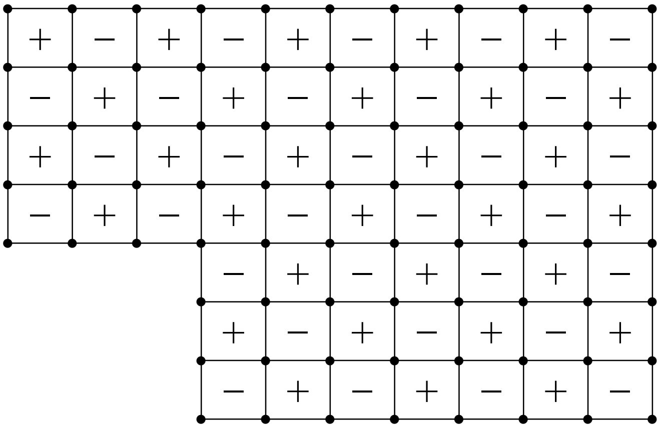


Fig. 38. Checkerboard instability

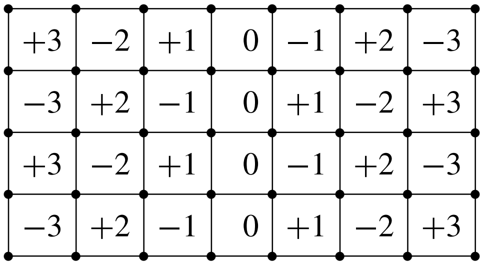
7.1 Remark. The inf-sup condition is an *analytic* property, and should not be interpreted just as a purely *algebraic* one. This fact becomes clear from the modification of Q_1-P_0 elements needed to achieve stability. We start with a reduction of the space M_h so that the kernel of B_h' becomes trivial. Since Ω is assumed to be connected, $\ker B_h' = \text{span}[\rho]$ has dimension 1. The mapping $B_h' : \mathcal{R}_h \longrightarrow X_h'$ is injective on the space

$$\mathcal{R}_h := \rho^\perp = \{q \in M_h;\ (q,\rho)_{0,\Omega} = 0\}.$$

Unfortunately this is not sufficient for full stability.

There is a constant $\beta_1 > 0$ such that

$$\sup_{v \in X_h} \frac{b(v,q)}{\|v\|_1} \ge \beta_1 h \|q\|_0 \quad \text{for } q \in \mathcal{R}_h \tag{7.3}$$



+3	−2	+1	0	−1	+2	−3
−3	+2	−1	0	+1	−2	+3
+3	−2	+1	0	−1	+2	−3
−3	+2	−1	0	+1	−2	+3

Fig. 39. Nearly instable pressure

for the pair X_h, $\mathcal{R}_h$ (see, e.g., Girault and Raviart [1986]). However, the factor h in (7.3) cannot be avoided. Indeed, suppose Ω is a rectangle of width $B = (2n+1)h$ and height $2mh$ with $n \geq 4$. The pressure

$$q^*_{i+1/2,j+1/2} := i\,(-1)^{i+j} \quad \text{for } -n \leq i \leq +n,\ 1 \leq j \leq 2m \tag{7.4}$$

(see Fig. 39) lies in $\mathcal{R}_h$.[8] Then

$$\begin{aligned} \|q^*\|^2_{0,\Omega} &= h^2 \sum_{i=-n}^{+n} \sum_{j=1}^{2m} i^2 = h^2 \frac{1}{3} n(n+1)(2n+1)2m = \frac{1}{3} n(n+1)\mu(\Omega) \\ &\geq \frac{1}{16} B^2 h^{-2} \mu(\Omega). \end{aligned} \tag{7.5}$$

In addition, obviously

$$(\nabla_1 q^*)_{ij} = 0, \qquad (\nabla_2 q^*)_{ij} = (-1)^{i+j} \frac{1}{h}\,.$$

We now return to the node-oriented sum (7.2). We want to reorder it to get an element-oriented sum as in (7.1). To this end, we reassign one-quarter of each summand associated with an interior node to each of the four neighboring squares:

$$\begin{aligned} \int_\Omega q^* \operatorname{div} v\, dx &= h \sum_{i,j} (-1)^{i+j} w_{ij} \\ &= \frac{h}{4} \sum_{i,j} (-1)^{i+j} [w_{i+1,j} - w_{i,j+1} - w_{i+1,j+1} + w_{i,j}]. \end{aligned} \tag{7.6}$$

[8] Similarly, if the width is $B = 2nh$, we set

$$q^*_{i+1/2,j+1/2} = (-1)^{i+j}\Big(i + \frac{1}{2}\Big) \quad \text{for } -n \leq i \leq n-1,\ 1 \leq j \leq 2m.$$

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

For a bilinear function $\hat{w}$ on the reference square $[0,1]^2$, the derivative $\partial_2 \hat{w}$ is linear in ξ. With $\hat{\phi}(\xi) = 2\xi - 1$, simple integration gives

$$\int_{[0,1]^2} \hat{\phi}(\xi)\partial_2 \hat{w}\, d\xi d\eta = \frac{1}{6}[\hat{w}(1,1) - \hat{w}(1,0) - \hat{w}(0,1) + \hat{w}(0,0)].$$

For a bilinear function w, affine transformation to a square T with edges of length h and vertices a, b, c, d (in cyclic order) gives

$$\int_T \phi \partial_2 w\, dxdy = \frac{h}{6}[w(a) - w(b) - w(c) + w(d)].$$

Here ϕ is a function with $\|\phi\|_{0,T}^2 = \mu(T)/3$. Repeating this computation for each square of the partition of Ω and using (7.6), we get

$$\int_\Omega q^* \operatorname{div} v\, dx = \frac{3}{2}\int_\Omega \phi \partial_2 w\, dx. \tag{7.7}$$

Here $\|\phi\|_0^2 = \mu(\Omega)/3$. With the help of the Cauchy–Schwarz inequality, (7.6) and (7.7) imply

$$\begin{aligned}\left|\int_\Omega q^* \operatorname{div} v\, dx\right| &\le \frac{3}{2}\|\phi\|_{0,\Omega}\, \|\partial_2 w\|_{0,\Omega} \le \mu(\Omega)^{1/2}\|v\|_{1,\Omega} \\ &\le 4B^{-1}h\|q^*\|_{0,\Omega}\, \|v\|_{1,\Omega}\,.\end{aligned}$$

In fact,

$$\sup_{v\in X_h} \frac{b(v,q^*)}{\|v\|_{1,\Omega}} \le 4B^{-1}h\|q^*\|_{0,\Omega}\,. \tag{7.8}$$

Thus, the inf-sup condition only holds for some constant depending on h. This clearly shows that *we cannot check the inf-sup condition by merely counting degrees of freedom and using dimensional arguments.*

In order to verify the Brezzi condition with a constant independent of h, we have to further restrict the space $\mathcal{R}_h$. This can be done by combining four neighboring squares into a macro-element. The functions sketched in Fig. 40 form a basis on the level of the macro-elements for the functions which are constant on every small square.

If we eliminate those functions in each macro-element which correspond to the pattern in Fig. 40d, we get the desired stability independent of h; cf. Girault and Raviart [1986], p. 167 or Johnson and Pitkäranta [1982]. However, in doing so, we lose much of the simplicity of the original approximations. Therefore, the stabilized Q_1-P_0 elements are not considered to be competitive.

Specifically, the following pair of subspaces of X_h and M_h is stable:

$$\begin{aligned} \tilde{X}_h \; &:= \{v \in X_h;\; (\operatorname{div} v, q) = 0 \\ &\qquad \text{for all } q \;\text{ spanned by the functions in Fig. 40d on macroelements}\}, \\ \tilde{M}_h \; &:= \{q \in M_h,\;\; \text{spanned by the functions in Fig. 40a–c on macroelements}\}, \end{aligned}$$

The pair $(\tilde{X}_h, \tilde{M}_h)$ is chosen such that the kernel is the same as for the pair (X_h, M_h). However $\tilde{X}_h$ is fixed as a subspace such that a smaller space of Lagrange multipliers is required, and those Lagrange multipliers are eliminated that were an obstacle for the inf-sup condition. Note that $\tilde{X}_h \subset X_{2h}$. Since the element is stable, we can apply Fortin interpolation. If $u \in H^1(\Omega)$ and $\operatorname{div} u = 0$, then

$$\inf_{v_h \in \tilde{V}_h} \|u - v_h\|_1 \le c \inf_{v_{2h} \in X_{2h}} \|u - v_{2h}\|_1 ,$$

and the usual approximation properties can be used.

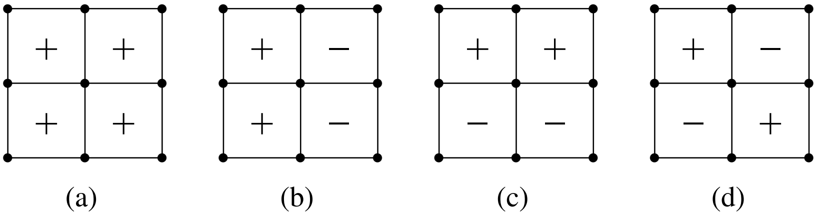


Fig. 40 a–d. Basis functions in M_h for the macro-element

The Taylor–Hood Element

The Taylor–Hood element is an often-used triangular element where the velocity polynomial has a higher degree than the pressure polynomial. The pressure is taken to be continuous:

$$\begin{aligned} X_h \; &:= (\mathcal{M}^2_{0,0})^d \quad\;\; = \{v_h \in C(\bar{\Omega})^d \cap H^1_0(\Omega)^d;\; v_h|_T \in \mathcal{P}_2 \quad \text{for } T \in \mathcal{T}_h\}, \\ M_h \; &:= \mathcal{M}^1_0 \cap L_{2,0} = \{q_h \in C(\Omega) \cap L_{2,0}(\Omega);\; q_h|_T \in \mathcal{P}_1 \quad \text{for } T \in \mathcal{T}_h\}. \end{aligned}$$

Here $\mathcal{T}_h$ is a partition of Ω into triangles. For a proof of the inf-sup condition, see Verfürth [1984] and the book of Girault and Raviart [1986].

Another stable element can be obtained by a simple modification. For the velocities we use piecewise linear functions on the triangulation obtained by dividing each triangle into four congruent subtriangles:

$$\begin{aligned} X_h &:= \mathcal{M}^1_{0,0}(\mathcal{T}_{h/2})^2 = \{v_h \in C(\bar{\Omega})^2 \cap H^1_0(\Omega)^2;\; v_h|_T \in \mathcal{P}_1 \quad \text{for } T \in \mathcal{T}_{h/2}\}, \\ M_h &:= \mathcal{M}^1_0 \cap L_{2,0} = \{q_h \in C(\Omega) \cap L_{2,0}(\Omega);\; q_h|_T \in \mathcal{P}_1 \quad \text{for } T \in \mathcal{T}_h\}. \quad (7.9) \end{aligned}$$

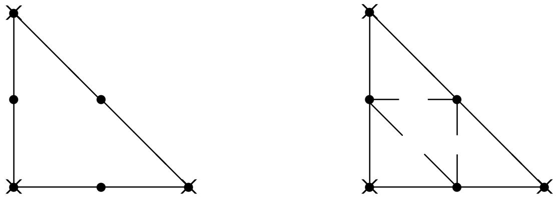


Fig. 41. The Taylor–Hood element and its variant. Here u is given at the nodes (●) and p at the nodes ($\times$)

Thus, the number of degrees of freedom is the same as for the Taylor–Hood element. Often this variant is also called the *modified Taylor–Hood element* in the literature (see Fig. 41).

The approximation properties for the velocities can be obtained directly from results for piecewise quadratic functions. For the approximation of the pressure, we have to verify that the restriction (6.3) to functions with zero integral mean does not reduce the order. Let $\tilde{q}_h$ be an interpolant to $q \in L_{2,0}(\Omega)$. In general, $\int_\Omega \tilde{q}_h dx \neq 0$. By the Cauchy–Schwarz inequality,

$$\left|\int_\Omega \tilde{q}_h dx\right| = \left|\int_\Omega (q - \tilde{q}_h) dx\right| \leq \mu(\Omega)^{1/2} \|q - \tilde{q}_h\|_{0,\Omega}.$$

Thus adding a constant of order $\|q - \tilde{q}_h\|_{0,\Omega}$ gives an approximation in the desired subspace $L_{2,0}(\Omega)$ with the same approximation order.

The MINI Element

One disadvantage of the Taylor–Hood element is that the nodal values of velocity and pressure occur on different triangulations. This complication is avoided with the so-called *MINI element*; see Arnold, Brezzi, and Fortin [1984].

The key idea for the MINI element is to include a *bubble function* in the space X_h for the velocities. Let λ_1, λ_2, and λ_3 be the barycentric coordinates of a triangle (e.g., x_1, x_2, and $(1 - x_1 - x_2)$ in the unit triangle). Then

$$b(x) = \lambda_1 \lambda_2 \lambda_3 \tag{7.10}$$

vanishes on the edges of the triangle. The addition of such a *bubble function* does not affect the continuity of the elements:

$$\begin{aligned} X_h &:= [\mathcal{M}_{0,0}^1 \oplus B_3]^2, \qquad M_h := \mathcal{M}_0^1 \cap L_{2,0}(\Omega) \\ \text{with } B_3 &:= \{v \in C^0(\bar{\Omega});\ v|_T \in \text{span}[\lambda_1\lambda_2\lambda_3] \quad \text{for } T \in \mathcal{T}_h\}. \end{aligned} \tag{7.11}$$

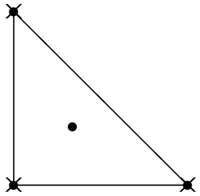


Fig. 42. MINI element. u is given at the nodes ($\bullet$) and p at the nodes ($\times$)

Since the support of a bubble is restricted to the element, we can eliminate the associated variable from the resulting system of linear equations by static condensation. The MINI element requires less computation than the Taylor–Hood element and its variant, but according to many reports, it yields a poorer approximation of the pressure.

7.2 Theorem. *Assume that Ω is convex or has a smooth boundary. Then the MINI element (7.11) satisfies the inf-sup condition.*

Proof. In order to apply Fortin's criterion, we will use arguments introduced in II.7.8 in our treatment of the boundedness of the L_2-projector. We will restrict ourselves to uniform meshes, and note that the extension to shape-regular triangulations is possible by the use of Clément's approximation process.

Let $\pi_h^0 : H_0^1(\Omega) \to \mathcal{M}_{0,0}^1$ be the L_2-projector. From Corollary II.7.8 we know that $\|\pi_h^0 v\|_1 \le c_1\|v\|_1$ and $\|v - \pi_h^0 v\|_0 \le c_2 h\|v\|_1$. Moreover, we fix a linear mapping $\pi_h^1 : L_2(\Omega) \to B_3$ such that

$$\int_T (\pi_h^1 v - v)dx = 0 \quad \text{for each } T \in \mathcal{T}_h. \tag{7.12}$$

We may interpret the map π_h^1 as a process with two steps. First, we apply the L_2-projection onto the space of piecewise constant functions. Afterwards, in each triangle the constant is replaced by a bubble function with the same integral. In this way we get $\|\pi_h^1 v\|_0 \le c_3\|v\|_0$.

Now we set

$$\Pi_h v := \pi_h^0 v + \pi_h^1(v - \pi_h^0 v). \tag{7.13}$$

By construction,

$$\int_T (\Pi_h v - v)dx = \int_T (\pi_h^1 - \mathrm{id})(v - \pi_h^0 v)dx = 0 \quad \text{for each } T \in \mathcal{T}_h. \tag{7.14}$$

The definition of the mapping Π_h is now extended to vector-valued functions. Specifically, each component is to be treated as specified in (7.13).

Since p is continuous, we can apply Green's formula. We recall (7.14), and that the gradient of the pressure is piecewise constant:

$$\begin{aligned} b(v - \Pi_h v, q_h) &= \int_\Omega \operatorname{div}(v - \Pi_h v) q_h dx \\ &= \int_{\partial\Omega} (v - \Pi_h v) \cdot n q_h ds - \int_\Omega (v - \Pi_h v) \cdot \operatorname{grad} q_h dx = 0. \end{aligned}$$

The boundedness of Π_h now follows from (7.12) and an inverse estimate for bubble functions

$$\begin{aligned}\|\Pi_h v\|_1 &\le \|\pi_h^0 v\|_1 + \|\pi_h^1(v - \pi_h^0 v)\|_1\\ &\le c_1\|v\|_1 + c_4 h^{-1}\|\pi_h^1(v - \pi_h^0 v)\|_0\\ &\le c_1\|v\|_1 + c_4 h^{-1} c_3\|v - \pi_h^0 v\|_0\\ &\le c_1\|v\|_1 + c_4 c_3 c_2\|v\|_1.\end{aligned}$$

Now by Fortin's criterion an inf-sup condition holds. □

The Divergence-Free Nonconforming P_1 Element

The Crouzeix–Raviart element plays a special role. We can select from the nonconforming P_1 elements those functions which are piecewise divergence-free, and we can get by without the pressure. We choose

$$\begin{aligned}X_h := \{v \in L_2(\Omega)^2;\ & v|_T \text{ is linear and divergence-free for every } T \in \mathcal{T}_h,\\ & v \text{ is continuous at the midpoints of the triangle edges,}\\ & v = 0 \text{ at the midpoints of the triangle edges in } \partial\Omega\},\end{aligned}$$

i.e., $X_h := \{v \in (\mathcal{M}^1_{*,0})^2;\ \operatorname{div} v = 0 \text{ on every } T \in \mathcal{T}_h\}$. As in the scalar case in §1, we set

$$a_h(u, v) := \sum_{T\in\mathcal{T}_h} \int_T \nabla u \cdot \nabla v\, dx.$$

We seek $u_h \in X_h$ with

$$a_h(u_h, v) = (f, v)_0 \quad \text{for all } v \in X_h.$$

For a convergence proof, see Crouzeix and Raviart [1973].

It is easy to construct a basis for X_h by geometric means. By the Gauss integral theorem, for $v \in X_h$

$$0 = \int_T \operatorname{div}\, v\, dx = \int_{\partial T} v \cdot n\, ds = \sum_{e\in\partial T} v(e_m) n\, \ell(e), \tag{7.15}$$

for every triangle T. Here e_m is the midpoint of the edge e, and $\ell(e)$ is its length.

Since the tangential components do not enter into (7.15), we can prescribe them at the midpoint of each edge. For every interior edge e, we get one basis function $v = v_e$ in X_h with

$$\begin{aligned} v(e_m)\cdot t &= 1,\\ v(e_m)\cdot n &= 0,\\ v(e'_m) \quad &= 0 \quad \text{for } e' \ne e.\end{aligned} \tag{7.16}$$

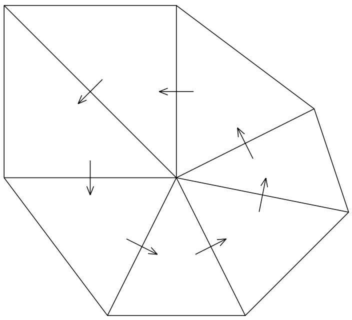


Fig. 43. Basis functions of the nonconforming P_1 element associated with one node. The normal components indicated by arrows have nonzero values

Let p be an arbitrary vertex of a triangle. Suppose the edges connected to p are oriented as follows: If we move around the point p in the mathematically positive direction, we cross the edges in the directions of the normal vectors. Clearly, (7.15) holds if

$$\begin{aligned} v(e_m)\cdot n &= \frac{1}{\ell(e)} && \text{for all edges connected to } p,\\ v(e_m)\cdot n &= 0 && \text{for all other edges,}\\ v(e_m)\cdot t &= 0 && \text{for all edges,} \end{aligned} \tag{7.17}$$

see Fig. 43. The functions in (7.16) and (7.17) are linearly independent, and a dimension count shows that they form a basis if the domain is simply connected. Otherwise an additional basis function with non-local support is required for each hole in the domain.

An analogous quadrilateral element was developed and studied by Rannacher and Turek [1992].

Problems

7.3 For Q_1-P_0 elements, the pressure is

$$q_{i+1/2,j+1/2} = \begin{cases} +(-1)^{i+j} & \text{for } i < i_0,\\ -(-1)^{i+j} & \text{for } i \ge i_0, \end{cases}$$

using the same notation as in (7.5). Here $-n < i_0 < +n$. This gives a checkerboard pattern – up to a shift. Show that

$$\left|\int_\Omega q \operatorname{div} v dx\right| \le c\sqrt{h}\|v\|_{1,\Omega}.$$

Note that in order to get a constant in the inf-sup condition which is independent of h, we have to put increasingly stronger restrictions on M_h as $h \to 0$.

§ 8. A Posteriori Error Estimates

It frequently happens in practical problems that due to the nature of the data in certain subdomains, a solution of a boundary-value problem is less regular. In this case we would like to increase the accuracy of the finite element approximation without using too many additional degrees of freedom. One way to do this is to *adaptively* perform *local grid refinement* in those subdomains where it is needed. We first carry out the finite element calculations on a provisional grid, and then compute an *a posteriori estimate* for the error whose purpose is to indicate what part of the grid induces large errors. Using this information, we then locally refine the grid, and repeat the finite element computation. If necessary, the process can be repeated several times.

To simplify our discussion, we restrict ourselves to the case of the Poisson equation

$$-\Delta u = f \tag{8.1}$$

with homogeneous Dirichlet boundary conditions. Moreover, we consider only conforming elements, although this still involves arguments which are usually associated with the analysis of nonconforming elements. This is why we have not presented a posteriori estimates earlier.

Let $\mathcal{T}_h$ be a shape-regular triangulation. In addition, suppose u_h is a finite element solution lying in $S_h := \mathcal{M}^2_{0,0}$ (or in $\mathcal{M}^1_{0,0}$). Suppose Γ_h is the set of all *inter-element boundaries*, i.e., edges of the triangles $T \in \mathcal{T}_h$ which lie in the interior of Ω.

If we insert u_h into the differential equation in its classical form, we get a residual. Moreover, u_h differs from the classical solution in that $\operatorname{grad} u_h$ has jumps on the edges of elements. Both the area-based residuals

$$R_T := R_T(u_h) := \Delta u_h + f \quad \text{for } T \in \mathcal{T}_h \tag{8.2}$$

and the edge-based jumps

$$R_e := R_e(u_h) := [\![\frac{\partial u_h}{\partial n}]\!] \quad \text{for } e \subset \Gamma_h \tag{8.3}$$

enter either directly or indirectly into many estimators; cf. Remark 8.2. [Note that both the jump $[\![\nabla u_h]\!]$ and the normal direction change if we reverse the orientation of the edge e, but that the product $[\![\frac{\partial u_h}{\partial n}]\!] = [\![\nabla u_h]\!] \cdot n$ remains fixed.] Moreover we need the following notation for the neighborhoods of elements and edges:

$$\begin{aligned} \omega_T &:= \bigcup \{T' \in \mathcal{T}_h;\ T \text{ and } T' \text{ have a common edge or } T' = T\}, \\ \omega_e &:= \bigcup \{T' \in \mathcal{T}_h;\ e \subset \partial T'\}. \end{aligned} \tag{8.4}$$

There are five popular approaches to building a posteriori estimators.

1. Residual estimators.

We bound the error on an element T in terms of the size of the residual R_T and the jumps R_e on the edges $e \subset \partial T$. These estimators are due to Babuška and Rheinboldt [1978a].

2. Estimators based on local Neumann problems

On every triangle T we solve a local variational problem which is a discrete analog of

$$\begin{aligned} -\Delta z &= R_T \quad \text{in } T, \\ \frac{\partial z}{\partial n} &= R_e \quad \text{on } e \subset \partial T. \end{aligned} \tag{8.5}$$

We choose the approximating space to contain polynomials whose degrees are higher than those in the underlying finite element space. These estimators are obtained using the energy norm $\|z\|_{1,T}$, and are due to Bank and Weiser [1985]. See also the comment before Theorem 9.5.

3. Estimators based on a local Dirichlet problem.

For every element T, we solve a variational problem on the set ω_T:

$$\begin{aligned} -\Delta z &= f \quad \text{in } \omega_T, \\ z &= u_h \quad \text{on } \partial\omega_T. \end{aligned} \tag{8.6}$$

Again, we expand the approximating space to include polynomials of higher degree than in the actual finite element space. Following Babuška and Rheinboldt [1978b], the norm of the difference $\|z - u_h\|_{1,\omega_T}$ provides an estimator.

4. Estimators based on averaging.

We construct a continuous approximation σ_h of ∇u_h by a two-step process. At every node of the triangulation, let σ_h be a weighted average of the gradients ∇u_h on the neighboring triangles, where the weight is proportional to the areas of the triangles. We then extend σ_h to the whole element by linear interpolation. Then following Zienkiewicz and Zhu [1987], we use the difference between ∇u_h and σ_h as an estimator. An analysis without restrictive assumptions was done by Rodriguez [1994] and by Carstensen and Bartels [2002].

5. Hierarchical estimators.

In principle the difference from a finite element approximation on an expanded space is estimated. The difference can be estimated by using a strengthened Cauchy inequality; see Deuflhard, Leinen, and Yserentant [1989]. The procedure fits into Carl Runge's old and general concept *(Runge's rule)*. The error of a numerical result is estimated by comparing it with the result of a more accurate formula.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

An estimator based on different ideas will be presented in the next §. Moreover, *goal-oriented estimators* are the topic of the book by Bangerth and Rannacher [2003]. Their aim is a small error in a given functional of the solution rather than a small norm of the error.

To get started, following Dörfler [1996] we first show how to extract a part of the expression (8.2) which can be determined already before computing the finite element solution. Let

$$f_h := P_h f \in S_h \tag{8.7}$$

be the L_2-projection of f onto S_h. Since $(f - f_h, v_h)_{0,\Omega} = 0$ for $v_h \in S_h$, the variational problems corresponding to f and f_h lead to the same finite element approximation in S_h. Thus, the a priori computable quantity

$$h_T \|f - f_h\|_{0,T} \tag{8.8}$$

appears in many estimates. As usual, h_T denotes the diameter of T. Similarly, h_e is the length of e. The term $\|f - f_h\|_{0,T}$ and analogous expressions are called *data oscillation*. In particular, clearly

$$\|\Delta u_h + f\|_{0,T} \le \|\Delta u_h + f_h\|_{0,T} + \|f - P_h f\|_{0,T}. \tag{8.9}$$

As an alternative to (8.7), we can define and use f_h as the projection onto piecewise constant functions.

Residual Estimators

To get residual estimators, we use the functions introduced in (8.2) and (8.3) to compute the local quantities

$$\eta_{T,R} := \Big\{h_T^2 \|R_T\|_{0,T}^2 + \frac{1}{2}\sum_{e\subset\partial T} h_e \|R_e\|_{0,e}^2\Big\}^{1/2} \quad \text{for } T \in \mathcal{T}_h. \tag{8.10}$$

Summing the squares over all triangles, we get a global quantity:

$$\eta_R := \Big\{\sum_{T\in\mathcal{T}_h} h_T^2 \|R_T\|_{0,T}^2 + \sum_{e\subset\Gamma_h} h_e \|R_e\|_{0,e}^2\Big\}^{1/2}. \tag{8.11}$$

8.1 Theorem. *Let $\mathcal{T}_h$ be a shape-regular triangulation with shape parameter κ. Then there exists a constant $c = c(\Omega, \kappa)$ such that*

$$\|u - u_h\|_{1,\Omega} \le c\,\eta_R = c\Big\{\sum_{T\in\mathcal{T}_h} \eta_{T,R}^2\Big\}^{1/2} \tag{8.12}$$

and

$$\eta_{T,R} \le c\Big\{\|u - u_h\|_{1,\omega_T}^2 + \sum_{T'\subset\omega_T} h_T^2 \|f - f_h\|_{0,T'}^2\Big\}^{1/2} \tag{8.13}$$

for all $T \in \mathcal{T}_h$.

The upper bound (8.12) means that the estimator η_R is *reliable* and the lower bound (8.13) that it is also *efficient*.

Proof of the upper estimate (8.12). We start by using a duality argument to find

$$|u - u_h|_1 = \sup_{|w|_1=1,\, w\in H_0^1} (\nabla(u-u_h), \nabla w)_0. \tag{8.14}$$

We make use of the following formula which also appeared in establishing the Céa Lemma:

$$(\nabla(u-u_h), \nabla v_h)_0 = 0 \quad \text{for } v_h \in S_h. \tag{8.15}$$

We now consider the functional ℓ corresponding to (8.14), apply Green's formula, and insert the residuals (8.2) and (8.3):

$$\begin{aligned} \langle \ell, w\rangle &:= (\nabla(u-u_h), \nabla w)_{0,\Omega} \\ &= (f, w)_{0,\Omega} - \sum_T (\nabla u_h, \nabla w)_{0,T} \\ &= (f, w)_{0,\Omega} - \sum_T \Big\{ (-\Delta u_h, w)_{0,T} + \sum_{e\subset\partial T} (\nabla u_h \cdot n, w)_{0,e} \Big\} \\ &= \sum_T (\Delta u_h + f, w)_{0,T} + \sum_{e\subset\Gamma_h} \Big([\![\frac{\partial u_h}{\partial n}]\!], w \Big)_{0,e} \\ &= \sum_T (R_T, w)_{0,T} + \sum_{e\subset\Gamma_h} (R_e, w)_{0,e}. \end{aligned} \tag{8.16}$$

By Clément's results on approximation, cf. II.6.9, for given $w \in H_0^1(\Omega)$ there exists an element $I_h w \in S_h$ with

$$\|w - I_h w\|_{0,T} \le c h_T \|\nabla w\|_{0,\tilde{\omega}_T} \quad \text{for all } T \in \mathcal{T}_h, \tag{8.17}$$

$$\|w - I_h w\|_{0,e} \le c h_e^{1/2} \|\nabla w\|_{0,\tilde{\omega}_T} \quad \text{for all } e \subset \Gamma_h. \tag{8.18}$$

Here $\tilde{\omega}_T$ is the neighborhood of T specified in (II.6.14) which is larger than ω_T. Since the triangulations are assumed to be shape regular, $\bigcup\{\tilde{\omega}_T;\ T \in \mathcal{T}_h\}$ covers Ω only a finite number of times. Hence, (8.15) implies

$$\begin{aligned} \langle \ell, w\rangle &= \langle \ell, w - I_h w\rangle \\ &\le \sum_T \|R_T\|_{0,T}\, \|w - I_h w\|_{0,T} + \sum_{e\subset\Gamma_h} \|R_e\|_{0,e}\, \|w - I_h w\|_{0,e} \\ &\le c \sum_T h_T \|R_T\|_{0,T}\, |w|_{1,T} + c \sum_{e\subset\Gamma_h} h_e^{1/2} \|R_e\|_{0,e}\, |w|_{1,\omega_e} \\ &\le c \sum_T \eta_{T,R} |w|_{1,T} \le c\, \eta_R\, |w|_{1,\Omega}. \end{aligned} \tag{8.19}$$

The last inequality follows from the Cauchy–Schwarz inequality for finite sums. Combining (8.18) and (8.19) with Friedrichs' inequality and the duality argument (8.14), we get the global upper error bound (8.12). □

8.2 Remarks. (1) The general procedure that led to Theorem 8.1 is also used for deriving residual error estimators for other finite element discretizations. Here an isomorphism $L : H_0^1(\Omega) \to H^{-1}(\Omega)$ was associated to the given variational problem in §3, and we have

$$u - u_h = L^{-1}\ell$$

with ℓ given by (8.16). The representation (8.16) of the H^{-1} function ℓ in terms of integrals enables us to establish computable bounds of $\|\ell\|_{-1}$. When a posteriori error estimates for saddle point problems are studied, the isomorphism L in Theorem 4.3 and the residues of the corresponding equations are used in an analogous way; see Hoppe and Wohlmuth [1997].

(2) When $\Delta u = 0$ is numerically solved with P_1 elements, we have *piecewise* $\Delta u_h = 0$ and the special case where the complete area-based estimator R_T vanishes. As observed by Carstensen and Verfürth [1999], this term is dominated by the edge term and the data oscillation also for $\Delta u \neq 0$ provided that the grids have a certain regularity. Specifically, they showed H^1-stability of the L_2-orthogonal projector Q_h under weaker assumptions than assumed in Corollary II.7.8 and Lemma II.7.9. Note that $(R_T, w - Q_h w)_{0,\Omega} = (f, w - Q_h w)_{0,\Omega} = (f - f_h, w - Q_h w)_{0,\Omega}$ if $f_h \in \mathcal{M}_{0,0}^1$, and this volume term is indeed bounded by the data oscillation.

(3) If $|u - u_h|_1$ is estimated in (8.12) instead of $\|u - u_h\|_1$, then the constant c depends only on the shape parameter κ and not on Ω since the Clément interpolation is a local process. The dependence on Ω enters merely due to Friedrichs' inequality at the end of the proof of Theorem 8.1.

Lower Estimates

The lower estimate (8.13) provides information on local properties of the discretization. It can be obtained using test functions with local support. The following cutoff functions ψ_T and ψ_e are essential tools: ψ_T is the well-known bubble function associated with the triangle T, so that

$$\psi_T \in B_3,\ \operatorname{supp}\psi_T = T,\ 0 \le \psi_T \le 1 = \max\psi_T. \tag{8.20}$$

ψ_e has support on a pair of neighboring triangles sharing the edge e, and consists of quadratic polynomials joined together continuously so that

$$\psi_e \in \mathcal{M}_0^2,\ \operatorname{supp}\psi_e = \omega_e;\ 0 \le \psi_e \le 1 = \max\psi_e. \tag{8.21}$$

We also need a mapping $E : L_2(e) \to L_2(\omega_e)$ which extends any function defined on an edge e to the pair of neighboring triangles making up ω_e. We take

$$E\sigma(x) := \sigma(x') \text{ in } T,\ \text{ if } x' \in e \text{ is the point in } e \text{ with } \lambda_j(x') = \lambda_j(x).$$

Here λ_j is one of the two barycentric coordinates in T which are not constant on the edge e; see Fig. 44.

https://doi.org/10.1017/CBO9780511618635.006 Published online by Cambridge University Press

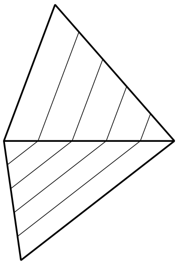


Fig. 44. Level curves for the extension of a function from e to ω_e. The level curves in the triangle lying below [1,2] are given by $\lambda_1 = const$, where λ_1 is the barycentric coordinate w.r.t. the point 1. The level curves in the triangle lying on the opposite side [1,2] are similarly described using λ_2.

8.3 Lemma. *Let $\mathcal{T}_h$ be a shape-regular triangulation. Then there exists a constant c which depends only on the shape parameter κ such that*

$$\begin{aligned}
\|\psi_T v\|_{0,T} &\le \|v\|_{0,T} && \textit{for all } v \in L_2(T), && (8.22)\\
\|\psi_T^{1/2} p\|_{0,T} &\ge c\|p\|_{0,T} && \textit{for all } p \in \mathcal{P}_2, && (8.23)\\
\|\nabla(\psi_T p)\|_{0,T} &\le c h_T^{-1}\|\psi_T p\|_{0,T} && \textit{for all } p \in \mathcal{P}_2, && (8.24)\\
\|\psi_e^{1/2}\sigma\|_{0,e} &\ge c\|\sigma\|_{0,e} && \textit{for all } \sigma \in \mathcal{P}_2, && (8.25)\\
ch^{1/2}\|\sigma\|_{0,e} \le \|\psi_e E\sigma\|_{0,T} &\le c h_e^{1/2}\|\sigma\|_{0,e} && \textit{for all } \sigma \in \mathcal{P}_2, && (8.26)\\
\|\nabla(\psi_e E\sigma)\|_{0,T} &\le c h_T^{-1}\|\psi_e E\sigma\|_{0,T} && \textit{for all } \sigma \in \mathcal{P}_2, && (8.27)
\end{aligned}$$

for all $T \in \mathcal{T}_h$ and all $e \subset \partial T$.

The inequality (8.22) follows directly from $0 \le \psi_T \le 1$. For a fixed reference triangle, the others are obvious because of the finite dimensionality of $\mathcal{P}_2$. The assertions for arbitrary triangles then follow from the usual scaling argument. Details have been elaborated by Verfürth [1994] and by Ainsworth and Oden [2000]. □

Proof of (8.13). Let $T \in \mathcal{T}_h$. In view of (8.9), in analogy with (8.2) we introduce

$$R_{T,\mathrm{red}} := \Delta u_h + f_h. \tag{8.28}$$

By construction, $R_{T,\mathrm{red}} \in \mathcal{P}_2$. Let

$$w := w_T := \psi_T \cdot R_{T,\mathrm{red}}.$$

Then (8.16), (8.23), and $\operatorname{supp} w = T$ imply

$$\begin{aligned}
c^{-1}\|R_{T,\mathrm{red}}\|_{0,T}^2 &\le \|\psi_T^{1/2} R_{T,\mathrm{red}}\|_{0,T}^2\\
&= (R_{T,\mathrm{red}}, w)_{0,T}\\
&= (R_T, w)_{0,T} + (f - f_h, w)_{0,T}\\
&= \langle \ell, w\rangle + (f - f_h, w)_{0,T}\\
&\le |u - u_h|_{1,T} \cdot |w|_{1,T} + \|f - f_h\|_{0,T}\|w\|_{0,T}.
\end{aligned}$$

Obviously (8.22) yields $\|w\|_{0,T} \leq \|R_{T,\text{red}}\|_{0,T}$. Now using Friedrichs' inequality and the inverse inequality (8.24), after dividing by $\|R_{T,\text{red}}\|_{0,T}$ we obtain

$$\|R_{T,\text{red}}\|_{0,T} \leq c(h_T^{-1}\|u - u_h\|_{1,T} + \|f - f_h\|_{0,T}).$$

By (8.9), this gives

$$h_T\|R_T\|_{0,T} \leq c(\|u - u_h\|_{1,T} + h_T\|f - f_h\|_{0,T}). \tag{8.29}$$

We show now that the edge-based terms in the error estimator can be treated in a similar way. Let $e \subset \Gamma_h$. We consider the extension of R_e and define

$$w := w_e := \psi_e \cdot E(R_e).$$

In particular, $\operatorname{supp} w = \omega_e$, and $R_e \in \mathcal{P}_2(e)$. Using (8.16), (8.25), we have

$$\begin{aligned} c\|R_e\|_{0,e}^2 &\leq \|\psi_e^{1/2} R_e\|_{0,e}^2 \\ &= (R_e, w)_{0,e} = \langle \ell, w\rangle - \sum_{T' \subset \omega_e} (R_{T'}, w)_{0,T'} \\ &\leq |u - u_h|_{1,e}|w|_{1,\omega_e} + \sum_{T' \subset \omega_e} \|R_{T'}\|_{0,T'}\|w\|_{0,T'}. \end{aligned} \tag{8.30}$$

From (8.27) it follows that $|w|_{1,T'} \leq ch_{T'}^{-1}\|w\|_{0,T'}$, while (8.26) yields the bound $\|w\|_{0,T'} \leq h_e^{1/2}\|R_e\|_{0,e}$. Thus (8.30) leads to

$$\|R_e\|_{0,e} \leq ch_e^{-1/2}|u - u_h|_{1,\omega_e} + ch_e^{1/2} \sum_{T' \subset \omega_e} \|R_{T'}\|_{0,T'},$$

which combined with (8.29) gives

$$h_e^{1/2}\|R_e\|_{0,e} \leq c|u - u_h|_{1,\omega_e} + \sum_{T' \subset \omega_e} h_{T'}\|f - f_h\|_{0,T'}. \tag{8.31}$$

Combining (8.29) and (8.31) and observing that $\omega_T = \bigcup\{\omega_e;\, e \subset \partial T\}$, we get the desired assertion (8.13). □

Remark on Other Estimators

As noted in Remark 8.2(1), the residual estimator gives rise to a convenient bound of $\|L(u - u_h)\|_{-1} = \|\ell\|_{-1}$. It makes use of the fact that $L : H^1(\Omega) \to H^{-1}(\Omega)$, given by the variational problem, is an isomorphism. Therefore a large gap is expected between the lower and upper bounds when the condition number of L is large, i.e., when

$$\frac{\sup\{\|Lu\|_{-1};\ \|u\|_1 = 1\}}{\inf\{\|Lu\|_{-1};\ \|u\|_1 = 1\}} \gg 1.$$

In this case it is assumed to be better to compute $\|\tilde{L}^{-1}\ell\|_1$ instead of $\|\ell\|_{-1}$, where $\tilde{L}^{-1}$ is an approximate inverse. Such an approximate inverse is implicitly used with hierarchical estimators. Although (8.5) and (8.6) define also approximate inverses, these (local) approximate inverses provide estimators that are equivalent to residual estimators; see Verfürth [1996]. Thus it is not clear whether they are more efficient than residual estimators in the case of large condition numbers.

Babuška, Durán, and Rodríguez [1992] show that the efficiency of estimators can be much worse for unstructured grids than for regular ones.

Local Mesh Refinement and Convergence

In finite element computations using local grid refinement, we generally start with a coarse grid and continue to refine it successively until the estimator $\eta_{T,R}$ is smaller than a prescribed bound for all elements T. In particular, those elements where the estimators give large values are the ones which are refined. The geometrical aspects have already been discussed in Ch. II, §8.

This leads to a triangulation for which the estimators have approximately equal values in all triangles. Numerical results obtained using this simple idea are quite good.

The above approach can be justified heuristically. Suppose the domain Ω in d-space is divided into m (equally large) subdomains where the derivatives of the solution have different size. Suppose an element with mesh size h_i in the i-th subdomain contributes $c_i h_i^\alpha$ to the error, where $\alpha > d$. The subdomains involve different factors c_i, but are all associated with the same exponent α. If the i-th subdomain is divided into n_i parts, then $h_i = n_i^{1/d}$, and the total error is of order

$$\sum_i n_i c_i h_i^\alpha = \sum_i c_i h_i^{\alpha-d}. \tag{8.32}$$

Our aim is to minimize the expression (8.32) subject to $\sum_i n_i = \sum_i h_i^{-d} = \text{const}$. The optimum is a stationary point of the Lagrange function

$$\mathcal{L}(h,\lambda) := \sum_i c_i h_i^{\alpha-d} + \lambda\Big(\sum_i h_i^{-d} - \text{const}\Big). \tag{8.33}$$

If we relax the requirement that the n_i be integers, then we can find the optimum by differentiating (8.33), which leads to $c_i h_i^\alpha = \frac{d\lambda}{\alpha-d}$. This is just the condition that the contributions of all elements be equal. □

The convergence of the finite element computations with the refinement strategy above is not obvious. A first proof was established by Dörfler [1996], and it was extended later by Morin, Nochetto, and Siebert [2002]. The general scheme that is also encountered in the analysis of other variational problems is as follows. Let $\mathcal{T}_h$ be a refinement of $\mathcal{T}_H$ which is not required to be as fine as $\mathcal{T}_{H/2}$. Let Γ_H, Γ_h denote the associated sets of interior edges and let u_H, u_h be the solutions with linear finite elements on these triangulations. The dominance of the edge terms of the residual estimator implies that

$$|u - u_H|_1^2 \le c \sum_{e \subset \Gamma_H} h_e \, \Big\| \Big[\!\!\Big[\frac{\partial u_h}{\partial n} \Big]\!\!\Big] \Big\|_0^2 + \text{ higher order terms.}$$

The Galerkin orthogonality (II.4.7) yields

$$|u - u_h|_1^2 = |u - u_H|_1^2 - |u_h - u_H|_1^2 .$$

The essential step is the *discrete local efficiency*,

$$|u_h - u_H|_1^2 \ge c^{-1} \sum_{e \subset \Gamma_H \to \Gamma_h} h_e \, \Big\| \Big[\!\!\Big[\frac{\partial u_h}{\partial n} \Big]\!\!\Big] \Big\|_0^2 + \text{ higher order terms,} \tag{8.34}$$

where the sum runs over those edges of Γ_H that are refined for getting Γ_h. The three inequalities imply that

$$|u - u_h|_1 \le c' |u - u_H|_1 + \text{ higher order terms}$$

with $c' < 1$. Thus convergence is guaranteed.

Optimal convergence rates were established in the framework of wavelets by Binev, Dahmen, and De Vore [2004]. Their procedure contains not only refinements, but also coarsenings. The latter can be dropped due to Gantumur, Harbrecht, and Stevenson [2007].

Problems

8.4 Most of the estimates in Lemma 8.3 refer to quadratic polynomials. Consider the generalization to polynomials of degree k. Show by a simple argument that the constant c cannot be independent of k.

§ 9. A Posteriori Error Estimates via the Hypercircle Method

The a posteriori error estimators in the preceding section provide a bound of the error up to a generic constant; cf. Theorem 8.1. The theorem of Prager and Synge (Theorem 5.1) admits the computation of an error bound *without such a generic constant*. The essential idea is that a comparison of an approximate solution of the primal variational problem with a feasible function of the dual mixed problem (cf. $(5.5)_v$) yields an estimate. An elaborated theory was provided by Neittaanmäki and Repin [2004]. Since Theorem 5.1 looks like Pythagoras' rule in an infinite dimensional space, the term *hypercircle method* is frequently found.

We consider the Poisson equation (Example II.2.10) as the simplest case. Given a finite element solution u_h of the primal problem, for applying Theorem 5.1 an auxiliary function $\sigma \in H(\mathrm{div})$ with

$$\operatorname{div} \sigma = -f \tag{9.1}$$

is required. Following Braess and Schöberl [2006] we demonstrate that such a function σ can be constructed by the solution of cheap local problems based on the knowledge of u_h.

Let $\mathcal{T}_h$ be a triangulation of a polygonal domain $\Omega \subset \mathbb{R}^2$. A crucial step is the evaluation of the error with respect to the solution for the differential equation with a right-hand side f_h that is piecewise constant on the triangulation. We know from (8.8) that there is only an extra term $ch\|f - f_h\|$ if we approximate a given function $f \in L_2(\Omega)$ by $f_h \in \mathcal{M}^0$. Obviously, the extra term is a term of higher order. [We suggest for a first reading to assume that f is piecewise constant. In this case the data oscillation vanishes and (9.5) below can be replaced by a simpler expression.]

The mixed method by Raviart–Thomas yields a bound that is optimal in a certain sense.

9.1 Lemma. *Let $u_h \in \mathcal{M}_0^1(\mathcal{T}_h)$ and $f_h \in \mathcal{M}^0(\mathcal{T}_h)$. Moreover, let (σ_h, w_h) be a solution of the mixed variational problem with the Raviart–Thomas element of lowest order in $RT_0(\mathcal{T}_h) \times \mathcal{M}^0(\mathcal{T}_h)$. Then*

$$\|\nabla u_h - \sigma_h\|_0 = \min \Big\{ \|\nabla u_h - \tau_h\|_0 ;\ \tau_h \in RT_0(\mathcal{T}_h),\ \operatorname{div} \tau_h + f_h = 0 \Big\}. \tag{9.2}$$

Proof. The Lagrange function for the minimization problem (9.2) is

$$\mathcal{L}(\tau, v) = \frac{1}{2}\|\tau\|_0^2 - (\nabla u_h, \tau)_0 + (v, \operatorname{div} \tau + f_h)_0,$$

with v being the Lagrange multiplier. Note that $\operatorname{div}\tau_h + f_h \in \mathcal{M}^0$ for all $\tau_h \in RT_0$. Thus the minimizing function σ_h and the Lagrange multiplier w_h are characterized by the equations

$$\begin{aligned}(\sigma_h, \tau)_0 + (\operatorname{div}\tau, w_h)_0 &= (\nabla u_h, \tau)_0 && \text{for all } \tau \in RT_0,\\ (\operatorname{div}\sigma_h, v)_0 &= -(f_h, v)_0 && \text{for all } v \in \mathcal{M}^0.\end{aligned} \tag{9.3}$$

By Green's formula we obtain $(\nabla u_h, \tau)_0 = -(u_h, \operatorname{div}\tau)_0$ since the boundary terms vanish. Let Q_h be the L_2 projector onto $\mathcal{M}^0$. Then $(u_h, \operatorname{div}\tau)_0 = (Q_h u_h, \operatorname{div}\tau)_0$ for all $\tau \in RT_0$, and (9.3) can be rewritten

$$\begin{aligned}(\sigma_h, \tau)_0 + (\operatorname{div}\tau, w_h + Q_h u_h)_0 &= 0 && \text{for all } \tau \in RT_0,\\ (\operatorname{div}\sigma_h, v)_0 &= -(f_h, v)_0 && \text{for all } v \in \mathcal{M}^0.\end{aligned}$$

The pair $(\sigma_h, w_h + Q_h u_h)$ is a solution of the mixed method with the Raviart–Thomas element.

Finally, we note that $(9.3)_2$ implies that $\operatorname{div}\sigma_h = -f_h$ holds in the strong sense. Indeed, the expressions on both sides belong to $\mathcal{M}^0$, and the relation $(9.3)_2$ is tested with functions in the same space. □

Of course, the numerical solution of the mixed method is too expensive when only an error estimate is desired. An approximation $\sigma \in RT_0$ will be sufficient and will be constructed from u_h by a simple postprocess.

We consider the space of *broken Raviart–Thomas functions*

$$RT_{-1} := \left\{\tau \in L_2(\Omega)^2;\ \tau|_T = \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T \begin{pmatrix} x \\ y \end{pmatrix},\ a_T, b_T, c_T \in \mathbb{R} \text{ for } T \in \mathcal{T}_h\right\}.$$

The normal components are not required to be continuous, and $RT_0 = RT_{-1} \cap H(\operatorname{div})$. The degrees of freedom of the finite element functions in RT_{-1} are the normal components on the edges, but the values at the two sides of interior edges may differ. Therefore, two degrees of freedom are associated to each inner edge; see Fig. 45 below.

Note that ∇u_h and σ_h belong to RT_{-1}. Moreover, in each triangle $\operatorname{div}\nabla u_h = 0$. (We do not consider $\operatorname{div}\nabla u_h$ as a global function on Ω.) We construct a suitable σ by determining a suitable $\sigma^\Delta := \sigma - \nabla u_h$. *Find* $\sigma^\Delta \in RT_{-1}$ *such that*

$$\begin{aligned}\operatorname{div}\sigma^\Delta &= -f_h && \textit{in each } T \in \mathcal{T}_h,\\ [\![\sigma^\Delta \cdot n]\!] &= -[\![\nabla u_h \cdot n]\!] && \textit{on each interior edge } e.\end{aligned} \tag{9.4}$$

Let z be a node of the triangulation. The construction will be performed on patches

$$\omega_z := \bigcup \{T;\ z \in \partial T\}.$$

As a preparation we have

9.2 Theorem. *Let $z \in \Omega\backslash\partial\Omega$ be a node of the triangulation, and let $\psi_z \in \mathcal{M}_0^1$ be the nodal base function with $\psi_z(z) = 1$ and $\psi_z(x) = 0$ for $x \in \Omega\backslash\omega_z$. Then*

$$\frac{1}{2}\sum_{e\subset\omega_z}\int_e [\![\frac{\partial u_h}{\partial n}]\!]\,ds = \sum_{T\subset\omega_z}\int_T f\psi_z\,dx. \tag{9.5}$$

Proof. Since u_h is the finite element solution in $\mathcal{M}_0^1$, we have

$$\int_{\omega_z} \nabla u_h \nabla\psi_z dx = \int_{\omega_z} f\psi_z dx. \tag{9.6}$$

We apply partial integration to the left-hand side of (9.6) and note that $\partial u_h/\partial n$ is constant on the edges. Since the mean value of ψ_z on the edges is $1/2$, we obtain

$$\begin{aligned}\int_{\omega_z} \nabla u_h \nabla\psi_z dx &= \sum_{T\subset\omega_z}\int_{\partial T}\frac{\partial u_h}{\partial n}\psi_z dx \\ &= \sum_{e\subset\omega_z}\int_e [\![\frac{\partial u_h}{\partial n}]\!]\psi_z ds = \frac{1}{2}\sum_{e\subset\omega_z}\int_e [\![\frac{\partial u_h}{\partial n}]\!]\,ds.\end{aligned}$$

We insert the last expression in (9.6), and the proof is complete. □

Note that the integrals $\int_T f\psi_z\,dx$ are evaluated in the finite element computations when the finite element equations are put in the matrix-vector form (II.4.5). If f is constant on T (e.g., if $f = f_h$), then the integral above equals $\frac{1}{3}f|T| = \frac{1}{3}R_T|T|$, which makes $(9.7)_1$ simpler.

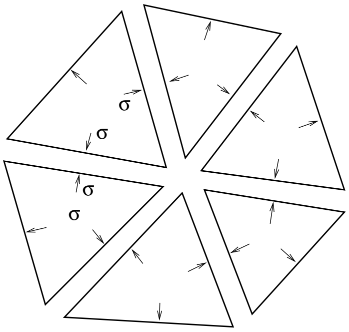


Fig. 45. Fluxes in a patch around a vertex z. $\sigma_{i,r}$ and $\sigma_{i,l}$ are the normal components of the fluxes that leave the triangle T_i on the right and left side, respectively. The triangles are enumerated counterclockwise and $e_i = \partial T_i \cap \partial T_{i+1}$ (with indices modulo the number of triangles)

Given a node z, the following algorithm yields a $\sigma_{\omega_z} \in RT_{-1}$ with support in ω_z such that

$$\begin{aligned} \operatorname{div}\sigma_{\omega_z} &= -\frac{1}{|T|}\int_T f\psi_z\,dx && \text{in each } T\subset\omega_z, \\ [\![\sigma_{\omega_z}\cdot n]\!] &= -(1/2)\,[\![\nabla u_h\cdot n]\!] && \text{on each edge } e\subset\omega_z\,, \\ \sigma_{\omega_z}\cdot n &= 0 && \text{on } \partial\omega_z\,. \end{aligned} \tag{9.7}$$

The notation for the algorithm is specified in Fig. 45.

9.3 Algorithm.

```
set $\sigma_{1,r} = 0$;
for $i = 1, 2, \ldots$, until an entire circuit around $z$ is completed
    {
    fix $\sigma_{i,l}$ such that $\int_{e_i} \sigma_{\omega_z} \cdot n ds = \int_{T_i} f \psi_z dx - \int_{e_{i-1}} \sigma_{\omega_z} \cdot n ds$;
    fix $\sigma_{i+1,r}$ such that $[\![\sigma_{\omega_z} \cdot n]\!] = -\frac{1}{2}[\![\nabla u_h \cdot n]\!]$ on $e_i$;
    }
```
□

It follows from Lemma 9.2 that the old and the new value of the normal component $\sigma_{1,r}$ coincide when an entire circuit around z has been completed.

If z is a node on $\partial\Omega$, the construction has to be modified in an obvious way. Here $\partial\omega_z$ shares two edges with $\partial\Omega$. We start at one of them and proceed as in Algorithm 9.3 until we get to the other edge on $\partial\Omega$. There is no problem since we do not return to the edge of departure.

9.4 Theorem. *Let u_h be the finite element solution with P_1 elements and*

$$\sigma^\Delta = \sum_z \sigma_{\omega_z}$$

where the functions σ_{ω_z} are constructed by Algorithm 9.3. Then

$$\|\nabla(u - u_h)\|_0 \le \|\sigma^\Delta\|_0 + ch\|f - f_h\|_0 . \tag{9.8}$$

Proof. Each edge has two nodes, and each triangle has three nodes. Moreover, we have $\sum_z \psi_z(x) = 1$ in each triangle T, and

$$\sum_z \int_T f \psi_z \, dx = \int_T f \, 1 \, dx = \int_T f_h \, dx .$$

Therefore it follows from (9.7) that the sum σ^Δ satisfies (9.4). The influence of the data oscillation was discussed in §8, and the theorem of Prager and Synge yields (9.8). □

The theory that led to Theorem 9.4 differs from the theory in the previous section. The estimator with local Neumann problems is also based on saddle point problems, but an additional discretization is required. Nevertheless, the error estimator $\|\sigma^\Delta\|_0$ is comparable to the residual error estimator (8.11). Consequently, the new estimator is also efficient.

9.5 Theorem. *There is a constant c that depends only on the shape of the triangles such that*

$$\|\sigma^\Delta\|_0 \le c\,\eta_R + ch\|f - f_h\|_0 \tag{9.9}$$

$$\le c|u - u_h|_1 + ch\|f - f_h\|_0 . \tag{9.10}$$

Proof. For convenience, we restrict ourselves to the case $f = f_h$ since the effect of the data oscillation is absorbed by the second terms in the inequalities. Only the residuals R_T and R_e defined in (8.2) and (8.3) enter into Algorithm 9.3. Therefore, the normal components of σ_{ω_z} on the edges are bounded,

$$|\sigma_{\omega_z} \cdot n| \le c\Big(h \sum_{T\subset\omega_z} |R_T| + \sum_{e\subset\omega_z} |R_e|\Big).$$

Since the broken Raviart–Thomas functions are piecewise polynomials with a fixed number of degrees of freedom, we have

$$\|\sigma_{\omega_z}\|_{0,T} \le ch_T \sum_{e\subset\omega_z} |\sigma_{\omega_z} \cdot n|$$

with c being a constant that depends only on the shape parameter. Hence,

$$\|\sigma_{\omega_z}\|_0^2 \le c \sum_{T\subset\omega_z} \eta_{T,R}^2$$

and another summation yields

$$\|\sigma^\Delta\|_0 \le c\,\eta_R\,.$$

This proves (9.9). The efficiency of the residual estimators (8.13) implies (9.10). □

As a byproduct we obtain the comparison of the performance of the P_1 element and of the mixed method with the Raviart–Thomas element that was stated in §5.

Proof of Theorem 5.6. We conclude from the preceding discussion that

$$\begin{aligned}\|\nabla u - \sigma_h\|_0^2 + \|\nabla(u-u_h)\|_0^2 &\le 2\,\|\sigma^\Delta\|_0^2 + ch^2\|f - f_h\|_0^2 \\ &\le c\|\nabla(u-u_h)\|_0^2 + ch^2\|f-f_h\|_0^2,\end{aligned}$$

and the proof is complete. □

Problem

9.6 Consider the Helmholtz equation

$$\begin{aligned}-\Delta u + \alpha u &= f \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \partial\Omega\end{aligned}$$

with $\alpha > 0$. Let $v \in H_0^1(\Omega)$ and $\sigma \in H(\text{div},\Omega)$ satisfy $\operatorname{div}\sigma + f = \alpha v$. Show the inequality of Prager–Synge type with a computable bound

$$\begin{aligned}&|u-v|_1^2 + \alpha\|u-v\|_0^2 \\ &+ \|\operatorname{grad} u - \sigma\|_0^2 + \alpha\|u-v\|_0^2 = \|\operatorname{grad} v - \sigma\|_0^2\,.\end{aligned} \tag{9.11}$$

Recall the energy norm for the Helmholtz equation in order to interpret (9.11).