MATH 8446, University of Minnesota, Spring 2010

# Numerical Analysis of Differential Equations, 2nd semester

Instructor's notes

Douglas N. Arnold

# Contents

CHAPTER 6

# $C^1$ finite element spaces

## 1. Review of finite elements

We begin with a brief review of finite elements as presented last semester. We considered the solution of boundary value problems for PDE that could be put into a *weak formulation* of the following sort: find $u \in V$ such that $b(u, v) = F(v)$ for all $v \in V$. Here $V$ is a Hilbert space, $b$ a bounded bilinear form, $F$ a bounded linear form. In the case where $b$ is symmetric and coercive, this weak formulation is equivalent to the variational problem

$$u = \operatorname*{argmin}_{v \in V} \left[ \frac{1}{2} b(v, v) - F(v) \right].$$

Such a weak formulation is well-posed if $b$ is coercive, or, more generally, if the *inf-sup condition* and *dense range condition* hold.

The numerical methods we considered were *Galerkin methods*, which means we seek $u_h$ in a finite dimensional subspace $V_h \subset V$ satisfying $b(u_h, v) = F(v)$ for all $v \in V_h$. If $b$ is coercive, this method is automatically *stable* with the stability constant $C_s$ bounded by the reciprocal of the coercivity constant. More generally, if the inf-sup condition holds on the discrete level, $C_s$ is bounded by the reciprocal of the inf-sup constant.

The *consistency error* for a Galerkin method is the *approximation error* for the space $V_h$ times the bound of $b$. From this we got the fundamental quasioptimal error estimate for Galerkin's method

$$\|u - u_h\|_V \leq (1 + C_s \|b\|) \inf_{v \in V_h} \|u - v\|_V.$$

For *finite element methods*, the spaces $V_h$ are constructed to be spaces of piecewise polynomials with respect to some simplicial decomposition of the domain, based on *shape functions* and *degrees of freedom*. For the case where $V$ is $H^1(\Omega)$, a very natural family of finite element spaces are the *Lagrange finite elements*, for which the shape functions on a simplex $T$ are the polynomials $\mathcal{P}_r(T)$ for some $r \geq 1$.

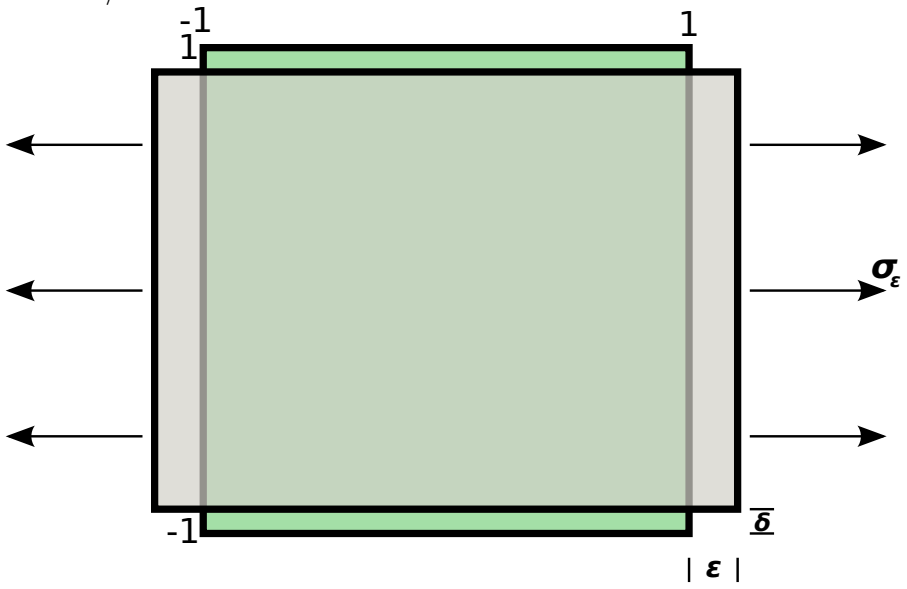
We bounded the approximation error for the Lagrange finite element spaces $V_h$ using the Bramble–Hilbert lemma and scaling. Putting together the above considerations, for the model scalar second order elliptic PDE, $-\operatorname{div} a \operatorname{grad} u + cu = f$, we obtained $H^1$ error estimates. We then used the *Aubin–Nitsche duality argument* to obtain error estimates of one higher order in $L^2$.

Finally, we introduced the Clément interpolant into the Lagrange finite element spaces, and used it to derive a posteriori error estimates, and error indicators which could be used in adaptive mesh refinement algorithms.

## 2. The plate problem

An elastic plate is a thin elastic body. First we recall that an elastic body is a sort of three-dimensional analogue of a spring. When a spring is extended it generates an internal restoring force, and in the simplest case, it satisfies Hooke's law: the force is proportional to extension. For an elastic body, a deformation in any direction provokes corresponding internal forces in the body, in all directions. In the simplest case of a linearly elastic material, the internal forces, or stresses are linear in the deformation. The simplest case is an *homogeneous* and *isotropic* elastic material. In this case the response of the material can be characterized in terms of two parameters, Young's modulus $E$ and Poisson's ratio $\nu$. Young's modulus is also called the tensile modulus, since it measures the tension (restoring force) in a length of the material subject to longitudinal stretching. In other words, if a sample in the form of a rectangular parallelpiped of width $L$ in one direction is stretched by pulling on the two opposite sides to increase their separation to $L(1+\epsilon)$, then the restoring force per unit area generated in the opposite direction will be $E\epsilon$. Thus $E$ is like the spring constant in Hooke's law. It has units of psi (pounds per square inch) in customary US units, or pascals (newtons per square meter) in international units. Aluminum, for instance, has $E$ around $1.0 \times 10^7$ psi, or $6.9 \times 10^{10}$ pascals.

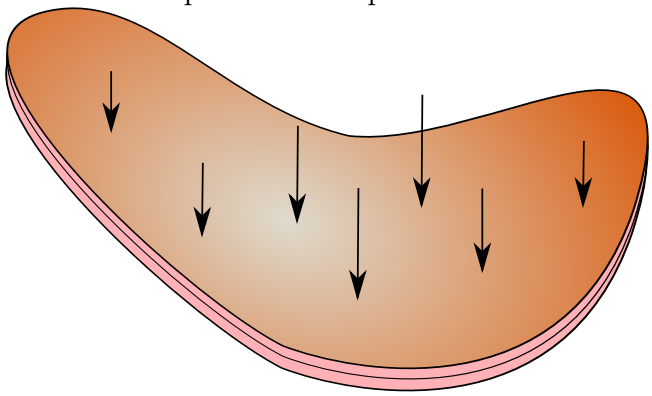
FIGURE 6.1. Elastic cube under tension $\sigma_\epsilon$. Strain is $\epsilon$ in the direction of tension, $-\delta$ in the normal directions. Young's modulus is $E = \sigma_\epsilon/\epsilon$. Poisson ratio is $\nu = \delta/\epsilon$.




Under the same tension test, Poisson's ratio is the ratio of the the compression in the orthogonal directions, to the extension in the given direction. Thus Poisson's ratio is dimensionless. The statement that if a material is stretched its volume does not decrease leads to $\nu \leq 1/2$. For most materials, $\nu \geq 0$, which we shall assume. For aluminum a value of about .33 is typical. For materials which are nearly incompressible, like rubber, the value is close to $1/2$.

We shall return to elasticity later in the course, but now we consider the transverse deflection of an elastic plate.

FIGURE 6.2. Thin plate under a transverse loading. Its deformation is measured by the vertical displacement of points on the middle plane.



Specifically, we suppose that our elastic body occupies the region $\Omega \times (-t/2, t/2)$ where $\Omega \subset \mathbb{R}^2$ is a domain (of roughly unit size) giving the crosssection of the plate, and $t << 1$ is the thickness. We assume that the plate is subject to a vertical load per unit area $g$, and let $w : \Omega \to \mathbb{R}$ denote the resulting vertical displacement of the middle surface. Then the classic *Kirchhoff plate bending model* says that $w$ minimizes the energy

$$\frac{1}{2}\frac{Et^3}{12(1-\nu^2)}\int_\Omega [(1-\nu)|\nabla^2 w|^2 + \nu|\Delta w|^2]\,dx - \int_\Omega gw\,dx.$$

The quantity $D = Et^3/[12(1-\nu^2)]$ is called the *bending modulus* of the plate. By $\nabla^2 w$ we mean the $2 \times 2$ Hessian matrix of $w$. (Warning: sometimes the notation $\nabla^2$ is used for the Laplacian, but we do not follow this usage.) For a matrix $\tau$ we write $|\tau|$ for the Frobenius norm $(\sum_{i=1}^2 \sum_{j=1}^2 \tau_{ij}^2)^{1/2}$ associated to the Frobenius inner product of matrices $\tau : \rho = \sum_{i=1}^2 \sum_{j=1}^2 \tau_{ij}\rho_{ij}$. Thus in the plate energy

$$|\nabla^2 w|^2 = \sum_{i,j}\left|\frac{\partial^2 w}{\partial x_i \partial x_j}\right|^2, \quad |\Delta w|^2 = \left|\sum_i \frac{\partial^2 w}{\partial x_i^2}\right|^2$$

The minimization of Kirchhoff's energy must be subject to boundary conditions, such as $w = \partial w/\partial n = 0$ on $\partial\Omega$ for a *clamped* plate, or just $w = 0$ for a *simply-supported* plate. Thus, if we define a bilinear form $b$ over $H^2(\Omega)$ by

$$b(w,v) = D\int_\Omega [(1-\nu)\nabla^2 w : \nabla^2 v + \nu\Delta w \Delta v]\,dx,$$

and the linear form $F(v) = \int_\Omega gv\,dx$, the clamped plate problem is to find $w \in V := \mathring{H}^2(\Omega)$ such that

$$b(w,v) = F(v), \quad v \in V.$$

The simply-supported plate problem has the same form, but with $V = H^2(\Omega) \cap \mathring{H}^1(\Omega)$.

Clearly $b(v,v) \geq D(1-\nu)|v|_2^2$ (the Sobolev $H^2$ seminorm), and there is a Poincaré type inequality which says that $\|v\|_2 \leq c_\Omega |v|_2$ for all $v \in H^2(\Omega) \cap \mathring{H}^1(\Omega)$, so $b$ is coercive over $V$

(for both the clamped and simply-supported cases) and so the weak formulation of the plate problem is well-posed.

Next we compute the strong form of the boundary value problems. First, for any smooth $u$ and $v$, we may integrate by parts twice and get Green's second identity:

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

Taking $u=\Delta w$ and $v\in\mathring{H}^2$, we get

$$\int_\Omega \Delta w\Delta v\,dx = \int_\Omega \Delta^2 w\,v\,dx,$$

while for $v\in H^2\cap\mathring{H}^1$,

$$\int_\Omega \Delta w\Delta v\,dx = \int_\Omega \Delta^2 w\,v\,dx + \int_{\partial\Omega}\Delta w\frac{\partial v}{\partial n}ds.$$

Now we consider the Hessian term. For a vector field $\phi$, let $\operatorname{grad}\phi$ denote the Jacobian matrix field $(\partial\phi_i/\partial x_j)$, and for a matrix field $\tau$, let $\operatorname{div}\tau$ denote the vector field $(\partial\tau_{i1}/\partial x_1+\partial\tau_{i2}/\partial x_2)$. Then

$$\begin{aligned}\int_\Omega \tau:\nabla^2 v\,dx &= \int_\Omega \tau:\operatorname{grad}\operatorname{grad} v\,dx = -\int_\Omega \operatorname{div}\tau\cdot\operatorname{grad} v\,dx + \int_{\partial\Omega}\tau n\cdot\operatorname{grad} v\,ds\\ &= \int_\Omega \operatorname{div}\operatorname{div}\tau\,v\,dx - \int_{\partial\Omega}(\operatorname{div}\tau\cdot n)v\,ds + \int_{\partial\Omega}\tau n\cdot\operatorname{grad} v\,ds.\end{aligned}$$

Also, if $s$ denotes the unit tangent, $\operatorname{grad} v = \frac{\partial v}{\partial n}n + \frac{\partial v}{\partial s}s$ and, if $v\in\mathring{H}^1$, $\frac{\partial v}{\partial s}=0$. Thus, for $v\in H^2\cap\mathring{H}^1$,

$$\int_\Omega \tau:\nabla^2 v\,dx = \int_\Omega \operatorname{div}\operatorname{div}\tau\,v\,dx + \int_{\partial\Omega} n\cdot\tau n\,\frac{\partial v}{\partial n}\,ds.$$

Taking $\tau=\nabla^2 w=\operatorname{grad}\operatorname{grad} w$, we get

$$\int_\Omega \nabla^2 w:\nabla^2 v\,dx = \int_\Omega \operatorname{div}\operatorname{div}\nabla^2 w\,v\,dx + \int_{\partial\Omega}\frac{\partial^2 w}{\partial n^2}\,\frac{\partial v}{\partial n}\,ds.$$

Now

$$\operatorname{div}\operatorname{div}\nabla^2 w = \sum_i \frac{\partial}{\partial x_i}\sum_j \frac{\partial}{\partial x_j}\frac{\partial^2 w}{\partial x_i\partial x_j} = \sum_i \frac{\partial^2}{\partial x_i^2}\sum_j \frac{\partial^2 w}{\partial x_j^2} = \Delta^2 w.$$

Putting all this together, we get for $w\in H^4$, $v\in\mathring{H}^2$,

$$b(w,v) = \int_\Omega D\Delta^2 w\,v\,dx,$$

while for $v\in H^2\cap\mathring{H}^1$,

$$b(w,v) = \int_\Omega D\Delta^2 w\,v\,dx + \int_{\partial\Omega} D[(1-\nu)\frac{\partial^2 w}{\partial n^2} + \nu\Delta w]\frac{\partial v}{\partial n}\,ds.$$

Therefore the strong form of the clamped plate problem is

$$D\Delta^2 w = f \text{ in } \Omega, \quad w = \frac{\partial w}{\partial n} = 0 \text{ on } \partial\Omega.$$

In this case both boundary conditions are *essential*.

The simply supported plate problem is

$$D\Delta^2 w = f \text{ in } \Omega, \quad w = D[(1-\nu)\frac{\partial^2 w}{\partial n^2} + \nu\Delta w] = 0 \text{ on } \partial\Omega.$$

In this case, the second boundary condition (which physically means that the *bending moment* vanishes), is natural.

REMARK. As an interesting digression, we describe the *Babuška plate paradox*. Suppose that we want to solve the Dirichlet problem for Poisson's equation on a smoothly bounded domain, such as the unit disc. We might triangulate the domain, and then use standard finite elements. The triangulation involves an approximation of the domain with a nearby polygon, e.g., an inscribed polygon in the disc. It is true, and not surprising, that the solution to the boundary value problem on the polygon converges to the solution on the disc, as more sides are added to the polygon, so that it approaches the disc. However consider a circular simply-supported plate (so the domain $\Omega$ is the unit disc). For simplicity we take the Poisson ratio equal to 0. Then the plate equations are

$$\Delta^2 w = f \text{ in } \Omega, \quad w = \frac{\partial^2 w}{\partial n^2} = 0 \text{ on } \partial\Omega. \tag{6.1}$$

Now consider the same system on the domain $\Omega_m$ which is an $m$-sided regular polygon inscribed in the unit disc, and let $w_m$ be the corresponding solution. Then the paradox is that $\bar{w} := \lim_{m\to\infty} w_m$ exists but is different from $w$. In fact, in the case of a uniform load $f = D$, $\bar{w}(0,0)$ is 40% smaller than $w(0,0)$.

To see how this comes about, we consider the boundary conditions. On a straight edge we may write

$$\Delta u = \frac{\partial^2 u}{\partial n^2} + \frac{\partial^2 u}{\partial s^2},$$

and, if $u = 0$ on the edge, then the second term vanishes. Thus on a straight portion of the boundary the simply-supported plate boundary conditions $u = \partial^2 u/\partial n^2 = 0$ are the same as $u = \Delta u = 0$. It can be shown rigorously that the same is true on a polygonal domain, in which the boundary is straight everywhere except at finitely many points. Thus

$$\Delta^2 w_m = f \text{ in } \Omega_m, \quad w_m = \Delta w_m = 0 \text{ on } \partial\Omega_m.$$

So it is not surprising that the limit $\bar{w}$ of the $w_m$ satisfies the problem

$$\Delta^2 \bar{w} = f \text{ in } \Omega, \quad \bar{w} = \Delta\bar{w} = 0 \text{ on } \partial\Omega. \tag{6.2}$$

This can be proven rigorously using the fact that this problem decouples as two Poisson problems. However, the expression for the Laplacian in polar coordinates is

$$\Delta w = \frac{\partial^2 w}{\partial r^2} + \frac{1}{r}\frac{\partial w}{\partial r} + \frac{1}{r^2}\frac{\partial^2 w}{\partial \theta^2},$$

so, on the boundary of the unit disc, for $w$ vanishing there,

$$\Delta w = \frac{\partial^2 w}{\partial n^2} + \frac{\partial w}{\partial n}.$$

Thus (6.1) becomes

$$\Delta^2 w = f \text{ in } \Omega, \quad w = \Delta w - \frac{\partial w}{\partial n} = 0 \text{ on } \partial\Omega,$$

which is a different problem from (6.2).

In fact, in the case $f \equiv 1$, the exact solution of (6.1) is $w = (r^4 - 6r^2 + 5)/64$, while the exact solution to (6.2) is $\bar{w} = (r^4 - 4r^2 + 3)/64$.

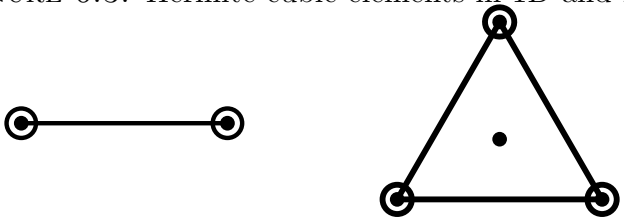
## 3. Conforming finite elements for the plate problem

Since the weak formulation of the plate problem (with either Dirichlet or simply-supported boundary conditions) is coercive over $H^2$. Therefore, we may use the Galerkin method with any subspace of $H^2$ (satisfying the essential boundary conditions), and get quasioptimal approximation in $H^2$. Therefore we now consider finite element subspaces of $H^2$.

As we know, a piecewise smooth function with respect to a triangulation belongs to $H^1$ if and only if it is continuous. (Thus, for example, the space of all piecewise polynomials of degree at most $r$ is exactly the Lagrange finite element space of degree $r$, since it consists precisely of the continuous piecewise polynomials of degree at most $r$.) A function belongs to $H^2$ only if it and all its first derivatives belong to $H^1$, so a piecewise smooth function belongs to $H^2$ if and only if it is $C^1$. This means that a finite element Galerkin method for the plate bending problem requires $C^1$ finite elements. This motivated a search to find shape functions and degrees of freedom which would ensure $C^1$ continuity.

**3.1. Hermite quintic elements.** In one-dimension it is not difficult to find $C^1$ finite elements (we could use these to solve the problem of the bending of an elastic bar). The simplest are the Hermite cubic elements, illustrated in Figure 6.3, with $\mathcal{P}_3$ shape functions and the values and first derivatives as DOFs on each interval. So let's consider the 2D analogue of these. On a triangle the Hermite cubic elements use $\mathcal{P}_3$ shape functions. Guided by 1D, we take as degrees of freedom the values and the values of the first derivatives at each vertex. Since there are two first derivatives, this gives 9 DOFs, leaving one more to be chosen. For this we take the value at the barycenter (Figure 6.3, right).

FIGURE 6.3. Hermite cubic elements in 1D and 2D.

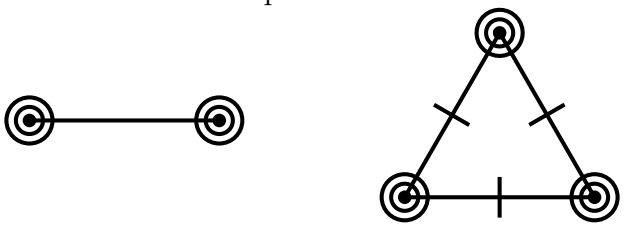


First we show the unisolvence of the proposed DOFs. Suppose $u \in \mathcal{P}_3(T)$ for some triangle $T$, and all the DOFs for $u$ vanish. For an edge $e$ of $T$, let $v = u|_e$. Using the distance along $e$ as a coordinate, we may view $e$ as an interval, and $v$ belongs to $\mathcal{P}_3(e)$,

and both $v$ and its derivative vanish at the end points. Therefore (by unisolvence of the Hermite cubic in 1D), $v$ vanishes, i.e., $u$ vanishes on $e$. This holds for all three edges, so $u$ is divisible by the bubble function $\lambda_1\lambda_2\lambda_3$. Since $u$ is cubic, it is a constant multiple. Since $u$ also vanishes at the barycenter (where the bubble function is positive), the constant must be zero, so $u \equiv 0$.

Our argument also showed that the DOFs associated to an edge $e$ determine $u$ on the edge $e$, so the resulting assembled finite element space will be $C^0$. Let us try to show it is $C^1$. This means that we must show that $\partial u/\partial n$ is determined by the DOFs on $e$. But the DOFs only determine $\partial u/\partial n$ at the two endpoints of $e$, and it is a polynomial of degree 2, which requires 3 values to be uniquely determined. Thus the Hermite cubic space is *not* $C^1$ in more than one dimension. (For a specific counterexample, consider two triangles with a common edge and define a piecewise polynomial which vanishes on one of the triangles and is equal to the bubble function on the other. This belongs to the Hermite cubic finite element space, but is not $C^1$.)

FIGURE 6.4. Hermite quintic elements in 1D and 2D.



Continuing our search for $C^1$ finite elements, we look to the Hermite quintic space. In 1D this gives a $C^2$ finite element. We shall show that in 2D it gives a $C^1$ space. The shape functions are, of course, $\mathcal{P}_5(T)$, a space of dimension 21. The DOF are the values of function and all its first and second derivatives at the vertices, and the values of the normal derivatives at the midpoints of each edge, which comes to 21 DOFs. This finite element is often called the *Argyris triangle*. Unisolvence is straightforward. If all the DOFs for $u$ vanish, then by the unisolvence of the Hermite quintic in 1D, $u$ vanishes on each edge. But also, on an edge $\partial u/\partial n$ is a quartic polynomial which vanishes along with its derivative at the endpoints, and, moreover, it vanishes in the midpoint of the edge. This is a unisolvent set of DOFs for a quartic in 1D, and hence the normal derivative vanishes on each edge as well. But a polynomial and its normal derivative vanish on the line $\lambda_i = 0$ if and only if it is divisible by $\lambda_i^2$. Thus $u$ is a multiple of $\lambda_1^2\lambda_2^2\lambda_3^2$ which is a polynomial of degree 6, and hence $u$, a polynomial of degree at most 5, must vanish.

Note that in the course of proving unisolvence we showed that $u$ and its normal derivative are determined on an edge by the degrees of freedom associated to the edge and its endpoints. Consequently the assembled finite element space belongs to $C^1$.

It is important to note that the assembled finite elements are, in fact, smoother than just $C^1$. They are, by definition, also $C^2$ at the vertices. The assembled Hermite quintic finite element space is precisely

$$\{\, u \in C^1(\Omega) \,|\, u|_T \in \mathcal{P}_5(T)\ \forall T, u \text{ is } C^2 \text{ at all vertices} \,\}.$$

This extra restriction in the space is a mild shortcoming of the Hermite quintic element as a $C^1$ (or $H^2$) finite element. In addition, with 21 degrees of freedom per triangle, of several

different types (values, first derivatives, second derivatives, normal derivatives), the element is regarded as quite complicated, especially in earlier days of finite element analysis. It is, nonetheless, an important element for actual computation.

If we use the Hermite quintic finite element space $V_h \subset V$, we get the quasioptimal estimate

$$\|w - w_h\|_2 \le c \inf_{v \in V_h} \|w - v\|_2. \tag{6.3}$$

So next we consider the approximation error for the space. From the DOFs we can define a projection operator $I_h : H^4(\Omega) \to V_h$. (It is bounded on $H^4$, but not on $H^3$, because it requires point values of the 2nd derivative.) $I_h$ is built from projections which preserve quintics on each triangle, so we would expect that we could use Bramble–Hilbert and scaling to get

$$\inf_{v \in V_h} \|w - v\|_2 \le ch^r \|w\|_{r+2}, \quad r = 2, 3, 4.$$

There is one complication. For Lagrange elements, we used the Bramble–Hilbert lemma to get an estimate only on the unit triangle, and then for an arbitrary triangle, we used affine scaling to the unit triangle. We found that the scaling brought in the correct powers of $h$ as long as we stuck to shape regular triangulations. To show this we needed the fact that the interpolant of the affinely scaled function is the affine scaling of the interpolant. This last fact does not hold when the interpolant is taken to be the Hermite quintic interpolant. The reason is that normals are not mapped to normals (and normal derivatives to normal derivatives) for general affine maps.

That is, given a triangle $T$ and $C^2$ function $u$ on $T$, let $I_T u \in \mathcal{P}_6(T)$ denote its Hermite quintic interpolant. If $\hat{T}$ is another triangle and $F$ an affine map taking $\hat{T}$ to $T$, we let $\hat{u} = u \circ F$. Then $(I_{\hat{T}} \hat{u}) \circ F^{-1}$ need not coincide with $I_T u$. For this reason, rather than general affine maps, we shall consider only dilations ($F\hat{x} = h\hat{x}$). As long as $F$ belongs to this class, it is easy to see check that $I_T u = (I_{\hat{T}} \hat{u}) \circ F^{-1}$.

For $\theta > 0$, define $\mathcal{S}_\theta$ to be the set of all triangles of diameter 1 all of whose angles are bounded below by $\theta$. Also let $\mathcal{S}'_\theta$ denote the elements of $\mathcal{S}_\theta$ which are normalized in the sense that their longest edge lies on the interval from 0 to 1 on the $x$-axis and its third vertex lies in the upper half plane. Note that the possible positions for the third vertex of $\hat{T} \in \mathcal{S}'_\theta$ lie inside a compact subset of the upper half plane. See Figure 6.5.

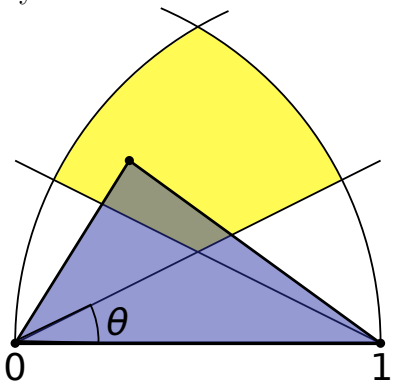
Now for any triangle $\hat{T}$, we know by the Bramble–Hilbert lemma that

$$|u - I_{\hat{T}} u|_r \le c|u|_s, \tag{6.4}$$

for $0 \le r \le s$, $s = 4, 5, 6$ (the lower bound on $s$ comes from the need for point values of the second derivative). Moreover a single constant $c$ works for all $\hat{T} \in \mathcal{S}'_\theta$, since the best constant depends continuously on the third vertex, which varies in a compact set. Of course the estimate is unchanged if we transform $\hat{T}$ by a rigid motion. Therefore, (6.4) holds with $c$ uniform over all $\hat{T} \in \mathcal{S}_\theta$.

Now let $T$ by any triangle with least angle $\ge \theta$. Set $h_T = \operatorname{diam} T$, and define $\hat{T} = h_T^{-1} T$, which belongs to $\mathcal{S}_\theta$. Note that $|T| = h_T^2 |\hat{T}|$. Given a function $u$ on $T$, define $\hat{u}(\hat{x}) = u(h_T \hat{x})$, $\hat{x} \in \hat{T}$. As we mentioned above, $I_{\hat{T}} \hat{u}(\hat{x}) = I_T u(h_T \hat{x})$. Of course, we have

FIGURE 6.5. The blue triangle belongs to $\mathcal{S}'_\theta$, i.e., its longest edge runs from 0 to 1 on the $x$-axis, its third vertex lies in the upper half plane, and all its angles are bounded below by $\theta$. Consequently the third vertex must lie in the compact region shown in yellow.




$D^\beta \hat{u}(\hat{x}) = h_T^{|\beta|} D^\beta u(x)$. Thus we get from

$$|u - I_T u|_{H^r(T)} = h_T^{-r} h_T |\hat{u} - I_{\hat{T}} \hat{u}|_{H^r(\hat{T})} \le c h_T^{-r} h_T |\hat{u}|_{H^s(\hat{T})} = c h_T^{s-r} |u|_{H^s(T)}.$$

Thus, through the usual approach of Bramble–Hilbert and scaling, but this time limiting the scaling to dilation, we have proved the expected estimates for the Hermite quintic interpolant:

$$|u - I_T u|_r \le c h_T^{s-r} |u|_s,$$

where $c$ only depends on the shape regularity of the triangle $T$. For a mesh of triangles, all satisfying the shape regularity constraint and with $h = \max h_T$, we can apply this element by element, square, and add. In this way we get

$$|u - I_h u|_r \le c h^{s-r} |u|_s, \quad u \in H^s(\Omega),$$

for $0 \le r \le 2$, $4 \le s \le 6$ (the upper bound on $r$ comes from the requirement that $I_h u \in H^r(\Omega)$.

Combining with the quasioptimality estimate (6.3), we immediately obtain error estimates for the finite element solution.

$$\|w - w_h\|_2 \le c h^{s-2} |w|_s,$$

where $w$ is the exact solution and $w_h$ the finite element solution. In particular, if $w$ is smooth, then $\|w - w_h\|_2 = O(h^4)$.

Concerning the smoothness of the exact solution, we run into a problem that we also ran into when we considered the Poisson equation. If the domain $\Omega$ has a smooth boundary and the data $f$ is smooth, then the theory of elliptic regularity insures that $w$ is smooth as well. However, since we have assumed that our domain can be triangulated, it is a polygon and therefore its boundary is not smooth. So in practice $w$ may not be smooth enough to imply $O(h^4)$ convergence.

Lack of regularity of the domain is also a problem when we try to apply an Aubin–Nitsche duality argument to get high order convergence in $H^1$ or $L^2$, because this requires an elliptic

regularity estimate, which will not hold on an arbitrary polygonal domain. For example, suppose we try to prove an $L^2$ estimate. We define $\phi \in V$ by

$$b(u,\phi) = \int u(w - w_h)\,dx, \quad u \in V.$$

Then $\phi$ satisfies the plate problem with $D\Delta^2\phi = w - w_h$. Taking $u = w - w_h$, we get

$$\|w - w_h\|^2 = b(w - w_h, \phi) = \inf_{v\in V_h} b(w - w_h, \phi - v) \le c\|w - w_h\|_2 \inf_{v\in V_h} \|\phi - v\|_2.$$

If we knew that $\phi \in H^4$ and $\|\phi\|_4 \le c\|w - w_h\|$, we could then complete the argument: But

$$\inf_{v\in V_h} \|\phi - v\|_2 \le ch^2\|\phi\|_4 \le ch^2\|w - w_h\|,$$

so $\|w - w_h\| \le ch^2\|w - w_h\|_2$. Unfortunately such 4-regularity of the plate problem does not hold on a general polygon, or even a general convex polygon.

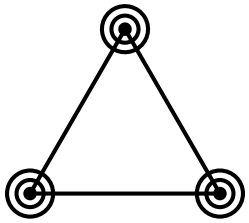
**3.2. Reduced Hermite quintic.** The difficulties with Hermite quintic elements (many DOFs, need for second derivatives, complicated) motivate the search for simpler elements. It turns out that one slight simplification can be made fairly easily. Define

$$\mathcal{P}_5'(T) = \{\, u \in \mathcal{P}_5(T) \mid u|_e \in \mathcal{P}_4(e) \text{ for each edge} \,\}.$$

Then $\dim \mathcal{P}_5'(T) \ge 18$. Indeed if we write out a general element of $\mathcal{P}_5(T)$ in terms of 21 coefficients, then each of the conditions $u|_e \in \mathcal{P}_4(e)$ is a homogeneous linear equation which must be satisfied the coefficients, so we get a system of 3 homogeneous linear equations in 21 unknowns. Now consider the 18 DOFs at the vertices we used for the Hermite quintic (but ignore the 3 DOFs at the edge midpoints). If these 18 DOFs vanish for an element $u \in \mathcal{P}_5'(T)$, then $u$ must vanish, by the same argument we used for $\mathcal{P}_5$. This implies that $\dim \mathcal{P}_5'(T) \le 18$, so we have equality, and we have a unisolvent set of degrees of freedom.

This finite element is called the *reduced Hermite quintic* or *Bell's triangle*. Its advantage over the full Hermite quintic is that it is in some ways simpler: it has 18 rather than 21 DOFs and all are values of the function or its derivatives at the vertices. The disadvantange is that the shape functions contain all of $\mathcal{P}_4(T)$, but not all of $\mathcal{P}_5(T)$. Therefore the rate of approximation for smooth functions is one order lower.

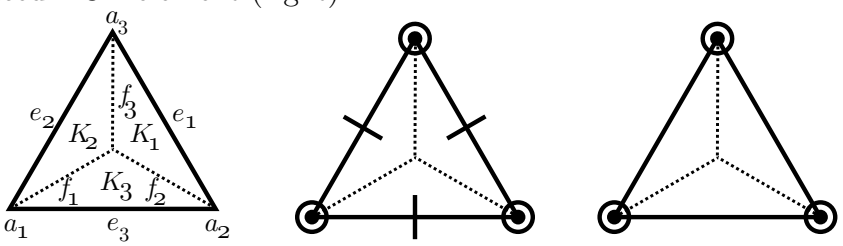
FIGURE 6.6. Reduced Hermite quintic element.



**3.3. Hsieh–Clough–Tocher composite elements.** It is not possible to design simpler conforming finite elements for the plate equation using polynomial shape functions. But in the early 1960s the civil engineer R. Clough (who, incidentally, invented the term "finite elements") and his students J. Tocher and T. K. Hsieh designed an element using *piecewise polynomial* shape functions on each triangle. To describe this HCT element, consider an arbitrary triangle $T$, partitioned into 3 subtriangles by connecting each vertex to a point $b$ in the center. It is natural, but not necessary, to take $b$ to be the barycenter of $T$, as we shall do. Let $K_1, K_2, K_3$ denote the 3-subtriangles. Then we shall use as the space of shape functions on $T$

$$\{\, u \in C^1(T) \,|\, u|_{K_i} \in \mathcal{P}_3(K_i), i = 1, 2, 3 \,\}.$$

FIGURE 6.7. A subdivided triangle (left), the HCT element (middle), and the reduced HCT element (right).




Our first task is to find the dimension of the space of shape functions. Each of the spaces $\mathcal{P}_3(K_i)$ is of dimension 10. We then impose the condition that $u|_{K_1}$ agrees with $u|_{K_2}$ and $u|_{K_3}$ at $b$ (which gives two homogeneous linear equations on the coefficients). We do similarly for $\partial u/\partial x_1$ and $\partial u/\partial x_2$, so we obtain in this way 6 equations in all. Next we take any two distinct points in the interior of the edge separating $K_1$ and $K_2$ and impose the equation that $u|_{K_1}$ and $u|_{K_2}$ agree at these two points and similarly for $\partial u/\partial n$. In this way we obtain 4 more equations. Doing this for all three interfaces, we obtain, altogether 18 homogeneous linear equations the 30 coefficients must satisfy in order that they join together to make a $C^1$ function. Thus the dimension of the space of shape functions is $\geq 12$. We now take as DOFs the 12 quantities indicated in the center of Figure 6.7 and show that if all vanish, then $u$ vanishes. This will imply that the dimension is exactly 12 and the DOFs are unisolvent.

The argument, which is taken from the monograph of Ciarlet, begins in the usual way. Let $u_i$ be the polynomial given by $u|_{K_i}$. On the edge of $T$ contained in $K_1$, $u_i$ is cubic and the 4 DOFs on that edge imply that $u_i$ vanishes on the edge. Similarly we get that $\partial u_i/\partial n$ vanishes on the edge. Hence the polynomial $u_i$ is divisible by $\mu_i^2$, where $\mu_i$ is the barycentric coordinate function on $K_i$ which is 1 at $b$ and vanishes on the two vertices of $T$ in $K_i$. Thus $u_i = p_i\mu_i$, where $p_i \in \mathcal{P}_1$. Now $\mu_1$ and $\mu_2$ agree on $f_3$. Since $p_1\mu_1^2$ and $p_2\mu_2^2$ must also agree on $f_3$ (by the continuity of $u$), we conclude that $p_1 = p_2$ on $f_3$. In this way we conclude that the piecewise linear which equals $p_i$ on $K_i$ is continuous.

Now consider the continuity of $\nabla u$ across $f_3$. This gives

$$(\nabla p_1)\mu_1^2 + 2p_1\mu_1\nabla\mu_1 = (\nabla p_2)\mu_2^2 + 2p_2\mu_2\nabla\mu_2 \text{ on } f_3.$$

On $f_3$, $\mu_1 = \mu_2 \neq 0$, so we can divide by this polynomial and recall that $p_1 = p_2$ on $f_3$ to get that

$$(\nabla p_1)\mu_1 + 2p_1\nabla\mu_1 = (\nabla p_2)\mu_2 + 2p_2\nabla\mu_2 \text{ on } f_3.$$

Passing to the vertex $a_3$ of $f_3$, where $\mu_1 = \mu_2 = 0$,

$$p_1 \nabla \mu_1 = p_1 \nabla \mu_2 \text{ at } a_3.$$

Now $\nabla \mu_1$ is a constant vector normal to $e_1$ and $\nabla \mu_2$ is a constant vector normal to $e_2$. So the above equation implies that $p_1(a_3) = 0$. Of course we get $p_1(a_2) = 0$ in the same way, so the linear polynomial $p_1$ vanishes on $e_1$, so is a constant multiple of $\mu_1$. Thus we have shown the $u_1 = C\mu_1^3$ for some constant $C$, which must be $u(b)$. In the same way we get $u_2 = C\mu_2^3$ and $u_3 = C\mu_3^3$. Then we equate $\nabla u_1$ and $\nabla u_2$ on $f_3$ and conclude that $C$ must be zero.

Thus the HCT element is unisolvent. While the space of shape functions does not include only polynomials (rather piecewise polynomials), it does include the space $\mathcal{P}_3(T)$. Therefore the interpolant associated to the DOFs preserves cubics, and we can use a Bramble–Hilbert argument with dilation, as for the Hermite quintics, and prove that $\inf_{v \in V_h} \|u - v\|_2 \leq Ch^2 \|u\|_4$ when $V_h$ is the HCT space.

It is also possible to define a reduced HCT space, a finite element space with 9 DOFs, just as we defined a reduced Hermite quintic space. The DOFs are shown in Figure 6.7.

CHAPTER 7

# Nonconforming elements

The complexity of finite element subspaces of $H^2$ motivates the development of *nonconforming* finite elements. These are finite elements for which the assembled space $V_h$ is not contained in $H^2$ (i.e., not contained in $C^1$). For this reason $\Delta v$ and $\nabla^2 v$ do not make sense (or at least are not $L^2$ functions) for $v \in V_h$. However, on each element $T \in \mathcal{T}_h$ $\nabla^2 v$ is well-defined, so we can define $w_h \in V_h$ by

$$\sum_{T\in\mathcal{T}_h} \int_T \nabla^2 w_h : \nabla^2 v \, dx = \int_\Omega f v \, dx, \quad v \in V_h.$$

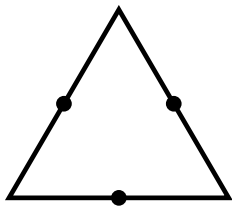
Not surprisingly, this method does not work in general. However, as we shall see, if we take elements which are in some sense "nearly $C^1$", we obtain a convergent method.

## 1. Nonconforming finite elements for Poisson's equation

First we will examine the idea of nonconforming finite elements in the simpler situation of Poisson's equation, which we will solve with finite element spaces which are not contained in $H^1$. Although we are doing this just to guide us in the more complicated case of $H^2$ elements, it turns out that the non-conforming $H^1$ elements are useful in some contexts.

We now define the space of non-conforming $\mathcal{P}_1$ finite elements. The shape functions are $\mathcal{P}_1(T)$, like for Lagrange $\mathcal{P}_1$ elements, but the DOFs are the values at the midpoints of the edges.

FIGURE 7.1. Nonconforming $\mathcal{P}_1$ finite element.



Consider now the Dirichlet problem

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

As a finite element space $V_h$ we use the nonconforming $\mathcal{P}_1$ space with all the DOFs on the boundary set equal to zero. Thus $\dim V_h$ is the number of interior edges of the mesh. The finite element method is to find $u_h \in V_h$ such that

$$\sum_{T\in\mathcal{T}_h} \int_T \operatorname{grad} u_h \cdot \operatorname{grad} v \, dx = \int_\Omega f v \, dx, \quad v \in V_h.$$

Writing $b_h(w,v) = \sum_{T\in\mathcal{T}_h} \int_T \operatorname{grad} w \cdot \operatorname{grad} v \, dx$ for any piecewise smooth $w$ and $v$, we may write the finite element method as: find $u_h \in V_h$ such that

$$b_h(u_h, v) = \int_\Omega fv \, dx, \quad v \in V_h. \tag{7.1}$$

When we try to analyze this, the first difficulty we encounter is that the true solution does not satisfy the discrete equations. That is, the equation

$$\sum_{T\in\mathcal{T}_h} \int_T \operatorname{grad} u \cdot \operatorname{grad} v \, dx = \int_\Omega fv \, dx,$$

holds if $v \in \mathring{H}^1(\Omega)$, but need not hold for $v \in V_h$.

To understand better what is going on, we multiply the differential equation by a test function $v \in \mathcal{P}_1(T)$ and integrate by parts over $T$:

$$\int_T fv \, dx = -\int_T \Delta u \, v \, dx = \int_T \operatorname{grad} u \cdot \operatorname{grad} v \, dx - \int_{\partial T} \frac{\partial u}{\partial n_T} v \, ds.$$

Next we add over $T$:

$$\sum_T \int_T \operatorname{grad} u \cdot \operatorname{grad} v \, dx - \sum_T \int_{\partial T} \frac{\partial u}{\partial n_T} v \, ds = \int_\Omega fv \, dx.$$

In other words

$$b_h(u, v) = \int_\Omega fv \, dx + E_h(u, v), \quad v \in V_h, \tag{7.2}$$

where

$$E_h(u, v) = \sum_T \int_{\partial T} \frac{\partial u}{\partial n_T} v \, ds.$$

Note that $E_h(u,v)$ measures the extent to which the true solution $u$ fails to satisfy the finite element equations, so it measures a kind of *consistency error*. This is different from the consistency error we saw in conforming methods, which comes from the approximation properties of the trial functions. Of course that sort of approximation error is also present for nonconforming methods. But nonconforming methods also feature the consistency error given by $E_h(u,v)$, which is due to the fact that the test functions do not belong to the space of test functions on the continuously level. (Note that it is the test functions, not the trial functions that matter here.)

In order to analyze this method we introduce some notation. Define the space of piecewise $H^1$ functions with respect to the triangulation,

$$H^1(\mathcal{T}_h) = \{\, v \in L^2(\Omega) \,|\, v|_T \in H^1(T),\ T \in \mathcal{T}_h \,\},$$

Note that both $H^1 \subset H^1(\mathcal{T}_h)$ and $V_h \subset H^1(\mathcal{T}_h)$, so this is a space in which we can compare the exact solution and the finite element solution. We also define the piecewise gradient $\operatorname{grad}_h : H^1(\mathcal{T}_h) \to L^2(\Omega, \mathbb{R}^2)$, given by

$$(\operatorname{grad}_h v)|_T = \operatorname{grad}(v|_T), \quad v \in H^1(\mathcal{T}_h),\ T \in \mathcal{T}_h.$$

Then the bilinear form $b_h(w,v) = \int \operatorname{grad}_h w \cdot \operatorname{grad}_h v \, dx$ is defined for all $w, v \in H^1(\mathcal{T}_h)$, and the associated seminorm, the *broken $H^1$ seminorm*, is

$$\|v\|_h := \|\operatorname{grad}_h v\|.$$

Although it is just a seminorm on $H^1(\mathcal{T}_h)$, on the subspace $\mathring{H}^1 + V_h$, it is a norm. Indeed if $\|v\|_h = 0$, then $v$ is piecewise constant. Since it is continuous at the midpoint of each edge, it is globally constant, and since it vanishes at the midpoint of each boundary edge, it vanishes altogether.

We clearly have the bilinear form is bounded and coercive with respect to this norm:

$$|b_h(w,v)| \le M\|w\|_h\|v\|_h, \quad b_h(v,v) \ge \gamma\|v\|_h^2, \quad w, v \in H^1(\mathcal{T}_h),$$

(in fact, with $M = \gamma = 1$).

Subtracting (7.1) from (7.2) we obtain the error equation.

$$b_h(u - u_h, v) = E_h(u, v), \quad v \in V_h.$$

Let $r_h u \in V_h$ be an approximation of $u$ (to be specified later). Then

$$b_h(r_h u - u_h, v) = b_h(r_h u - u, v) + E_h(u, v), \quad v \in V_h.$$

Taking $v = r_h u - u_h$, we get

$$\|r_h u - u_h\|_h^2 \le \|r_h u - u\|_h \|r_h u - u_h\|_h + |E_h(u, r_h u - u_h)|.$$

We shall prove:

THEOREM 7.1 (Bound on consistency error for $\mathcal{P}_1$ nonconforming FE). *There exists a constant $c$ such that*

$$|E_h(u,v)| \le ch\|u\|_2\|v\|_h, \quad v \in \mathring{H}^1 + V_h.$$

Using this result it is easy to complete the argument. We immediately get

$$\|r_h u - u_h\|_h \le \|r_h u - u\|_h \ + ch\|u\|_2,$$

and so

$$\|u - u_h\|_h \le 2\|r_h u - u\|_h + ch\|u\|_2.$$

For the approximation error $r_h u - u$ we could take $r_h u$ to be the interpolant into $V_h$ and use a Bramble–Hilbert argument. But even easier, we take $r_h u$ to be the interpolant of $u$ into the Lagrange $\mathcal{P}_1$ space, which is a subspace of $V_h$, and for which we already known $\|r_h u - u\|_h \le ch\|u\|_2$. Thus we have proven (modulo Theorem 7.1) the following error estimates that for the $\mathcal{P}_1$ nonconforming finite element method.

THEOREM 7.2 (Convergence of $\mathcal{P}_1$ nonconforming FE). *Let $u$ solve the Dirichlet problem for Poisson's equation and let $u_h$ be the finite element solution computed using $\mathcal{P}_1$ nonconforming finite elements on a mesh of size $h$. Then*

$$\|u - u_h\|_h \le ch\|u\|_2.$$

It remains to prove the bound on the consistency error given in Theorem 7.1. The theorem follows immediately from the following lemma (by taking $\phi = \operatorname{grad} u$).

LEMMA 7.3. *There exists a constant $c$ such that*

$$\left|\sum_{T\in\mathcal{T}_h}\int_{\partial T}(\phi\cdot n_T)v\,ds\right| \le Ch\|\operatorname{grad}\phi\|_0\|\operatorname{grad}_h v\|_0, \quad \phi\in H^1(\Omega;\mathbb{R}^2), \quad v\in \mathring{H}^1(\Omega)+V_h.$$

To see why a result like this should be true, think of each of the integrals over $\partial T$ as a sum of three integrals over the three edges of $T$. When we sum over all $T$, we will get two terms which are integrals over each edge $e$ in the interior of $\Omega$, and one term for each edge in $\partial\Omega$. For an interior edge $e$, let $T_+$ and $T_-$ be the triangles sharing the edge $e$ and let $n_e$ denote the unit normal pointing out of $T_+$ into $T_-$, (so $n_e = n_{T_+} = -n_{T_-}$ on $e$). Define $v_+$ and $v_-$ to be the restriction of $v$ to $T_+$ and $T_-$, and set $[\![\, v \,]\!] = v_+ - v_-$ on $e$, the *jump* of $v$ across $e$. Then the contribution to the sum from $e$ is $\int_e(\phi\cdot n_e)[\![\, v \,]\!]\,ds$. For $e$ an edge contained in $\partial\Omega$, the contribution to the sum is just $\int_e(\phi\cdot n_T)v\,ds$, so for such edges we define $n_e$ to be $n_T$ (the unit normal pointing exterior to $\Omega$) and define $[\![\, v \,]\!]$ to be $v|_e$. With this notation, we have

$$\sum_{T\in\mathcal{T}_h}\int_{\partial T}(\phi\cdot n_T)v\,ds = \sum_e\int_e(\phi\cdot n_e)[\![\, v \,]\!]\,ds,$$

where the second sum is over all edges. Now, if $v\in\mathring{H}^1$, then $[\![\, v \,]\!]$ vanishes, but for $v\in V_h$ it need not. It is a linear polynomial on the edge $e$. However, it is not just any linear polynomial: it is a linear polynomial on $e$ (an edge of length at most $h$) which vanishes at the midpoint of $e$. Therefore, roughly, we expect $v$ to be of size $h$, which explains where the factor of $h$ arises in Lemma 7.3.

To prove Lemma 7.3, we need a new approximation estimate. Let $T$ be a triangle and $e$ an edge of $T$. Let $P_e : L^2(e)\to\mathbb{R}$ be the $L^2(e)$ projection, i.e., the constant $P_e\psi$ is the average value of $\psi\in L^2(e)$.

LEMMA 7.4. *Let $T$ be a triangle and $e$ and edge. There exists a constant depending only on the shape constant for $T$ such that*

$$\|\phi|_e - P_e(\phi|_e)\|_{L^2(e)} \le ch_T^{1/2}\|\operatorname{grad}\phi\|_{L^2(T)}, \quad \phi\in H^1(T).$$

PROOF. The operator $\phi\mapsto\phi|_e - P_e(\phi|_e)$ is a bounded linear operator $H^1(T)\to L^2(e)$ which vanishes on constants. From the Bramble–Hilbert lemma, we find

$$\|\phi|_e - P_e(\phi_e)\|_{L^2(e)} \le c_T\|\operatorname{grad}\phi\|_{L^2(T)}, \quad \phi\in H^1(T).$$

We apply this result on the unit triangle $\hat{T}$, and then use affine scaling to get it on an arbitrary element, leading to the claimed estimate. □

PROOF OF LEMMA 7.3. Let $e$ be an edge. Then

$$\left|\int_e(\phi\cdot n_e)[\![\, v \,]\!]\,ds\right| = \left|\int_e[\phi\cdot n_e - P_e(\phi\cdot n_e)][\![\, v \,]\!]\,ds\right| \le \|\phi\cdot n_e - P_e(\phi\cdot n_e)\|_{L^2(e)}\|[\![\, v \,]\!]\|_{L^2(e)}.$$

From the preceding lemma we obtain the bound

$$\|\phi\cdot n_e - P_e(\phi\cdot n_e)\|_{L^2(e)} \le ch^{1/2}\|\operatorname{grad}(\phi\cdot n_e)\|_{L^2(e^*)},$$

where $h$ is the maximum triangle diameter and $e^*$ is the union of the one or two triangles containing $e$ (actually, here we could use either triangle, rather than the union, if we wished).

Next we bound $\|[\![ v ]\!]\|_{L^2(e)}$. On an interior edge, we may write

$$[\![ v ]\!] = [\![ v ]\!] - P_e[\![ v ]\!] = [v_+|_e - P_e(v_+|_e)] - [v_-|_e - P_e(v_-|_e)].$$

Applying the previous lemma to each piece to get

$$\|[\![ v ]\!] - P_e[\![ v ]\!]\|_{L^2(e)} \le ch^{1/2} \|\operatorname{grad}_h v\|_{L^2(e^*)}.$$

The same holds on a boundary edge, by a similar argument. Putting the bounds together, we get

$$\left| \int_e (\phi \cdot n_e)[\![ v ]\!]\, ds \right| \le ch \|\operatorname{grad} \phi\|_{L^2(e^*)} \|\operatorname{grad}_h v\|_{L^2(e^*)},$$

where $h$ is the maximum element size. Then we sum over all edges $e$, using

$$\begin{aligned}\sum_e \|\operatorname{grad} \phi\|_{L^2(e^*)} \|\operatorname{grad} v\|_{L^2(e^*)} &\le \left[ \sum_e \|\operatorname{grad} \phi\|^2_{L^2(e^*)} \right]^{1/2} \left[ \sum_e \|\operatorname{grad} v\|^2_{L^2(e^*)} \right]^{1/2} \\ &\le 3 \|\operatorname{grad} \phi\|_{L^2(\Omega)} \|\operatorname{grad}_h v\|_{L^2(\Omega)}.\end{aligned}$$

where the 3 comes from the fact that each triangle is contained in $e^*$ for 3 edges. Thus

$$\sum_{T \in \mathcal{T}_h} \int_{\partial T} (\phi \cdot n_T) v \, ds = \sum_e \int_e (\phi \cdot n_e)[\![ v ]\!]\, ds \le Ch \|\operatorname{grad} \phi\|_{L^2(\Omega)} \|\operatorname{grad} v\|_{L^2(\Omega)}.$$

□

We have proven $O(h)$ convergence for the nonconforming $\mathcal{P}_1$ FEM in the norm $\|\cdot\|_h$, i.e., the broken $H^1$ seminorm. On $\mathring{H}^1(\Omega)$, the $H^1$ seminorm bounds the $L^2$ norm (Poincaré–Friedrichs inequality), but this does not immediately apply to $V_h$. However we can use Lemma 7.3 to show that the analogue of the Poincaré–Friedrichs inequality does indeed hold.

THEOREM 7.5 (Discrete Poincaré–Friedrichs inequality). *There exists $c > 0$ such that*

$$\|v\| \le c \|v\|_h, \quad v \in \mathring{H}^1(\Omega) + V_h.$$

PROOF. Choose a function $\phi \in H^1(\Omega; \mathbb{R}^2)$ such that $\operatorname{div} \phi = v$ and $\|\phi\|_1 \le c\|v\|$ (e.g., take $\psi \in H^2$ with $\Delta \psi = v$ and set $\phi = \operatorname{grad} \psi$. Even if the domain is not convex, we can extend $v$ by zero to a larger convex domain and solve a Dirichlet problem there to get $\psi$). Then

$$\|v\|^2 = \int_\Omega \operatorname{div} \phi \, v \, dx = -\int_\Omega \phi \cdot \operatorname{grad}_h v \, dx + \sum_T \int_T (\phi \cdot n_h) v \, ds.$$

Clearly

$$\left| \int_\Omega \phi \cdot \operatorname{grad}_h v \, dx \right| \le \|\phi\| \|\operatorname{grad}_h v\| \le c\|v\| \|v\|_h.$$

By Lemma 7.3,

$$\left| \sum_T \int_T (\phi \cdot n_h) v \, ds \right| \le ch \|\phi\|_1 \|v\|_h \le c\|v\| \|v\|_h.$$

The theorem follows. □

We have shown the that the nonconforming $\mathcal{P}_1$ finite element method satisfies the same kind of $H^1$ bound as the conforming $\mathcal{P}_1$ finite element method. We now obtain a higher order error estimate in $L^2$ using a duality argument just as we did for the conforming method.

THEOREM 7.6. *Assuming (in addition to the hypotheses of Theorem 7.2) that the domain is convex,*

$$\|u - u_h\| \le ch^2 \|u\|_2.$$

PROOF. Define $\phi$ by the Dirichlet problem

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

Ellipitic regularity tells us that $\phi \in H^2$ and $\|\phi\|_2 \le c\|u - u_h\|$. Then

$$\|u - u_h\|^2 = -\int \Delta\phi(u - u_h)\,dx = \int \operatorname{grad}\phi \cdot \operatorname{grad}_h(u - u_h)\,dx - E_h(\phi, u - u_h). \tag{7.3}$$

Now let $v$ be any conforming finite element approximation in $\mathring{H}^1$, i.e., any continuous piecewise linear function vanishing on the boundary. Then

$$\int \operatorname{grad}_h(u - u_h) \operatorname{grad} v\,dx = 0.$$

Therefore we can bound the first term on the right hand side of (7.3):

$$\begin{aligned}\left|\int \operatorname{grad}\phi \cdot \operatorname{grad}_h(u - u_h)\,dx\right| &\le \left|\int \operatorname{grad}(\phi - v) \cdot \operatorname{grad}_h(u - u_h)\,dx\right| \\ &\le \|\operatorname{grad}(\phi - v)\| \|\operatorname{grad}_h(u - u_h)\|.\end{aligned}$$

Choosing $v$ to be the interpolant of $\phi$ gives

$$\left|\int \operatorname{grad}\phi \cdot \operatorname{grad}_h(u - u_h)\,dx\right| \le ch\|\phi\|_2 \|u - u_h\|_h \le ch\|u - u_h\| \|u - u_h\|_h.$$

For the second term on the right hand side of (7.3), we have by Theorem 7.1 that

$$|E_h(\phi, u - u_h) \le ch\|\phi\|_2 \|u - u_h\|_h \le ch\|u - u_h\| \|u - u_h\|_h.$$

Thus (7.3) becomes

$$\|u - u_h\|^2 \le ch\|u - u_h\| \|u - u_h\|_h,$$

which gives $\|u - u_h\| \le ch\|u - u_h\|_h$, and so the theorem. □

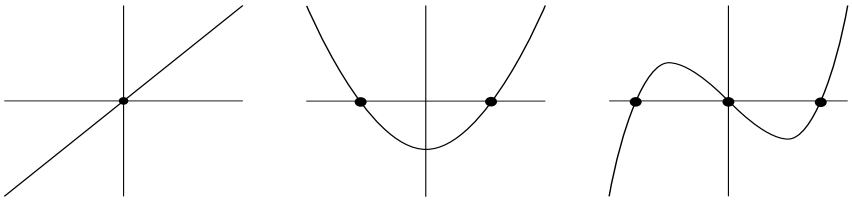
**1.1. Nonconforming spaces of higher degree.** We close this section by discussing the generalization to higher degree nonconforming elements. For $r > 0$, the nonconforming $\mathcal{P}_r$ space is defined

$$V_h = \{\, v \in L^2(\Omega) \mid v|_T \in \mathcal{P}_r(T)\ \forall T \in \mathcal{T}_h, \quad [\![\, v \,]\!] \perp \mathcal{P}_{r-1}(e)\ \forall \text{ edges } e \,\}. \tag{7.4}$$

For $r = 1$, this is the nonconforming piecewise linear space we just discussed, since a linear function is orthogonal to constants on an interval if and only if it vanishes at the midpoint. For $r = 2$, we can define a unique (up to a constant multiple) quadratic function on an interval $e$ which is orthogonal to $\mathcal{P}_1(e)$. This is the Legendre polynomial, and its zeros are the 2 Gauss points on the interval (if the interval is $[-1, 1]$ the Legendre polynomial is $(3x^2 - 1)/2$, and the 2 Gauss points are $\pm 1/\sqrt{3}$. It is easy to see that a quadratic polynomial is orthogonal to $\mathcal{P}_1$ if and only if it vanishes at the 2 Gauss points. More generally a

polynomial of degree $r$ is orthogonal to $\mathcal{P}_{r-1}$ if and only if it is a multiple of the $r$th degree Legendre polynomial, if and only if it vanishes at the $r$ Gauss points (zeros of the $r$th degree Legendre polynomial). See Figure 7.2.

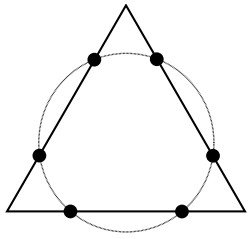
FIGURE 7.2. Legendre polynomials of degree 1, 2, and 3, and their roots, the Gauss points.



The analysis we gave above for nonconforming $\mathcal{P}_1$ extends easily to nonconforming $\mathcal{P}_r$.

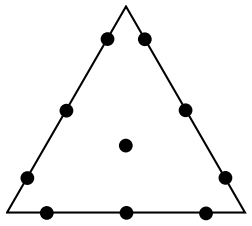
There is however one issue. The space $V_h$ defined by (7.4) is a finite element space, definable through shape functions and DOFs, for $r$ odd, but *not* for $r$ even. To see what goes wrong in the even case, take $r = 2$. The shape function space is, of course, $\mathcal{P}_2(T)$, and the natural choice of DOFs is the value at the 2 Gauss points on each edge. This gives $6 = \dim \mathcal{P}_2(T)$ DOFs, but *they are not unisolvent*. In fact, consider the case where the triangle is equilateral with its barycenter at the origin. Then all 6 of the Gauss points lie on a circle through the origin, so there is a nonzero quadratic polynomial, $x_1^2 + x_2^2 - c^2$, for which all the DOFs vanish. See Figure 7.3. (Despite the fact that the nonconforming $\mathcal{P}_2$ space is not a finite element space, in the strict sense of the word, it turns out that it is possible to implement it in a practical fashion, and it is occasionally used. It is called the Fortin-Soulie element [sic].)

FIGURE 7.3. The Gauss point values are not unisolvent over $\mathcal{P}_2(T)$.



This problem does not occur for nonconforming $\mathcal{P}_3$, for which we choose as DOFs the values as the 3 Gauss points on each side and the value of the barycenter (scaled to the interval $[-1, 1]$ the cubic Legendre polynomial is $(5x^3 - 3x)/2$ so the three Gauss points are $\pm\sqrt{3/5}$ and 0). See Figure 7.4. To see that these are unisolvent, suppose that a cubic vanishes $v$ at all of them. On each edge $e$, $v$ vanishes at the three Gauss points, so the restriction of $v$ to each edge is a constant multiple of the Legendre polynomial on the edge. Now let $p_i$, $i = 1, 2, 3$, denote the vertices. Since $v$ is a multiple of the Legendre polynomial on the edge from $p_1$ to $p_2$, $v(p_1) = -v(p_2)$. Similarly $v(p_2) = -v(p_3)$ and $v(p_3) = -v(p_1)$. Therefore $v(p_1) = -v(p_1)$, $v(p_1) = 0$. From this we easily get that $v \equiv 0$ on the boundary

FIGURE 7.4. The $\mathcal{P}_3$ nonconforming element.

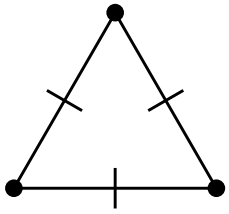


of $T$. Thus $v$ is a multiple of the bubble function on $T$, and so the DOF at the barycenter implies $v \equiv 0$.

## 2. Nonconforming finite elements for the plate equation

A number of different nonconforming finite element methods have been devised for the plate equation. (Some were proposed in the literature but later found not to converge or to converge only for special mesh families.) We shall consider only one here, the very clever Morley element. The shape functions for this element are $\mathcal{P}_2(T)$, and the DOFs are the values at the vertices and the normal derivatives at the midpoints of edges. To see that these DOFs

FIGURE 7.5. The Morley nonconforming plate element.



are unisolvent, suppose that $v \in \mathcal{P}_2(T)$ has vanishing DOFs. Note that a quadratic that vanishes at the endpoints of an interval has a vanishing derivative at the midpoint. Therefore at the midpoints of the edges, not just the normal derivatives vanish, but also the tangential derivatives, so the entire gradient vanishes. Each component of the gradient is a linear polynomial, which vanishes at the three midpoint, so the gradient vanishes. Therefore $v$ is constant, and so zero.

Now let $V_h$ denote the assembled Morley finite element space, approximating $\mathring{H}^2$ (so that we take all the DOFs on the boundary to be zero). For simplicity we consider the clamped plate problem with 0 Poisson ratio: $u \in \mathring{H}^2$,

$$\int \nabla^2 u : \nabla^2 v \, dx = \int f v \, dx, \quad v \in \mathring{H}^2.$$

The Morley finite element solution $u_h \in V_h$ is defined by

$$\int \nabla_h^2 u_h : \nabla_h^2 v \, dx = \int f v \, dx, \quad v \in V_h.$$

As before, the error analysis will hinge on the consistency error

$$E_h(u,v) := \sum_T \int_T \nabla^2 u : \nabla^2 v \, dx - \int f v \, dx, \quad v \in V_h.$$

Since $\operatorname{div} \nabla^2 u = \operatorname{grad} \Delta u$, we can write

$$\begin{aligned} E_h(u,v) = &\sum_T \left( \int_T \nabla^2 u : \nabla^2 v \, dx + \int_T \operatorname{div} \nabla^2 u \cdot \operatorname{grad} v \, dx \right) \\ &+ \sum_T \left( - \int_T \operatorname{grad} \Delta u \cdot \operatorname{grad} v \, dx - \int_T f \, v \, dx \right) =: E_1 + E_2. \end{aligned}$$

Note that $E_2$ vanishes if $v \in \mathring{H}^1(\Omega)$. For any $v$ belonging to the Morley space $V_h$, let $I_h v$ be the piecewise linear function with the same vertex values as $v$, so $I_h v \in \mathring{H}^1$. Therefore

$$E_2 = \sum_T \left( - \int_T \operatorname{grad} \Delta u \cdot \operatorname{grad}(v - I_h v) \, dx - \int_T f \, (v - I_h v) \, dx \right).$$

By standard approximation properties we have

$$\|v - I_h v\| \le c h^2 \|v\|_h, \quad \|\operatorname{grad}_h (v - I_h v)\| \le c h \|v\|_h.$$

Hence

$$|E_2| \le c(h\|u\|_3 + h^2 \|f\|) \|v\|_h.$$

For $E_1$, since

$$\int_T \nabla^2 u : \nabla^2 v \, dx = - \int_T \operatorname{div} \nabla^2 u \cdot \operatorname{grad} v \, dx + \int_{\partial T} (\nabla^2 u) n_T \cdot \operatorname{grad} v \, ds,$$

we get

$$E_1 = \sum_T \int_{\partial T} (\nabla^2 u) n_T \cdot \operatorname{grad} v \, ds$$

Now each component of $\operatorname{grad}_h v$ is a nonconforming $\mathcal{P}_1$, so we can applying Lemma 7.3 with $\phi$ replaced by $\nabla^2 u$ and $v$ replaced by $\operatorname{grad}_h v$ to get

$$|E_1| \le c h \|u\|_3 \|v\|_h.$$

Thus we have shown that

$$|E_h(u,v)| \le c h (\|u\|_3 + h\|f\|) \|v\|_h.$$

From this point the analysis is straightforward and leads to

$$\|u - u_h\|_h \le c h (\|u\|_3 + h\|f\|).$$

Note that the order $h$ estimate is what we would expect since the norm is a broken $H^2$ seminorm. The regularity required on $u$ is just a bit more than $u \in H^3$.

This result was established by Rannacher in 1979. In 1985 Arnold and Brezzi used a duality argument to prove an $O(h^2)$ broken $H^1$ estimate:

$$\|\operatorname{grad}_h (u - u_h)\| \le c h^2 (\|u\|_3 + \|f\|).$$

It is *not* true that $\|u - u_h\| = O(h^3)$.

CHAPTER 8

# Mixed finite element methods

The Kirchhoff plate problem is difficult to solve by finite elements since it is a fourth order PDE, leading to the need for finite element spaces contained in $H^2$. One way we might avoid this would be to formulate the fourth order PDE as a system of lower order PDEs. For example, we can write the biharmonic $\Delta^2 w = f$ as

$$M = \nabla^2 w, \quad \operatorname{div}\operatorname{div} M = f,$$

i.e.,

$$M_{ij} = \frac{\partial^2 w}{\partial x_i \partial x_j}, \quad \sum_{ij} \frac{\partial^2 M_{ij}}{\partial x_i \partial x_j} = f.$$

Actually, for plate problem with bending modulus $D$ and Poisson ratio $\nu$, a more physical way to do this—and one which will be more appropriate when supplying boundary conditions—is to define the *bending moment tensor*

$$M = D[(1-\nu)\nabla^2 w + \nu \Delta w) I],$$

i.e.,

$$M_{ij} = D[(1-\nu)\frac{\partial^2 w}{\partial x_i \partial x_j} + \nu(\Delta w)\delta_{ij}],$$

which, together with $\operatorname{div}\operatorname{div} M = f$ gives the plate equation. Of course, there are other ways to factor the fourth order problem into lower order problems, including the obvious $\phi = \Delta w$, $\Delta \phi = f$. We could even factor the problem into a system of four first order equations:

$$\theta = \operatorname{grad} w, \quad M = D[(1-\nu)\nabla\theta + \nu(\operatorname{div}\theta) I], \quad \zeta = \operatorname{div} M, \quad \operatorname{div}\zeta = f.$$

All the variables in this formulation are physically meaningful: $w$ is the vertical displacement of the plate, $\theta$ the rotation of vertical fibers, $M$ the bending moment tensor, and $\zeta$ the shear stress.

For any such factorization, we can introduce a weak formulation, and then try to discretize by finite elements. Such weak formulations are called *mixed* because they mix together fields of different types in the same equation. The resulting finite element methods are called mixed finite element methods.

In this chapter we will study mixed finite element methods, but for simpler problems, like Poisson's equation. Thus we will be reducing a second order equation to a system of first order equations. The motivation for doing this (besides as a way to gain insight for higher order problems) may not be clear, but it turns out that such mixed methods are of great use.

## 1. Mixed formulation for Poisson's equation

We start with the simplest problem

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \tag{8.1}$$

We have discussed finite element methods based on the corresponding weak formulation. The associated variational formulation is a minimization problem over $\mathring{H}^1(\Omega)$. Now we consider introducing a new variable $\sigma = \operatorname{grad} u$ (which is vector-valued), so we have the system

$$\sigma = \operatorname{grad} u \text{ in } \Omega, \quad -\operatorname{div} \sigma = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega.$$

To obtain a weak formulation, we multiply the first PDE by a vector-valued test function $\tau$ and the second by a scalar test function $v$, and integrate over $\Omega$. We now proceed as follows. First, we integrate the gradient in the first equation by parts, and use the boundary condition. This leads to the weak formulation: find $\sigma$ and $u$ such that

$$\int_\Omega \sigma \cdot \tau \, dx + \int_\Omega u \operatorname{div} \tau \, dx = 0 \quad \forall \tau, \qquad \int_\Omega \operatorname{div} \sigma v \, dx = -\int_\Omega f v \, dx \quad \forall v.$$

Note that we do not integrate by parts in the second equation, and we multiplied it by $-1$. The reason is to obtain a symmetric bilinear form. That is, if we add the two equations, we obtain a bilinear form acting on the trial function $(\sigma, u)$ and the test function $(\tau, v)$ which is symmetric: find $(\sigma, u)$ such that

$$B((\sigma, u), (\tau, v)) = \int_\Omega \sigma \cdot \tau \, dx + \int_\Omega u \operatorname{div} \tau \, dx + \int_\Omega \operatorname{div} \sigma v \, dx \quad \forall (\tau, v). \tag{8.2}$$

This reflects the fact that the original boundary value problem (8.1) is self-adjoint.

What are the correct spaces to use with this formulation? We see that the trial function $u$ and the corresponding test function $v$ enter undifferentiated. Therefore the appropriate Hilbert space is $L^2(\Omega)$. On the other hand, we need to integrate not products involving $\sigma$ and $\tau$, but also products involving $\operatorname{div} \sigma$ and $\operatorname{div} \tau$. Therefore we need $\tau \in L^2(\Omega; \mathbb{R}^2)$ and also $\operatorname{div} \tau \in L^2(\Omega)$ (and similarly for $\sigma$). We therefore define a new Hilbert space

$$H(\operatorname{div}) = H(\operatorname{div}, \Omega) = \{ \tau \in L^2(\Omega; \mathbb{R}^2) \mid \operatorname{div} \tau \in L^2(\Omega) \}.$$

As an example of a function in $H(\operatorname{div})$ we may take $\sigma = \operatorname{grad} u$, where $u \in H^1$ solves Poisson's equation $-\Delta u = f$ for some $f \in L^2$. Since $\operatorname{div} \sigma = -f$, we see that $\sigma \in H(\operatorname{div})$. It may be that $\sigma \notin H^1(\Omega; \mathbb{R}^2)$. This usually happens, for instance, for the Dirichlet problem for a nonconvex polygon.

Thus the mixed weak formulation of the Dirichlet problem for Poisson's equation is: Find $\sigma \in H(\operatorname{div})$ and $u \in L^2$ such that

$$\int_\Omega \sigma \cdot \tau \, dx + \int_\Omega u \operatorname{div} \tau \, dx = 0 \quad \forall \tau \in H(\operatorname{div}), \qquad \int_\Omega \operatorname{div} \sigma v \, dx = -\int_\Omega f v \, dx \quad \forall v \in L^2. \tag{8.3}$$

We have shown that the solution to the Dirichlet problem does indeed satisfy this system. We shall see below that there is a unique solution to this system for any $f \in L^2$. Thus this is a well-posed formulation of the Dirichlet problem. We may, of course, write it using a single bilinear form $B$, as in (8.2), and the Hilbert space $H(\operatorname{div}) \times L^2$.

The weak formulation is associated to a variational formulation as well. Namely if we define

$$\mathcal{L}(\tau, v) = \frac{1}{2}\int_\Omega |\tau|^2\,dx + \int_\Omega v \operatorname{div}\tau\,dx + \int_\Omega fv\,dx, \tag{8.4}$$

then $(\sigma, u)$ is the unique critical point of $\mathcal{L}$ over $H(\operatorname{div}) \times L^2$. In fact,

$$\mathcal{L}(\sigma, v) \le \mathcal{L}(\sigma, u) \le \mathcal{L}(\tau, u)\ \forall \tau \in H(\operatorname{div}), v \in L^2,$$

so $(\sigma, u)$ is a *saddle point* of $\mathcal{L}$.

Note that $\operatorname{div}\sigma = -f$, so $\mathcal{L}(\sigma, u) = \frac{1}{2}\int |\sigma|^2\,dx$. If $\tau \in H(\operatorname{div})$ is another function with $\operatorname{div}\tau = -f$, then $\mathcal{L}(\tau, u) = \frac{1}{2}\int |\tau|^2\,dx$. Thus

$$\frac{1}{2}\int |\sigma|^2\,dx \le \frac{1}{2}\int |\tau|^2.$$

The quantity $(1/2)\int |\tau|^2$ is called the complementary energy. We have just shown that, *subject to the constraint* $\operatorname{div}\tau = -f$ *the unique minimizer of the complementary energy* $(1/2)\int |\tau|^2$ *is* $\tau = \sigma$. Now recall how one computes the minimum of a function $J(\tau)$ subject to a constraint $L(\tau) = 0$. One introduces another variable $v$ of the same type as $L(\tau)$, and seeks a critical point of the extended function $J(\tau) + \langle L(\tau), v\rangle$ (where the angular brackets denote the inner product). If $(\tau, v) = (\sigma, u)$ is the critical point of the extended functional, that $\sigma$ is the minimizer of $J(\tau)$ subject to the constraint $L(\tau) = 0$. In our case, $L(\tau) = \operatorname{div}\tau + f \in L^2$, so the extended functional is

$$\frac{1}{2}\int |\tau|^2\,dx + \int (\operatorname{div}\tau + f)v\,dx, \quad \tau \in H(\operatorname{div}), v \in L^2,$$

which is exactly $\mathcal{L}(\tau, v)$. Thus we find that the variational formulation of the mixed method exactly characterizes $\sigma$ as the minimizer of the complementary energy, and $u$ as the Lagrange multiplier associated to the divergence constraint.

## 2. A mixed finite element method

A Galerkin method for the Poisson equation now proceeds as follows. We choose finite dimensional subspaces $V_h \subset H(\operatorname{div})$ and $W_h \subset L^2$, and seek $\sigma_h \in W_h$, $u_h \in V_h$ such that

$$\int_\Omega \sigma_h \cdot \tau\,dx + \int_\Omega u_h \operatorname{div}\tau\,dx = 0 \quad \forall \tau \in V_h, \qquad \int_\Omega \operatorname{div}\sigma_h v\,dx = -\int_\Omega fv\,dx \quad \forall v \in W_h. \tag{8.5}$$

This is simply Galerkin's method applied to the mixed formulation. However the bilinear form $B$ in the mixed formulation is not coercive, and so our theory thus far does not imply that this method is stable.

Let us try out the method in a simple case. We consider the problem on the unit square, with a uniform mesh of $n \times n$ subsquares, each divided in two by its positively sloped diagonal. For finite elements we consider three possibilities:

- continuous piecewise linear vector fields for $V_h$, continuous piecewise linear scalar fields for $W_h$;
- continuous piecewise linear vector fields for $V_h$, piecewise constants for $W_h$;
- the Raviart–Thomas elements, a subspace of $H(\operatorname{div})$ we shall study below for $V_h$, and piecewise constants for $W_h$.

The first possibility, Lagrange elements for both variables, is a complete failure, in the sense that the resulting matrix is singular. To see this, consider taking $u$ as the piecewise linear function with the vertex values shown in Figure 8.1, where $a$, $b$, and $c$ are any three real numbers adding to 0 (a 2-dimensional space). Then we have that $\int_T u\,dx = 0$ for each triangle $u$. Therefore $u$ is orthogonal to piecewise constants, and so $\int u \operatorname{div} \tau\,dx = 0$ for all continuous piecewise linear $\tau$. Therefore $(0, u) \in V_h \times W_h$ satisfies

$$B((0,u),(\tau,v)) = 0, \quad (\tau,v) \in V_h \times W_h,$$

i.e., $(0, u)$ belongs to the kernel of the stiffness matrix. Thus the stiffness matrix is singular.

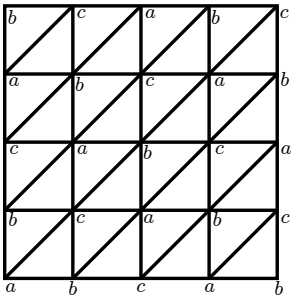



FIGURE 8.1. A piecewise linear which is orthogonal to all piecewise constants ($a + b + c = 0$).

The other two methods both lead to nonsingular matrices. To compare them, we choose a very simple problem: $u = x(1-x)y(1-y)$, so $f = 2[x(1-x)+y(1-y)]$. Figure 8.2 shows the variable $u$ for the two cases. Notice that the method using Lagrange elements for $\sigma$ gives complete nonsense. The solution is highly oscillatory on the level of the mesh, it ranges from $-0.15$ to 0.25, while the true solution is in the range from 0 to 0.0625, and it has a line of near zeros down the main diagonal, which is clearly an artifact of the particular mesh. The Raviart–Thomas method gives a solution $u$ that is a reasonably good approximation to the true solution (considering it is a piecewise constant).

Clearly the choice of elements for mixed methods is very important. This is not a question of approximation or consistency, but rather stability.

In fact, the issue already arises in one dimension. Consider the Poisson equation ($-u'' = f$) on an interval, say $(-1, 1)$, written as $\sigma = u'$, $-\sigma' = f$. Assuming homogeneous Dirichlet boundary conditions, we get the mixed formulation: find $\sigma \in H^1$, $u \in L^2$ such that

$$\int_{-1}^{1} \sigma\tau\,dx + \int_{-1}^{1} \tau' u\,dx = 0, \quad \tau \in H^1, \qquad \int_{-1}^{1} \sigma' v\,dx = -\int_{-1}^{1} fv\,dx, \quad v \in L^2.$$

Notice that in one dimension $H^1 = H(\mathrm{div})$. If we again consider the possibility of continuous piecewise linear functions for both variables, we again obtain a singular matrix. However in one dimension, the choice of continuous piecewise linears for $\sigma$ and piecewise constants for $u$ works just fine. In fact, this method is the 1-D analogue of the Raviart–Thomas method. In Figure 8.3 we compare this method, and the method we obtain by using continuous piecewise *quadratics* for $\sigma$ and piecewise constants for $u$. That method is clearly unstable. (Our test problem has $u(x) = \cos(\pi x/2)$.)

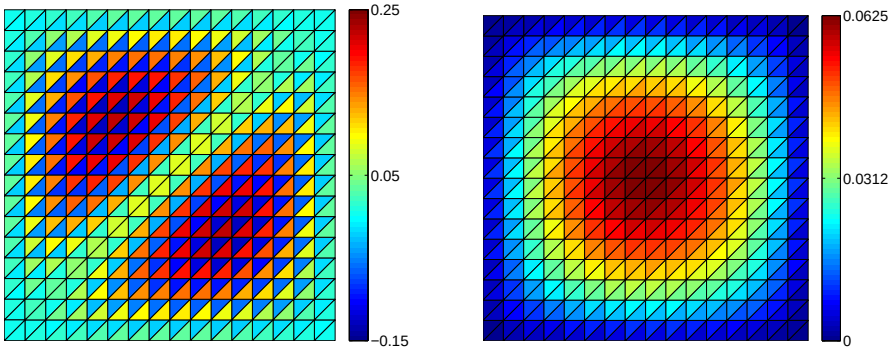



FIGURE 8.2. Approximation of the mixed formulation for Poisson's equation using piecewise constants for $u$ and for $\sigma$ using either continuous piecewise linears (left), or Raviart–Thomas elements (right). The plotted quantity is $u$ in each case.

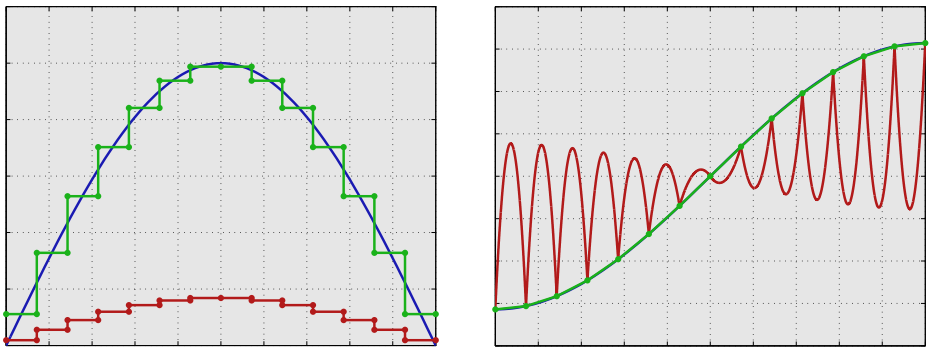


FIGURE 8.3. Approximation of the mixed formulation for $-u'' = f$ in one dimension with two choices of elements, piecewise constants for $u$ and piecewise linears for $\sigma$ (a stable method, shown in green), or piecewise constants for $u$ and piecewise quadratics for $\sigma$ (unstable, shown in red). The left plot shows $u$ and the right plot shows $\sigma$, with the exact solution in blue. (In the right plot, the blue curve essentially coincides with the green curve and hence is not visible.)

An important goal is to understand what is going on these examples. How can we tell which elements are stable for the mixed formulation? How can we find stable elements?

## 3. Inhomogeneous Dirichlet boundary conditions

Before continuing, we consider some other problems. Since the Dirichlet boundary condition is natural in the mixed form, an inhomogeneous Dirichlet condition $u = g$ on $\partial\Omega$, just modifies the right hand side. Here, to make things a bit more interesting, let us also introduce a coefficient $a$ in our equation:

$$-\operatorname{div} a \operatorname{grad} u = f.$$

We assume that $a(x)$ is bounded above and below by a positive constant. To obtain the weak formulation, we introduce the new variable $\sigma = a \operatorname{grad} u$. We write the system as

$$\alpha\sigma - \operatorname{grad} u = 0, \quad \operatorname{div} \sigma = -f,$$

where $\alpha = a^{-1}$. The reason for writing the first equation with $\alpha$ rather than $a$, is that this will lead to a symmetric system, associated to a variational principle. Now if we multiply the first equation by $\tau \in H(\operatorname{div})$, integrate by parts, and use the Dirichlet boundary condition, we get

$$\int \alpha\sigma \cdot \tau \, dx + \int \operatorname{div} \tau u \, dx = \int_{\partial\Omega} \tau \cdot n g \, dx, \quad \tau \in H(\operatorname{div}),$$

The equilibrium equation remains unchanged

$$\int \operatorname{div} \sigma v \, dx = -\int f v \, dx, \quad v \in L^2.$$

This is again of the form

$$B((\sigma, u), (\tau, v)) = F(\tau, v) \quad (\tau, v) \in H(\operatorname{div}) \times L^2,$$

but now the linear functional $F$ acts on both variables.

## 4. The Neumann problem

We next consider the Neumann boundary condition $a\, \partial u/\partial n = 0$. If we write the PDE as the first order system

$$\alpha\sigma - \operatorname{grad} u = 0, \quad \operatorname{div} \sigma = -f,$$

then the boundary condition is $\sigma \cdot n = 0$ on $\partial\Omega$. Now if we multiply the first equation by $\tau \in H(\operatorname{div})$ and integrate by parts, the boundary term $\int_{\partial\Omega} u\tau \cdot n \, ds$ will not vanishes unless $\tau \cdot n$ vanishes on the boundary. Thus we are led to incorporate the Neumann boundary condition into the space for $\sigma$ and $\tau$, and we define the space

$$\mathring{H}(\operatorname{div}) = \{\, \tau \in H(\operatorname{div}) \,|\, \tau \cdot n = 0 \text{ on } \partial\Omega \,\}.$$

To do so, we need to make sure that the normal trace $\tau \cdot n$ makes sense for $\tau \in H(\operatorname{div})$. We shall return to this point, but let us accept it for now.

In this way we obtain a weak formulation for the Neumann problem: find $\sigma \in \mathring{H}(\operatorname{div})$, $u \in L^2$ such that

$$\int \alpha\sigma \cdot \tau \, dx + \int \operatorname{div} \tau u \, dx = 0, \quad \tau \in \mathring{H}(\operatorname{div}), \qquad \int \operatorname{div} \sigma v \, dx = -\int f v \, dx, \quad v \in L^2.$$

This problem is not well-posed, nor should it be, since the Neumann problem is not well-posed. To have a solution we need $\int f = 0$ (take $v \equiv 1$), and then the solution is undetermined up to addition of a constant. To get a well-posed problem, we replace $L^2$ with

$$\hat{L}^2 = \{\, v \in L^2 \,|\, \int v = 0 \,\}.$$

This leads to a well-posed problem (as we shall see below). Thus the solution of the Neumann problem is a saddle point of $\mathcal{L}$ over $\mathring{H}(\operatorname{div}) \times \hat{L}^2$.

Note that the Neumann boundary conditions are built into the space used for the weak and variational form ($\mathring{H}(\operatorname{div})$). Thus they are essential boundary conditions, while Dirichlet

boundary conditions were natural. In this, the mixed formulation has the opposite behavior as the standard one.

To complete this section, we show how to define the normal trace $\tau \cdot n$ on $\partial\Omega$ for $\tau \in H(\mathrm{div})$. First we begin by giving a name to the trace space of $H^1(\Omega)$. Define

$$H^{1/2}(\partial\Omega) = \{\, u|_{\partial\Omega} \,|\, v \in H^1(\Omega) \,\}.$$

Then $H^{1/2}$ is a subspace of $L^2(\partial\Omega)$. If we define the norm

$$\|g\|_{H^{1/2}(\partial\Omega)} = \inf_{\substack{v\in H^1(\Omega)\\ u|_{\partial\Omega}=g}} \|v\|_1,$$

then, by definition, the trace operator is bounded $H^1(\Omega) \to H^{1/2}(\partial\Omega)$. This way of defining the trace space avoids many complications. Of course it would be nice to have a better intrinsic sense of the space. This is possible to obtain, but we will not pursue it here.

Now consider a vector function $\tau \in H^1(\Omega;\mathbb{R}^2)$, and a function $g \in H^{1/2}(\partial\Omega)$. We can find a function $v \in H^1(\Omega)$ with $v|_{\partial\Omega} = g$ and $\|v\|_1 \le 2\|g\|_{1/2,\partial\Omega}$ (we can even replace 2 by 1). Then

$$\int_{\partial\Omega} \tau \cdot n g \, ds = \int_\Omega \tau \cdot \operatorname{grad} v \, dx + \int_\Omega \operatorname{div} \tau v \, dx.$$

so

$$\left| \int_{\partial\Omega} \tau \cdot n g \, ds \right| \le c\|v\|_1 \|\tau\|_{H(\mathrm{div})} \le c\|g\|_{1/2,\partial\Omega} \|\tau\|_{H(\mathrm{div})}.$$

Now we define the $H^{-1/2}(\partial\Omega)$ norm of some $k \in L^2(\partial\Omega)$ by

$$\|k\|_{H^{-1/2}(\partial\Omega)} = \sup_{g\in H^{1/2}(\partial\Omega)} \frac{\int_{\partial\Omega} kg \, ds}{\|g\|_{H^{1/2}(\partial\Omega)}}.$$

Note that $\|k\|_{H^{-1/2}(\partial\Omega)} \le c\|k\|_{L^2(\partial\Omega)}$. With this definition we have that the map $\gamma : H^1(\Omega;\mathbb{R}^2) \to L^2(\partial\Omega)$ given by $\gamma\tau = \tau \cdot n$ satisfies

$$\|\gamma\tau\|_{H^{-1/2}(\partial\Omega)} \le c\|\tau\|_{H(\mathrm{div})}, \quad \tau \in H^1(\Omega;\mathbb{R}^2).$$

We can extend this result to all of $H(\mathrm{div})$ by density, but for this we need to define the space $H^{-1/2}(\partial\Omega)$ as the completion of $\gamma H^1(\Omega)$ in the $H^{-1/2}(\partial\Omega)$ norm. If we do that we have the following trace theorem.

THEOREM 8.1 (Trace theorem in $H(\mathrm{div})$). *The map $\gamma\tau = \tau \cdot n$ extends to a bounded linear map from $H(\mathrm{div})$ onto $H^{-1/2}(\partial\Omega)$.*

## 5. The Stokes equations

The Stokes equations seek a vector field $u$ and a scalar field $p$, such that

$$-\Delta u + \operatorname{grad} p = f, \quad \operatorname{div} u = 0.$$

No slip boundary conditions are $u = 0$ on the boundary, and no conditions on $p$. Note that in this equation $\Delta$ represents the vector Laplacian, applied to each component. We shall see that there is some similarity between this problem and the mixed Poisson equation, with $u$ here corresponding to $\sigma$ there and $p$ here to $u$ there.

The weak formulation of the Stokes equation is to find $u \in \mathring{H}^1(\Omega;\mathbb{R}^2)$, $p \in L^2$ such that

$$\int \operatorname{grad} u : \operatorname{grad} v \, dx - \int \operatorname{div} vp \, dx = \int fv \, dx, \quad v \in \mathring{H}^1(\Omega;\mathbb{R}^2),$$
$$\int \operatorname{div} uq \, dx = 0, \quad q \in L^2.$$

## 6. Abstract framework

All the problems considered in this section may be put in the following form. We have two Hilbert spaces $V$ and $W$, two bilinear forms

$$a : V \times V \to \mathbb{R}, \quad b : V \times W \to \mathbb{R},$$

and two linear forms

$$F : V \to \mathbb{R}, \quad G : W \to \mathbb{R}.$$

Then we consider the weak formulation, find $(\sigma, u) \in V \times W$ such that

$$\begin{aligned} a(\sigma,\tau) + b(\tau,u) &= F(\tau), \quad \tau \in V, \\ b(\sigma,v) &= G(v), \quad v \in W. \end{aligned} \tag{8.6}$$

For the Poisson equation, $V = H(\operatorname{div})$ and $a$ is the $L^2$ inner product (not the $H(\operatorname{div})$ inner product, or, in the case of a coefficient, a weighted $L^2$ inner product. For the Stokes equations, $V = H^1(\Omega;\mathbb{R}^2)$ and $a$ is the $H^1$ seminorm. In both cases $W = L^2$ and $b(\tau, v) = \int \operatorname{div} \tau v \, dx$. Besides these there are many other examples of this structure.

## 7. Duality

Before proceeding we recall some results from functional analysis. If $T : V \to W$ is a linear map between Hilbert (or Banach) spaces, then $T^* : W^* \to V^*$ is defined by

$$T^*(g)(v) = g(Tv).$$

Then $T^*$ is a bounded operator if $T$ is:

$$|T^*g(v)| = |g(Tv)| \le \|g\|_{W^*}\|Tv\|_W \le \|g\|_{W^*}\|T\|_{\mathcal{L}(V,W)}\|v\|_V,$$

so $\|T^*g\|_{V^*} \le \|g\|_{W^*}\|T\|_{\mathcal{L}(V,W)}$, which means that $\|T^*\|_{\mathcal{L}(W^*,V^*)} \le \|T\|_{\mathcal{L}(V,W)}$. Moreover if $S : W \to X$ is another bounded linear operator, then, directly from the definition, $(S \circ T)^* = T^* \circ S^*$. The dual of the identity operator $V \to V$ is the identity $V^* \to V^*$. This gives an immediate theorem about the dual of an invertible map.

THEOREM 8.2. *If a bounded linear operator $T : V \to W$ between Hilbert spaces is invertible, then $T^* : W^* \to V^*$ is invertible and $(T^*)^{-1} = (T^{-1})^*$.*

For the proof, we just take the dual of the equations $T \circ T^{-1} = I_W$ and $T^{-1} \circ T = I_V$.

Recall that a Hilbert space is reflexive: $(V^*)^* = V$ (where we think of $v \in V$ as acting on $V^*$ by $v(f) = f(v)$). Therefore $T^{**} = (T^*)^* : V \to W$. It is immediate that $T^{**} = T$: indeed for $v \in V$, $g \in W^*$, we have

$$g(T^{**}v) = (T^{**}v)g = v(T^*g) = T^*g(v) = g(Tv).$$

This allows us, whenever we have deduced a property of $T^*$ from a property of $T$ to reverse the situation, deducing a property of $T$ from one of $T^*$ just by applying the first result to

$T^*$ rather than $T$. For example,we have $\|T\|_{\mathcal{L}(V,W)} = \|T^{**}\|_{\mathcal{L}(V^{**},W^{**})} \leq \|T^*\|_{\mathcal{L}(W^*,V^*)}$, which gives the important result

$$\|T^*\|_{\mathcal{L}(W^*,V^*)} = \|T\|_{\mathcal{L}(V,W)}.$$

As another example, $T^*$ is invertible if *and only if* $T$ is invertible.

Now we introduce the notion of the annihilator of a subspace $Z$ in a Hilbert (or Banach) space $V$:

$$Z^a = \{\, f \in V^* \,|\, f(v) = 0 \ \forall v \in Z \,\} \subset V^*.$$

Note that the annihilator $Z^a$ is defined for any subspace of $V$, not just closed subspaces, but $Z^a$ is itself always closed. Of course we may apply the same notion to a subspace $Y$ of $V^*$ in which case the annihilator belongs to $V^{**} = V$ (in a Hilbert or reflexive Banach space) and can be written

$$Y^a = \{\, v \in V \,|\, f(v) = 0 \ \forall f \in Y \,\} \subset V.$$

If we start with a subspace $Z$ of $V$ and apply the annhilator twice, we obtain another subspace of $V$, this one closed. In fact

$$(Z^a)^a = \bar{Z},$$

the closure of $Z$ in $V$ (the smallest closed subspace containing $Z$). Indeed, it is obvious that $Z \subset (Z^a)^a$, and the latter is closed, so $\bar{Z} \subset (Z^a)^a$. On the other hand, if $v \in V$, $v \notin \bar{Z}$, then there exists $f \in V^*$ such that $f(z) = 0 \ \forall z \in Z$, but $f(v) \neq 0$, showing that $v \notin (Z^a)^a$.

Now suppose $T : V \to W$ is a bounded linear map of Hilbert spaces. Then the null space of $T$ is precisely the annihilator of the range of $T^*$:

$$\mathcal{N}(T) = \mathcal{R}(T^*)^a.$$

Indeed, for $v \in V$,

$$v \in \mathcal{N}(T) \iff Tv = 0 \iff g(Tv) = 0 \,\forall g \in W^* \iff T^*g(v) = 0 \,\forall g \in W^* \iff v \in \mathcal{R}(T^*)^a.$$

Replacing $T$ with $T^*$ we get $\mathcal{N}(T^*) = \mathcal{R}(T)^a$. Taking the annihilator of both sides we get

$$\overline{\mathcal{R}(T)} = \mathcal{N}(T^*)^a.$$

In summary:

THEOREM 8.3. *Let $T : V \to W$ be a bounded linear operator between Hilbert spaces. Then*

$$\mathcal{N}(T) = \mathcal{R}(T^*)^a \text{ and } \overline{\mathcal{R}(T)} = \mathcal{N}(T^*)^a.$$

COROLLARY 8.4. *$T$ is injective if and only if $T^*$ has dense range, and $T^*$ is injective if and only if $T$ has dense range.*

Thus far we have used the identification of $V$ with $V^{**}$, but we have not used the identification, given by the Riesz Representation Theorem, of $V$ with $V^*$. For this reason, the whole discussion so far carries over immediately to reflexive Banach spaces (and much of it to general Banach spaces). However we now use the identification of $V$ with $V^*$ given by the Riesz Representation Theorem, and really use the Hilbert space structure. This will allow us to give a very simple proof of the Closed Range Theorem (although the theorem is true for general Banach spaces). Let $Z$ be a closed subspace of a Hilbert space, with $i_Z : Z \to V$ and $\pi_Z : V \to Z$ the inclusion and the orthogonal projection, respectively.

What are $i_Z^* : V^* \to Z^*$ and $\pi_Z^* : Z^* \to V^*$? It is easy to see that the following diagrams commute

$$\begin{array}{ccc} Z & \xrightarrow{i_Z} & V \\ \downarrow\cong & & \downarrow\cong \\ Z^* & \xrightarrow{\pi_Z^*} & V^* \end{array} \qquad\qquad \begin{array}{ccc} V & \xrightarrow{\pi_Z} & Z \\ \downarrow\cong & & \downarrow\cong \\ V^* & \xrightarrow{i_Z^*} & Z^* \end{array}$$

where the vertical maps are the Riesz isomorphisms. This says, that $Z^*$ may be viewed simply as a subspace of $V^*$ with $\pi_Z^*$ the inclusion and $i_Z^*$ the orthogonal projection.

THEOREM 8.5 (Closed Range Theorem). *Let $T : V \to W$ be a bounded linear operator between Hilbert spaces. Then $\mathcal{R}(T)$ is closed in $W$ if and only if $\mathcal{R}(T^*)$ is closed in $V^*$.*

PROOF. Suppose $Y := \mathcal{R}(T)$ is closed in $W$. Let $Z = \mathcal{N}(T) \subset V$ and define the map $\tilde{T} : Z^\perp \to Y$ by restriction of both the domain and range ($\tilde{T}v = Tv \in Y$ for all $v \in Z^\perp$). Clearly the following diagram commutes:

$$\begin{array}{ccc} V & \xrightarrow{T} & W \\ \downarrow{\scriptstyle \pi_{Z^\perp}} & & \uparrow{\scriptstyle i_Y} \\ Z^\perp & \xrightarrow{\tilde{T}} & Y \end{array}$$

Taking duals we get the commuting diagram

$$\begin{array}{ccc} V^* & \xleftarrow{T^*} & W^* \\ \uparrow{\scriptstyle i_{(Z^\perp)^*}} & & \downarrow{\scriptstyle \pi_{Y^*}} \\ (Z^\perp)^* & \xleftarrow{\tilde{T}^*} & Y^* \end{array}$$

Now, $\tilde{T}$ is an isomorphism from $Z^\perp$ to $Y$, so $\tilde{T}^*$ is an isomorphism from $Y^*$ to $(Z^\perp)^*$. We can then read off the range of $T^*$ from the last diagram: it is just the closed subspace $(Z^\perp)^*$ of $V^*$.

Thus if $\mathcal{R}(T)$ is closed, $\mathcal{R}(T^*)$ is closed. Applying this result to $T^*$ we see if $\mathcal{R}(T^*)$ is closed, then $\mathcal{R}(T)$ is closed. □

COROLLARY 8.6. *$T$ is injective with closed range if and only if $T^*$ is surjective and vice versa.*

We close this section by remarking that, using the Riesz identification of $V$ and $V^*$, we may view the dual of $T : V \to W$ as taking $W \to V$ (this is sometimes called the Hilbert space dual, to distinguish it from the dual $W^* \to V^*$). In this view, Theorem 8.3 becomes

$$\mathcal{N}(T) = \mathcal{R}(T^*)^\perp \text{ and } \overline{\mathcal{R}(T)} = \mathcal{N}(T^*)^\perp.$$

A simple case is when $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ so $T$ can be viewed as an $m \times n$ matrix. Then clearly $\mathcal{N}(T)$ is the orthogonal complement of the span of the rows, i.e., the orthogonal complement of the span of columns of the transpose. Thus the fact that $\mathcal{N}(T) = \mathcal{R}(T^*)^\perp$ is completely elementary (but nonetheless very useful) in this case.

## 8. Well-posedness of saddle point problems

Consider now the abstract saddle point problem describe in Section 6. Associated to the bilinear forms $a$ and $b$, we have bounded bilinear operators $A : V \to V^*$ and $B : V \to W^*$, and the problem may be stated in operator form: given $F \in V^*$, and $G \in W^*$ find $\sigma \in V$, $u \in W$ such that

$$A\sigma + B^*u = F, \quad B\sigma = G.$$

We now establish when this problem is well-posed, i.e., for all $F$, $G$, there exists a unique solution $\sigma$, $u$, and there is a constant such that

$$\|\sigma\|_V + \|u\|_W \leq c(\|F\|_{V^*} + \|G\|_{W^*}). \tag{8.7}$$

THEOREM 8.7 (Brezzi's theorem in operator form). *Let $Z = \mathcal{N}(B)$ and define $A_{ZZ} : Z \to Z^*$ by $A_{ZZ} = \pi_{Z^*} \circ A|_Z$. The abstract saddle point problem is well-posed if and only if*

1. *$A_{ZZ}$ is an isomorphism of $Z$ onto $Z^*$.*
2. *$B$ maps $V$ onto $W^*$.*

*Moreover the well-posedness constant $c$ in* (8.7) *may be bounded above in terms of the $\|A\|$, $\|B\|$, $\|A_{ZZ}^{-1}\|$, and $\|B|_{Z^\perp}^{-1}\|$.*

PROOF. In addition to $A_{ZZ}$, define maps $A_{Z\perp} = \pi_{Z^{\perp *}} \circ A|_Z : Z \to Z^{\perp *}$ and, similarly, $A_{\perp Z}$ and $A_{\perp\perp}$. We also define $B_\perp = B|_{Z^\perp} : Z^\perp \to W^*$. (The corresponding $B_Z$ is just the zero map, so we don't introduce that notation.) If we partition $\sigma \in V = Z + Z^\perp$ as $\sigma_Z + \sigma_\perp$ and $F \in V^* = Z^* + Z^{\perp *}$ as $F_Z + F_\perp$, we may write the equations $A\sigma + B^*u = F$, $B\sigma = G$ in matrix form:

$$\begin{pmatrix} A_{ZZ} & A_{\perp Z} & 0 \\ A_{Z\perp} & A_{\perp\perp} & B_\perp^* \\ 0 & B_\perp & 0 \end{pmatrix} \begin{pmatrix} \sigma_Z \\ \sigma_\perp \\ u \end{pmatrix} = \begin{pmatrix} F_Z \\ F_\perp \\ G \end{pmatrix}. \tag{8.8}$$

Now reorder the unknowns, putting $u$ first, so the last column of the matrix moves in front of the first:

$$\begin{pmatrix} 0 & A_{ZZ} & A_{\perp Z} \\ B_\perp^* & A_{Z\perp} & A_{\perp\perp} \\ 0 & 0 & B_\perp \end{pmatrix} \begin{pmatrix} u \\ \sigma_Z \\ \sigma_\perp \end{pmatrix} = \begin{pmatrix} F_Z \\ F_\perp \\ G \end{pmatrix}.$$

Now reverse the first and second equation:

$$\begin{pmatrix} B_\perp^* & A_{Z\perp} & A_{\perp\perp} \\ 0 & A_{ZZ} & A_{\perp Z} \\ 0 & 0 & B_\perp \end{pmatrix} \begin{pmatrix} u \\ \sigma_Z \\ \sigma_\perp \end{pmatrix} = \begin{pmatrix} F_\perp \\ F_Z \\ G \end{pmatrix}. \tag{8.9}$$

From the upper triangular form of the matrix, we see that it is invertible if and only if all the three matrices on the diagonal are invertible. But $B_\perp$ is invertible if and only if $B$ is onto (since we restricted $B$ to the orthogonal complement of its kernel), and $B_\perp^*$ is invertible if and only if $B_\perp$ is. Therefore we have that (8.8) is invertible if and only if (1) and (2) hold.

When the conditions hold, we may write down the inverse matrix. Using the reordered form we have

$$\begin{pmatrix} B_\perp^{*-1} & -B_\perp^{*-1} A_{Z\perp} A_{ZZ}^{-1} & B_\perp^{*-1}(A_{Z\perp} A_{ZZ}^{-1} A_{\perp Z} - A_{\perp\perp}) B_\perp^{-1} \\ 0 & A_{ZZ}^{-1} & -A_{ZZ}^{-1} A_{\perp Z} B_\perp^{-1} \\ 0 & 0 & B_\perp^{-1} \end{pmatrix}$$

from which can give an explicit bound on the well-posedness constant. □

Now we return to the statement of the problem in terms of bilinear forms rather than operators. The operator $A_{ZZ}$ corresponds to the restriction of the bilinear form $a$ to $Z \times Z$. Thus we know that a sufficient condition for condition (1) above is that $a$ is coercive on $Z \times Z$, i.e., there exists $\gamma_1 > 0$ such that

$$a(z, z) \geq \gamma_1 \|z\|_V^2, \quad z \in Z. \tag{8.10}$$

This condition is referred to as *coercivity in the kernel* or the first Brezzi condition. It is not necessary, but usually sufficient in practice. If we prefer necessary and sufficient conditions, we need to use the inf-sup condition: for all $z_1 \in Z$ there exists $z_2 \in Z$ such that

$$a(z_1, z_2) \geq \gamma_1 \|z_1\| \|z_2\|,$$

together with the dense range condition: for all $0 \neq z_2 \in Z$ there exists $0 \neq z_1 \in Z$ such that

$$a(z_1, z_2) \neq 0.$$

Note that $\gamma_1^{-1}$ is a bound for $A_{ZZ}^{-1}$.

Next we interpret condition (2) of the theorem in terms of the bilinear form $b$. The condition is that $B$ maps $V$ onto $W^*$, which is equivalent to $B^*$ maps $W$ one-to-one onto a closed subspace of $V^*$, which is equivalent to the existence of a constant $\gamma_2 > 0$ with $\|B^* w\| \geq \gamma_2 \|w\|$ for all $w \in W$, which is equivalent to, that for all $w \in W$ there exists $0 \neq v \in V$ such that $b(v, w) = B^* w(v) \geq \gamma_2 \|w\| \|v\|$, or, finally:

$$\inf_{0 \neq v \in W_h} \sup_{0 \neq \tau \in V_h} \frac{b(\tau, v)}{\|\tau\| \|v\|} \geq \gamma_2. \tag{8.11}$$

In this case $\gamma_2^{-1}$ is a bound for $\|B_\perp^{-1}\|$. This is known as Brezzi's inf-sup condition, or the second Brezzi condition.

Putting things together we have proved:

THEOREM 8.8 (Brezzi's theorem). *The abstract saddle point problem is well-posed if*

1. *The bilinear form $a$ is coercive over the kernel, that is,* (8.10) *holds for some $\gamma_1 > 0$.*
2. *The Brezzi inf-sup condition* (8.11) *holds for some $\gamma_2 > 0$.*

*Moreover the well-posedness constant may be bounded above in terms of the $\|A\|$, $\gamma_1^{-1}$, and $\gamma_2^{-1}$.*

REMARK. Looking back at the inverse matrix we derived in the proof of Brezzi's theorem in operator form, we get explicit estimates:

$$\|\sigma\| \leq \gamma_2^{-1}(1+\|a\|\gamma_1^{-1})\|G\| + \gamma_1^{-1}\|F\|, \quad \|u\| \leq \gamma_2^{-2}\|a\|(1+\|a\|\gamma_1^{-1})\|G\| + \gamma_2^{-1}(1+\|a\|\gamma_1^{-1})\|F\|.$$

Let us now look at some examples. For the mixed form of the Dirichlet problem, $a : H(\mathrm{div}) \times H(\mathrm{div}) \to \mathbb{R}$ is $a(\sigma, \tau) = \int \alpha\sigma \cdot \tau \, dx$, and $b : H(\mathrm{div}) \times L^2 \to \mathbb{R}$ is $b(\tau, v) = \int \mathrm{div}\, \tau v \, dx$. Therefore $Z = \{ \tau \in H(\mathrm{div}) \mid \mathrm{div}\, \tau = 0 \}$, the space of divergence free vector fields. Clearly we have coercivity in the kernel:

$$a(\tau, \tau) \geq \underline{\alpha} \|\tau\|^2 = \underline{\alpha} \|\tau\|_{H(\mathrm{div})}^2.$$

Note that $a$ is *not* coercive on all of $H(\mathrm{div})$, just on the kernel.

For the second Brezzi condition we show that for any $v \in L^2$ we can find $\tau \in H(\mathrm{div})$ with $\operatorname{div}\tau = v$ and $\|\tau\|_{H(\mathrm{div})} \le c\|v\|$. There are many ways to do this. For example, we can extend $v$ by zero and then define a primitive:

$$\tau_1(x,y) = \int_0^x u(t,y)\,dt, \quad \tau_2 = 0.$$

Clearly $\operatorname{div}\tau = v$ and it is easy to bound $\|\tau\|$ in terms of $\|v\|$ and the diameter of the domain. Or we could solve a Poisson equation $\Delta u = v$ and set $\tau = \operatorname{grad} u$.

As a second example, we consider the Stokes problem. In this case we seek the vector variable (which we now call $u$) in $\mathring{H}^1(\Omega;\mathbb{R}^2)$. It is not true that div maps this space onto $L^2$, but almost. Clearly $\int \operatorname{div} u\,dx = 0$ for $u \in \mathring{H}^1$, so to have the surjectivity of $B$ we need to take the pressure space as

$$\hat{L}^2 = \{\, p \in L^2 \mid \int p = 0 \,\}.$$

For the Stokes problem, the coercivity in the kernel condition is trivial, because the $a$ form is coercive over all of $\mathring{H}^1(\Omega;\mathbb{R}^2)$. This accounts for the fact that this condition is less well-known than the second Brezzi condition. For the Stokes equations it is automatic, also on the discrete level.

For the second condition we need to prove that div maps $\mathring{H}^1$ onto $\hat{L}^2$. This result, usually attributed to Ladyzhenskaya, is somewhat technical due to the boundary conditions , and we do not give the proof.

## 9. Stability of mixed Galerkin methods

Now suppose we apply a Galerkin method to our abstract saddle point problem. That is, we choose finite dimensional subspaces $V_h \subset V$ and $W_h \subset W$ and seek $\sigma_h \in V_h$, $u_h \in W_h$ such that

$$\begin{aligned} a(\sigma_h,\tau) + b(\tau,u_h) &= F(\tau), \quad \tau \in V_h, \\ b(\sigma_h,v) &= G(v), \quad v \in W_h. \end{aligned} \tag{8.12}$$

We may apply Brezzi's theorem to this problem. Suppose that

$$a(z,z) \ge \gamma_{1,h}\|z\|_V^2, \quad z \in Z_h := \{\, \tau \in V_h \mid b(\tau,v) = 0,\ v \in W_h \,\}, \tag{8.13}$$

and

$$\inf_{0\ne v\in W_h} \sup_{0\ne\tau\in V_h} \frac{b(\tau,v)}{\|\tau\|\|v\|} \ge \gamma_{2,h}. \tag{8.14}$$

for some positive constants $\gamma_{1,h}$, $\gamma_{2,h}$. Then the discrete problem admits a unique solution and we have the stability estimate

$$\|\sigma_h\|_V + \|u_h\|_W \le c(\|F|_{V_h}\|_{V_h^*} + \|G|_{W_h}\|_{W_h^*}),$$

where $c$ depends only on $\gamma_{1,h}$, $\gamma_{2,h}$ and $\|a\|$. The general theory of Galerkin methods then immediately gives a quasioptimality estimate.

THEOREM 8.9. *Suppose that $(\sigma, u) \in V \times W$ satisfy the abstract saddle point problem* (8.6) *Let $V_h \subset V$ and $W_h \subset W$ be finite dimensional subspaces and suppose that the Brezzi conditions* (8.13) *and* (8.14) *hold for some $\gamma_{1,h}, \gamma_{2,h} > 0$. Then the discrete problem* (8.12) *has a unique solution $(\sigma_h, u_h) \in V_h \times W_h$ and*

$$\|\sigma - \sigma_h\|_V + \|u - u_h\|_W \le c(\inf_{\tau \in V_h} \|\sigma - \tau\|_V + \inf_{v \in W_h} \|u - v\|_W),$$

*where the constant $c$ depends only on $\gamma_{1,h}$, $\gamma_{2,h}$ and the norms of $a$ and $b$.*

This estimate is the fundamental estimate for mixed methods. In many cases it is too crude, since it couples the approximation of $\sigma$ and $u$, and often other useful estimates can be derived using a duality argument. We will see these in specific cases.

The major message from this theorem, however, is that, unlike for coercive formulations, for saddle point problems the Galerkin subspaces $V_h$ and $W_h$ have to be chosen with a view not only to approximation, but also stability, specifically, so that (8.13) and (8.14) hold.

## 10. Mixed finite elements for the Poisson equation

**10.1. Mixed finite elements in 1D.** As a simple example, let us return to the one-dimensional example shown in Figure 8.3. Here

$$a(\sigma, \tau) = \int_{-1}^{1} \sigma\tau \, dx, \quad b(\tau, v) = \int_{-1}^{1} \tau' v \, dx.$$

If we choose both $V_h$ and $W_h$ to be the space of continuous piecewise linears for some mesh, then $\gamma_{2,h} = 0$, because for $v$ a nonzero continuous piecewise linear which vanishes at each element midpoint, $\int \tau' v \, dx = 0$ for all $\tau \in V_h$. Thus this choice of elements violates the second Brezzi condition in the worst possible way, $\gamma_{2,h} = 0$, and does not even give a nonsingular discrete problem. One might consider removing this highly oscillatory function from $W_h$, e.g., by replacing $W_h$ by its orthogonal complement, but in that case it turns our $\gamma_{2,h} \to 0$ with $h$.

Next we make the choice shown in green in Figure 8.3, namely $V_h$ continuous piecewise linear, $W_h$ piecewise constant. Turning to the first Brezzi condition, $Z_h$ is the space of continuous piecewise linears with derivative orthogonal to piecewise constants, which means with vanishing derivative, i.e., $Z_h$ consists only of the constant functions. Clearly $a(\tau, \tau) = \int \tau^2 \, dx$ coerces (actually equals) the $H^1$ norm for a constant. So the first condition holds with $\gamma_{1,h} = 1$. For the second condition, given piecewise constant $v$, we let $\tau(x) = \int_0^x v(t) \, dt$, which is a continuous piecewise linear. Note that $\|\tau\|_0 \le \|v\|_0$ and $\tau' = v$, so $\|\tau\|_1^2 \le 2\|v\|_0^2$. We have

$$b(\tau, v) = \|v\|_0^2 \ge \frac{1}{\sqrt{2}} \|\tau\| \|v\|.$$

which establishes the inf-sup condition with $\gamma_{2,h} = 1/\sqrt{2}$. This proves the stability of the method and justifies the good approximation quality we see in the figure.

Finally, consider the same choice for $W_h$ but the use of continuous piecewise quadratics for $V_h$, which is shown in red in Figure 8.3. Increasing the size of $V_h$ only increases the inf-sup constant, so the second condition is fulfilled. However it also increases the size of $Z_h$, and so makes the coercivity in the kernel condition more difficult. Specifically, let $[\bar{x}, \bar{x} + h]$ be any mesh interval of length $h$ and consider $\tau(x) = (x - \bar{x})(x - \bar{x} - h)$ on this interval, 0

everywhere else. Then $\tau \in Z_h$, $\|\tau\|_0^2 = O(h^5)$, $\|\tau\|_1^2 = O(h^3)$, and so $a(\tau,\tau)/\|\tau_1\|_1^2 = O(h^2)$. Therefore, $\gamma_{1,h} \to 0$ as $h \to 0$, explaining the instability we see.

**10.2. Mixed finite elements in 2D.** Now we return to mixed finite elements for Poisson's equation in two dimensions; see (8.5). What spaces $V_h \subset H(\mathrm{div})$ and $W_h \subset L^2$ can we choose for stable approximation? We saw by numerical example that the choice of continuous piecewise linear elements for $V_h$ and piecewise constants for $W_h$, while stable in one dimension, are not stable in two dimensions.

The first stable spaces for this problem were provided by Raviart and Thomas in 1975. We begin with the description of the simplest finite elements in the Raviart–Thomas family. For the space $W_h$ we do indeed take the space of piecewise constants (so the shape functions on any triangle are simply the constants, and for each triangle $T$ we take the single DOF $v \mapsto \int_T v\,dx$). For the space $V_h$ we take as shape functions on a triangle $T$

$$\mathcal{P}_1^-(T;\mathbb{R}^2) := \{\, \tau(x) = a + bx \,|\, a \in \mathbb{R}^2, b \in \mathbb{R}, x = (x_1, x_2) \,\}.$$

In other words, the shape function space is spanned by the constant vector fields $(1,0)$ and $(0,1)$ together with the vector field $x = (x_1, x_2)$. Note that $\mathcal{P}_1^-(T;\mathbb{R}^2)$ is a 3-dimensional subspace of the 6-dimensional space $\mathcal{P}_1(T;\mathbb{R}^2)$. For example, the function $\tau(x) = (1 + 2x_1, 3 + 2x_2)$ is a shape function, but $\tau(x) = (1, x_2)$ is not.

For DOFs, we assign one to each edge of the triangle, namely to the edge $e$ of $T$ we assign

$$\tau \mapsto \int_e \tau \cdot n_e\,ds,$$

where $n_e$ is one of the unit normals to $e$. Let us show that these DOFs are unisolvent. Let $\tau = a + bx$, $a \in \mathbb{R}^2$, $b \in \mathbb{R}$, and suppose all three DOFs vanish for $\tau$. Note that $\operatorname{div}\tau = 2b$. Therefore

$$2|T|b = \int_T \operatorname{div}\tau\,dx = \int_{\partial T} \tau \cdot n\,ds = 0.$$

Thus $b = 0$ and $\tau = a$ is a constant vector. But the DOFs imply that $\tau \cdot n_e$ vanish for each of the three edges. Any two of these are linearly independent, so $\tau$ vanishes.

For any triangulation $\mathcal{T}_h$ we have thus defined a finite element space $V_h$. It consists of all the vector fields $\tau : \Omega \to \mathbb{R}^2$ such that $\tau|_T \in \mathcal{P}_1^-(T;\mathbb{R}^2)$ for all $T \in \mathcal{T}_h$ and, if $e$ is a common edge of $T_-, T_+ \in \mathcal{T}_h$, and $n_e$ is one of the normals to $e$, then

$$\int_e \tau|_{T_-} \cdot n_e\,ds = \int_e \tau|_{T_+} \cdot n_e\,ds. \tag{8.15}$$

Our next goal is to show that $V_h \subset H(\mathrm{div})$. Just as a piecewise smooth function with respect to a triangulation belongs to $H^1$ if and only if it is continuous across each edge, we can show that a piecewise smooth vector field belongs to $H(\mathrm{div})$ if and only if the normal component is continuous across each edge. This basically follows from the computation

$$-\int_\Omega \tau \cdot \operatorname{grad}\phi\,dx = \sum_T \int_T \operatorname{div}\tau\phi\,dx - \sum_T \int_{\partial T} \tau \cdot n_T \phi\,ds.$$

for any piecewise smooth $\tau$ and $\phi \in \mathring{C}^\infty(\Omega)$. If $\tau$ has continuous normal components, then we have cancellation, so

$$\sum_T \int_{\partial T} \tau \cdot n_T \phi \, ds = 0,$$

which means that

$$-\int_\Omega \tau \cdot \operatorname{grad} \phi \, dx = \int_\Omega \operatorname{div}_h \tau \phi \, dx,$$

where $\operatorname{div}_h \tau \in L^2(\Omega)$ is the piecewise divergence of $\tau$. This shows that the weak divergence of $\tau$ exists and belongs to $L^2$.

Now, by (8.15) we have for the Raviart–Thomas space $W_h$ that the jump of the normal component $\tau|_{T_-} \cdot n_e - \tau|_{T_+} \cdot n_e$ vanishes *on average* on $e$. However, for $\tau$ to belong to $H(\operatorname{div})$ we need this jump to vanish identically. This depends on a property of the space $\mathcal{P}_1^-(T;\mathbb{R}^2)$.

LEMMA 8.10. *Let $\tau \in \mathcal{P}_1^-(T;\mathbb{R}^2)$ and let $e$ be an edge of $T$. Then $\tau \cdot n_e$ is constant on $e$.*

PROOF. It is enough to consider the case $\tau(x) = x$ (since $\mathcal{P}_1^-$ is spanned by this $\tau$ and constants). Take any two points $x, y \in e$. Then $x - y$ is a vector tangent to $e$, so $(x-y)\cdot n_e = 0$, i.e., $\tau(x)\cdot n_e = \tau(y)\cdot n_e$. Thus $\tau \cdot n_e$ is indeed constant on $e$. □

We have thus defined the Raviart–Thomas space $V_h \subset H(\operatorname{div})$ and the space of piecewise constants $W_h \subset L^2$. Clearly we have $\operatorname{div} V_h \subset W_h$ (since the vector fields in $V_h$ are piecewise linear). From this we have that the discrete kernel

$$Z_h = \{\, \tau \in V_h \mid \int \operatorname{div} \tau v \, dx = 0 \ \forall v \in W_h \,\}$$

consists precisely of the divergence-free functions in $V_h$. From this the first Brezzi condition (coercivity over $Z_h$) holds (with constant 1).

The key point is prove the inf-sup condition. To this end we introduce the projection operator $\pi_h : H^1(\Omega;\mathbb{R}^2) \to V_h$ determined by the DOFs:

$$\int_e \pi_h \tau \cdot n_e \, ds = \int_e \tau \cdot n_e \, ds, \quad \tau \in H^1(\Omega;\mathbb{R}^2).$$

Note that we take the domain of $\pi_h$ as $H^1(\Omega;\mathbb{R}^2)$ rather than $H(\operatorname{div})$. The reason for this is that $\int_e \tau \cdot n_e \, ds$ need not be defined for $\tau \in H(\operatorname{div})$, but certainly is for $\tau \in H^1$, since then $\tau|_{\partial T} \in L^2(\partial T)$.

We also define $P_h : L^2(T) \to W_h$ by $\int_T P_h v \, dx = \int_T v \, dx$, i.e., the $L^2$ projection. Then we have the following very important result.

THEOREM 8.11. $\operatorname{div} \pi_h \tau = P_h \operatorname{div} \tau, \quad \tau \in H^1(\Omega;\mathbb{R}^2)$.

PROOF. The left hand side of the equation is a piecewise constant function, so it suffices to show that

$$\int_T \operatorname{div} \pi_h \tau \, dx = \int_T \operatorname{div} \tau \, dx.$$

But this is an easy consequence of Green's theorem:

$$\int_T \operatorname{div} \pi_h \tau \, dx = \int_{\partial T} \pi_h \tau \cdot n \, ds = \int_{\partial T} \tau \cdot n \, ds = \int_T \operatorname{div} \tau \, dx.$$

□

The theorem can be restated as the commutativity of the following diagram:

$$\begin{array}{ccc} H^1 & \xrightarrow{\operatorname{div}} & L^2 \\ \downarrow{\scriptstyle \pi_h} & & \downarrow{\scriptstyle P_h} \\ V_h & \xrightarrow{\operatorname{div}} & W_h. \end{array}$$

We shall also prove below that $\pi_h$ is bounded on $H^1$:

THEOREM 8.12. *There exists a constant independent of $h$ such that*

$$\|\pi_h \tau\|_{H(\operatorname{div})} \leq c\|\tau\|_1, \quad \tau \in H^1(\Omega; \mathbb{R}^2).$$

From these two results, together with the inf-sup condition on the continuous level, we get the inf-sup condition for the Raviart–Thomas spaces.

THEOREM 8.13. *There exists $\gamma > 0$ independent of $h$ such that*

$$\inf_{0 \neq v \in W_h} \sup_{0 \neq \tau \in W_h} \frac{\int \operatorname{div} \tau v \, dx}{\|\tau\|_{H(\operatorname{div})} \|v\|} \geq \gamma.$$

PROOF. It suffices to show that for any $v \in W_h$ we can find $\tau \in V_h$ with $\operatorname{div} \tau = v$ and $\|\tau\|_{H(\operatorname{div})} \leq c\|v\|$. First we find $\sigma \in H^1(\Omega; \mathbb{R}^2)$ with $\operatorname{div} \sigma = v$, $\|\sigma\|_1 \leq c\|v\|$. For example, we can extend $v$ by zero to a disc or other smooth domain and define $u \in H^2$ by $\Delta u = v$ with Dirichlet boundary conditions, and then put $\sigma = \operatorname{grad} u$. Finally, we let $\tau = \pi_h \sigma$. We then have

$$\operatorname{div} \tau = \operatorname{div} \pi_h \sigma = P_h \operatorname{div} \sigma = P_h v = v.$$

Moreover,

$$\|\tau\|_{H(\operatorname{div})} \leq c\|\sigma\|_1 \leq c\|v\|.$$

□

In view of Brezzi's theorem, we then get quasioptimality:

THEOREM 8.14. *If $(\sigma, u) \in H(\operatorname{div}) \times L^2$ solves the Poisson problem and $(\sigma_h, u_h) \in V_h \times W_h$ is the Galerkin solution using the Raviart–Thomas spaces, then*

$$\|\sigma - \sigma_h\|_{H(\operatorname{div})} + \|u - u_h\| \leq c(\inf_{\tau \in V_h} \|\sigma - \tau\|_{H(\operatorname{div})} + \inf_{v \in W_h} \|u - v\|).$$

For the second infimum, we of course have

$$\inf_{v \in W_h} \|u - v\| \leq ch\|u\|_1.$$

It remains to bound the first infimum, i.e., to investigate the approximation properties of the Raviart–Thomas space $V_h$.

We will approach this in the usual way. Namely, we will use the projection operator $\pi_h$ coming from the DOFs to provide approximation, and we will investigate this using Bramble–Hilbert and scaling. We face the same difficulty we did when we analyzed the Hermite quintic interpolant: $\pi_h$ is not invariant under affine scaling, because it depends on the normals to the triangle. Therefore, just as for the Hermite quintic, we shall only use scaling by dilation, together with a compactness argument.

For any triangle $T$, set $\pi_T : H^1(T;\mathbb{R}^2) \to \mathcal{P}_1^-(T;\mathbb{R}^2)$ denote the interpolant given by the Raviart–Thomas degrees of freedom. Since the constant vector fields belong to $\mathcal{P}_1^-$, we get, by the Bramble–Hilbert lemma, that

$$\|\tau - \pi_T\tau\|_{L^2(T)} \le c_T|\tau|_{H^1(T)}.$$

As in the Hermite quintic case, we denote by $\mathcal{S}(\theta)$ the set of all triangles of diameter 1 with angles bounded below by $\theta > 0$. By compactness we get that the constant $c_T$ can be chosen independent of $T \in \mathcal{S}(\theta)$. Then we dilate an arbitrary triangle $T$ by $1/h_T$ to get a triangle of diameter 1, and find that

$$\|\tau - \pi_T\tau\|_{L^2(T)} \le ch_T|\tau|_{H^1(T)},$$

where $c$ depends only on the minimum angle condition. Adding over the triangles, we have

$$\|\tau - \pi_h\tau\|_{L^2(\Omega)} \le ch|\tau|_{H^1(\Omega)}, \quad \tau \in H^1(\Omega),$$

where $h$ is the maximum triangle size.

We also have, by Theorem 8.11, that

$$\|\operatorname{div}(\tau - \pi_h\tau)\|_{L^2(\Omega)} = \|\operatorname{div}\tau - P_h\operatorname{div}\tau\|_{L^2(\Omega)} \le ch\|\operatorname{div}\tau\|_1 \le ch\|\tau\|_2.$$

THEOREM 8.15.

$$\|\tau - \pi_h\tau\| \le ch\|\tau\|_1, \quad \tau \in H^1(\Omega;\mathbb{R}^2),$$
$$\|\operatorname{div}(\tau - \pi_h\tau)\| \le ch\|\operatorname{div}\tau\|_1, \quad \tau \in H^1, \operatorname{div}\tau \in H^1.$$

We immediately deduce Theorem 8.12 as well:

$$\|\pi_h\tau\| \le \|\tau\| + \|\pi_h\tau - \tau\| \le c\|\tau\|_1,$$
$$\|\operatorname{div}\pi_h\tau\| = \|P_h\operatorname{div}\tau\| \le \|\operatorname{div}\tau\| \le \|\tau\|_1.$$

Putting together Theorem 8.15 and Theorem 8.14 we get

$$\|\sigma - \sigma_h\|_{H(\operatorname{div})} + \|u - u_h\| \le ch(\|\sigma\|_1 + \|\operatorname{div}\sigma\|_1 + \|u\|_1).$$

10.2.1. *Improved estimates for $\sigma$.* This theorem gives first order convergence for $\sigma$ in $L^2$, $\operatorname{div}\sigma \in L^2$, and $u \in L^2$, which, for each, is optimal. However, by tying the variables together it requires more smoothness than is optimal. For example, it is not optimal that the $L^2$ estimate for $\sigma$ or $u$ depend on the $H^1$ norm of $\operatorname{div}\sigma$. Here we show how to obtain improved estimates for $\sigma$ and $\operatorname{div}\sigma$, and below we obtain an improved estimate for $u$.

We begin with the error equations

$$\int (\sigma - \sigma_h)\cdot\tau\,dx + \int \operatorname{div}\tau(u - u_h)\,dx = 0, \quad \tau \in V_h, \tag{8.16}$$

$$\int \operatorname{div}(\sigma - \sigma_h)v\,dx = 0, \quad v \in W_h. \tag{8.17}$$

Now, from the inclusion $\operatorname{div}\sigma_h \in W_h$, we obtain

$$P_h\operatorname{div}\sigma - \operatorname{div}\sigma_h = P_h\operatorname{div}(\sigma - \sigma_h).$$

But (8.17) implies $P_h\operatorname{div}(\sigma - \sigma_h) = 0$. Thus

$$\operatorname{div}\sigma_h = P_h\operatorname{div}\sigma,$$

and we have a truly optimal estimate for $\operatorname{div}\sigma$:

$$\|\operatorname{div}(\sigma-\sigma_h)\| = \inf_{v\in V_h}\|\operatorname{div}\sigma - v\| \le ch\|\operatorname{div}\sigma\|_1.$$

Next we use the commuting diagram property of Theorem 8.11 to see that $\operatorname{div}(\pi_h\sigma-\sigma_h)=0$, so if we take $\tau=\pi_h\sigma-\sigma_h\in V_h$ in the first equation, we get

$$\int(\sigma-\sigma_h)\cdot(\pi_h\sigma-\sigma_h)\,dx = 0,$$

that is, $\sigma-\sigma_h$ is $L^2$-orthogonal to $\pi_h\sigma-\sigma_h$. It follows that

$$\|\sigma-\sigma_h\| \le \|\sigma-\pi_h\sigma\|,$$

and so,

$$\|\sigma-\sigma_h\| \le ch\|\sigma\|_1.$$

This is an optimal $L^2$ estimate for $\sigma$.

We shall obtain an optimal $L^2$ estimate for $u$ below.

**10.3. Higher order mixed finite elements.** We have thus far discussed the lowest order Raviart–Thomas finite element space, which uses the 3-dimensional space $\mathcal{P}_1^-(T)$ for shape functions. We now consider the higher order Raviart–Thomas elements, with shape functions

$$\mathcal{P}_r^- = \{\, a+bx \,|\, a\in\mathcal{P}_{r-1}(T;\mathbb{R}^2),\ b\in\mathcal{H}_{r-1}(T)\,\}.$$

Here $\mathcal{H}_{r-1}(T)$ is the space of *homogeneous* polynomials of degree $r-1$. We could allow $b$ to vary in $\mathcal{P}_{r-1}(T)$ instead of $\mathcal{H}_{r-1}(T)$, and the result space would be the same. Note that

$$\dim\mathcal{P}_r^-(T) = \dim\mathcal{P}_{r-1}(T;\mathbb{R}^2) + \dim\mathcal{H}_{r-1}(T) = (r+1)r + r = (r+2)r.$$

Before giving the DOFs and proving unisolvence, we establish some useful facts about polynomials.

THEOREM 8.16. *Let $b\in\mathcal{H}_r(\mathbb{R}^2)$ and $x=(x_1,x_2)$. Then* $\operatorname{div}(bx) = (r+2)b$

PROOF. It suffices to check this for a monomial $x_1^\alpha x_2^\beta$ with $\alpha+\beta=r$. Then

$$\begin{aligned}\operatorname{div}(bx) &= \operatorname{div}(x_1^{\alpha+1}x_2^\beta, x_1^\alpha x_2^{\beta+1}) = \frac{\partial}{\partial x_1}x_1^{\alpha+1}x_2^\beta + \frac{\partial}{\partial x_2}x_1^\alpha x_2^{\beta+1}\\ &= (\alpha+1)x_1^\alpha x_2^\beta + (\beta+1)x_1^\alpha x_2^\beta = (r+2)b.\end{aligned}$$

□

COROLLARY 8.17. *The divergence map* div *maps $\mathcal{P}_r(\mathbb{R}^2;\mathbb{R}^2)$ onto $\mathcal{P}_{r-1}(\mathbb{R}^2)$. In fact, it maps $\mathcal{P}_r^-(\mathbb{R}^2;\mathbb{R}^2)$ onto $\mathcal{P}_{r-1}(\mathbb{R}^2)$.*

PROOF. Given $f\in\mathcal{P}_{r-1}(\mathbb{R}^2)$ we have $f=\sum_{i=0}^{r-1}b_i$, $b_i\in\mathcal{H}_i(\mathbb{R}^2)$. We have

$$\operatorname{div}(\sum(i+2)^{-1}b_i x) = \sum(i+2)^{-1}\operatorname{div}(b_i x) = \sum b_i = f,$$

and

$$\sum_{i=0}^{r-1}(i+2)^{-1}b_i x = \sum_{i=0}^{r-2}b_i x + b_{r-1}x \in \mathcal{P}_{r-1}(\mathbb{R}^2;\mathbb{R}^2) + x\mathcal{H}_{r-1}(\mathbb{R}^2) = \mathcal{P}_r^-(\mathbb{R}^2;\mathbb{R}^2).$$

□

For a 2-vector $a = (a_1, a_2)$, we write $a^\perp = (-a_2, a_1)$ (rotation by $-\pi/2$). If $b$ is a function, we write $\operatorname{grad}^\perp b = (\operatorname{grad} b)^\perp = (-\partial b/\partial x_1, \partial b/\partial x_2)$.

THEOREM 8.18 (Polynomial de Rham sequence). *For any $r \geq 1$, the complex of maps*

$$\mathcal{P}_r(\mathbb{R}^2) \xrightarrow{\operatorname{grad}^\perp} \mathcal{P}_{r-1}(\mathbb{R}^2;\mathbb{R}^2) \xrightarrow{\operatorname{div}} \mathcal{P}_{r-2}(\mathbb{R}^2) \to 0$$

*is a resolution of the constants. In other words, the augmented complex*

$$0 \to \mathbb{R} \xrightarrow{\subset} \mathcal{P}_r(\mathbb{R}^2) \xrightarrow{\operatorname{grad}^\perp} \mathcal{P}_{r-1}(\mathbb{R}^2;\mathbb{R}^2) \xrightarrow{\operatorname{div}} \mathcal{P}_{r-2}(\mathbb{R}^2) \to 0.$$

*is exact. (For $r = 1$ we interpret $\mathcal{P}_{-1}(\mathbb{R}^2)$ as zero.*

The statement that the sequence of maps is a *complex* means that the composition of any two consecutive maps is zero, i.e., that the range of each map is contained in the kernel of the next map. In this case that means that $\operatorname{grad}^\perp$ kills the constant functions (which is obvious), and that $\operatorname{div} \circ \operatorname{grad}^\perp = 0$, which is easy to check. The statement that the complex is *exact* means that the range of each map precisely coincides with the kernel of the next map.

PROOF. Clearly the null space of the inclusion is zero, and the null space of $\operatorname{grad}^\perp$ is the space of constants. We have shown that the range of div is all of $\mathcal{P}_{r-2}$. So the only thing to be proven is that

$$\mathcal{R}(\operatorname{grad}^\perp) := \{\, \operatorname{grad}^\perp v \,|\, v \in \mathcal{P}_r(\mathbb{R}^2) \,\} = \mathcal{N}(\operatorname{div}) := \{\, \tau \in \mathcal{P}_{r-1}(\mathbb{R}^2;\mathbb{R}^2) \,|\, \operatorname{div}\tau = 0 \,\}.$$

We note that the first space is contained in the second, so it suffices to show that their dimensions are equal. For any linear map $L : V \to W$ between vector spaces, $\dim \mathcal{N}(L) + \dim \mathcal{R}(L) = \dim V$. Thus

$$\dim \mathcal{R}(\operatorname{grad}^\perp) = \dim \mathcal{P}_r(\mathbb{R}^2) - 1 = \frac{(r+1)(r+2)}{2} - 1 = \frac{(r+3)r}{2},$$
$$\dim \mathcal{N}(\operatorname{div}) = \dim \mathcal{P}_{r-1}(\mathbb{R}^2;\mathbb{R}^2) - \dim \mathcal{P}_{r-2}(\mathbb{R}^2) = r(r+1) - \frac{r(r-1)}{2} = \frac{(r+3)r}{2}.$$

□

By a very similar argument we get an exact sequence involving $\mathcal{P}_r^-$.

THEOREM 8.19. *For any $r \geq 1$, the complex of maps*

$$\mathcal{P}_r(\mathbb{R}^2) \xrightarrow{\operatorname{grad}^\perp} \mathcal{P}_r^-(\mathbb{R}^2;\mathbb{R}^2) \xrightarrow{\operatorname{div}} \mathcal{P}_{r-1}(\mathbb{R}^2) \to 0.$$

*is a resolution of the constants.*

We now give the degrees of freedom of the $\mathcal{P}_r^-$ finite element. These are

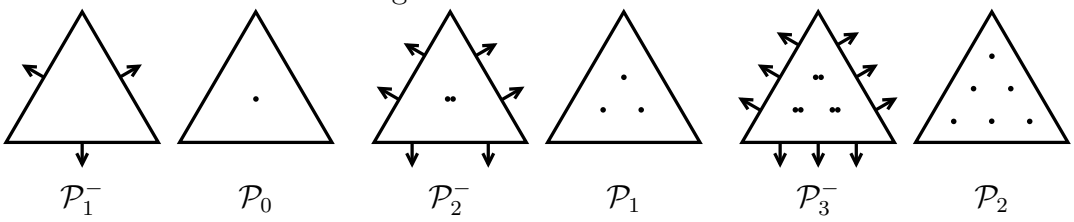
$$\tau \mapsto \int_e \tau \cdot n_e p(s)\, ds, \quad p \in \mathcal{P}_{r-1}(e), \tag{8.18}$$

and

$$\tau \mapsto \int_T \tau \cdot p(x)\, dx, \quad p \in \mathcal{P}_{r-2}(T;\mathbb{R}^2). \tag{8.19}$$

Note: strictly speaking what we have defined is the span of the DOFs on each edge and on $T$. By taking any basis of $\mathcal{P}_{r-1}(e)$ and for $\mathcal{P}_{r-2}(T)$ we get the DOFs. See Figure 8.4.

FIGURE 8.4. Higher order Raviart–Thomas elements.




THEOREM 8.20. *The DOFs given by* (8.18) *and* (8.19) *are unisolvent for* $\mathcal{P}_r^-(T;\mathbb{R}^2)$.

PROOF. First, we count the number of DOFs. There are $r$ per edge and $2 \times r(r-1)/2$ on the triangle, so $3r + r(r-1) = r(r+2) = \dim \mathcal{P}_{r-1}^1(T;\mathbb{R}^2)$ altogether. So, to show unisolvence, all we need to do is show that if all the DOFs vanish, then $\tau \in \mathcal{P}_{r-1}^1(T;\mathbb{R}^2)$ vanishes.

Now we know that $x \cdot n_e$ is constant on $n_e$, so this implies that for $\tau \in \mathcal{P}_r^-(T;\mathbb{R}^2)$, $\tau \cdot n_e \in \mathcal{P}_{r-1}(e)$. Therefore the DOFs in (8.18) imply that $\tau \cdot n$ vanishes on $\partial T$. We may then use integration by parts to find that

$$\int_T |\operatorname{div} \tau|^2 \, dx = -\int_T \tau \cdot \operatorname{grad} \operatorname{div} \tau \, dx = 0,$$

with the last equality coming from (8.19). Thus $\operatorname{div} \tau = 0$. Writing $\tau = a + bx$, $a \in \mathcal{P}_{r-1}(T;\mathbb{R}^2)$, $b \in \mathcal{H}_{r-1}(T)$ we conclude from Theorem 8.16 that $b = 0$, so $\tau \in \mathcal{P}_{r-1}(T;\mathbb{R}^2)$ and $\operatorname{div} \tau = 0$. The polynomial de Rham sequence Theorem (8.18) then tells us that $\tau = \operatorname{grad}^\perp \phi$, where $\phi \in \mathcal{P}_r(T)$ is determined up to addition of a constant. The condition $\tau \cdot n = 0$ means that $\partial\phi/\partial s = 0$ on each edge, so $\phi$ is equal to some constant on the boundary, which we can take equal to 0. Therefore $\phi = b\psi$, with $b \in \mathcal{P}_3(T)$ the bubble function and $\psi \in \mathcal{P}_{r-3}(T)$. Using the polynomial de Rham sequence again, we can write $\psi = \operatorname{div} \sigma$ with $\sigma \in \mathcal{P}_{r-2}(T;\mathbb{R}^2)$. Then

$$\begin{aligned} \int_T b\psi^2 \, dx = \int_T b\psi \operatorname{div} \sigma \, dx &= -\int_T \operatorname{grad}(b\psi) \cdot \sigma \, dx \\ &= -\int_T \operatorname{grad}^\perp(b\psi) \cdot \sigma^\perp \, dx = -\int_T \tau \cdot \sigma^\perp \, dx = 0, \end{aligned} \tag{8.20}$$

since $\sigma^\perp \in \mathcal{P}_{r-2}(T;\mathbb{R}^2)$. Thus $\psi = 0$ so $\tau = 0$ as claimed. □

Just as in the lowest order case, $r = 1$, considered previously, the choice of DOFs for the higher order Raviart–Thomas spaces are designed to make the proof of stability straightforward. First of all, they ensure that $\tau \cdot n_e$ is continuous across each edge $e$, so the assembled space is a subspace of $H(\operatorname{div})$. Let us denote the assembled $\mathcal{P}_r^-$ space by $V_h$ and denote by $W_h$ the space of all (not necessarily continuous) piecewise polynomials of degree $r-1$. We have $\operatorname{div} V_h \subset W_h$, so the first Brezzi condition is automatic. Again let $\pi_h : H^1(\Omega;\mathbb{R}^2) \to V_h$ be the projection determined by the DOFs, and let $P_h : L^2(\Omega) \to W_h$ be the $L^2$ projection.

Then the diagram

$$\begin{array}{ccc} H^1 & \xrightarrow{\operatorname{div}} & L^2 \\ \downarrow{\scriptstyle \pi_h} & & \downarrow{\scriptstyle P_h} \\ V_h & \xrightarrow{\operatorname{div}} & W_h. \end{array}$$

commutes, as follows directly from integration by parts and the DOFs. The inf-sup condition follows from this, just as in the lowest order case, and the quasioptimality estimate of Theorem 8.14 holds for all $r \geq 1$. Assuming a smooth solution, we thus get

$$\|\sigma - \sigma_h\|_{H(\operatorname{div})} + \|u - u_h\| = O(h^r).$$

The improved estimates for $\sigma$ and $\operatorname{div}\sigma$ carry through as well (since they only used the inclusion $\operatorname{div} V_h \subset V_h$ and the commuting diagram). Thus

$$\|\sigma - \sigma_h\| \leq ch^r \|\sigma\|_r, \quad \|\operatorname{div}(\sigma - \sigma_h)\| \leq ch^r \|\operatorname{div}\sigma\|_r.$$

We now use a duality argument to prove an improved estimate for $u$. As we have seen before, when using duality, we need 2-regularity of the Dirichlet problem, and hence we require that $\Omega$ be convex.

First we recall the error equations

$$\int (\sigma - \sigma_h) \cdot \tau \, dx + \int \operatorname{div}\tau (P_h u - u_h) \, dx = 0, \quad \tau \in V_h, \tag{8.21}$$

$$\int \operatorname{div}(\sigma - \sigma_h) v \, dx = 0, \quad v \in W_h. \tag{8.22}$$

Note that we have replaced $u$ with $P_h u$ in the first equation, which we can do, since $\operatorname{div}\tau \in W_h$ for $\tau \in V_h$. Now we follow Douglas and Roberts in defining $w$ as the solution of the Dirichlet problem

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

and set $\rho = \operatorname{grad} w$. By elliptic regularity, we have $\|w\|_2 + \|\rho\|_1 \leq c\|P_h u - u_h\|$.

Then

$$\begin{aligned} \|P_h u - u_h\|^2 &= (\operatorname{div}\rho, P_h u - u_h) = (\operatorname{div}\pi_h\rho, P_h u - u_h) = -(\sigma - \sigma_h, \pi_h\rho) \\ &= (\sigma - \sigma_h, \rho - \pi_h\rho) - (\sigma - \sigma_h, \rho) \\ &= (\sigma - \sigma_h, \rho - \pi_h\rho) + (\operatorname{div}(\sigma - \sigma_h), w) \\ &= (\sigma - \sigma_h, \rho - \pi_h\rho) + (\operatorname{div}(\sigma - \sigma_h), w - P_h w). \end{aligned}$$

This gives

$$\|P_h u - u_h\| \leq C(h\|\sigma - \sigma_h\| + h^2 \|\operatorname{div}(\sigma - \sigma_h)\|),$$

if $r > 1$, but for the lowest order elements, $r = 1$, it only gives

$$\|P_h u - u_h\| \leq Ch(\|\sigma - \sigma_h\| + \|\operatorname{div}(\sigma - \sigma_h)\|).$$

From this we easily get in the case $r > 1$ that

$$\|P_h u - u_h\| \leq Ch^{r+1}\|\sigma\|_r \leq Ch^{r+1}\|u\|_{r+1}$$

(so $u_h$ and $P_h$ are "super close", closer than either to $u$). For the case $r = 1$ we get

$$\|P_h u - u_h\| \leq Ch^2\|\sigma\|_1 + h\|\operatorname{div}(\sigma - \sigma_h)\| \leq Ch\|\sigma\|_1 \leq Ch\|u\|_2.$$

Using the triangle inequality to combine these with estimates for $\|u - P_h u\|$ we get these improved estimates for $u$:

$$\|u - u_h\| \le \begin{cases} Ch^r \|u\|_r, & r > 1, \\ Ch\|u\|_2, & r = 1. \end{cases}$$

Finally, we close this section by mentioning that the whole theory easily adapts to a second family of mixed elements, the BDM (Brezzi–Douglas–Marini) elements. Here the shape functions for $V_h$ are $\mathcal{P}_r(T;\mathbb{R}^2)$, $r \ge 1$, and the DOFs are

$$\tau \mapsto \int_e \tau \cdot n_e p(s)\, ds, \quad p \in \mathcal{P}_r(e),$$

and, if $r > 1$,

$$\tau \mapsto \int_T \tau \cdot p(x)^\perp \, dx, \quad p \in \mathcal{P}^-_{r-1}(T;\mathbb{R}^2).$$

## 11. Mixed finite elements for the Stokes equation

We return now to the Stokes equation, given in weak form: Find $u \in \mathring{H}^1(\Omega;\mathbb{R}^2)$, $\quad p \in \hat{L}^2(\Omega)$, such that

$$\int \operatorname{grad} u : \operatorname{grad} v \, dx - \int \operatorname{div} v \, p \, dx = \int f v \, dx, \quad v \in \mathring{H}^1(\Omega;\mathbb{R}^2),$$
$$\int \operatorname{div} u \, q \, dx = 0, \quad q \in \hat{L}^2.$$

Recall that $\hat{L}^2(\Omega)$ consists of the functions in $L^2$ with integral 0, and that we know that $\operatorname{div} \mathring{H}^1(\Omega;\mathbb{R}^2) = \hat{L}^2(\Omega)$, and so, for any $q \in \hat{L}^2$ there exists $v \in \mathring{H}^1(\Omega;\mathbb{R}^2)$ with $\operatorname{div} v = q$ and $\|v\|_1 \le c\|q\|$. This is equivalent to the inf-sup condition on the continuous level:

$$\inf_{0 \ne q \in \hat{L}^2} \sup_{0 \ne v \in \mathring{H}^1} \frac{\int \operatorname{div} v \, p \, dx}{\|v\|_1 \|p\|} \ge \gamma > 0.$$

Our goal is now to find stable finite element subspaces for Galerkin's method. Compared to the mixed Laplacian we see some differences.

- Because the bilinear form $a(u,v) = \int \operatorname{grad} u : \operatorname{grad} v \, dx$ is coercive over $\mathring{H}^1$, we do not have to worry about the first Brezzi condition. It holds for any choices of subspace.
- Since we need $V_h \subset H^1$ rather than $V_h \subset H(\operatorname{div})$, the finite elements we used for the mixed Laplacian do not apply. We need finite elements which are continuous across edges, not just with continuous normal component.
- The bilinear form $b(u,q) = \int \operatorname{div} uq \, dx$ is the same as for the mixed Laplacian, but the fact that we need the inf-sup condition with the $H^1$ norm rather than the $H(\operatorname{div})$ norm makes it more difficult to achieve.

We can rule out one simple choice of element which is vector-valued Lagrange $\mathcal{P}_1$ subject to the Dirichlet boundary conditions for $u$ and scalar Lagrange $\mathcal{P}_1$ elements subject to the mean value zero condition for $p$. We already saw that on a simple mesh there are nonzero

piecewise linears which are of mean value zero for which $\int \operatorname{div} v\, q\, dx = 0$ for all piecewise linear vector fields $v$.

We can rule out as well what may be regarded as the most obvious choice of elements, vector-valued Lagrange $\mathcal{P}_1$ for $u$ and piecewise constants for $p$. This method does not satisfy the inf-sup condition, as we saw in the case of the mixed Laplacian (for which the inf-sup condition is weaker).

However, we shall see that both these methods can be salvaged by keeping the same pressure space $W_h$ and enriching the velocity space $V_h$ appropriately.

**11.1. The $\mathcal{P}_2$-$\mathcal{P}_0$ element.** One of the simplest and most natural ways to prove the inf-sup condition is to construct a *Fortin operator*, by which we mean a linear operator $\pi_h : \mathring{H}^1(\Omega;\mathbb{R}^2) \to V_h$ satisfying

$$b(\pi_h v, q) = b(v, q), \quad q \in W_h, \tag{8.23}$$

and also the norm bound $\|\pi_h v\|_1 \le c\|v\|_1$. If we can find a Fortin operator, then we can deduce the inf-sup condition for $V_h \times W_h$ from the continuous inf-sup condition. Namely, given $q \in W_h$, we use the continuous inf-sup condition to find $v \in \mathring{H}^1$ with $\operatorname{div} v = q$, $\|v\|_1 \le \gamma^{-1}\|q\|$ for some $\gamma > 0$, so $b(v, q) = \|q\|^2 \ge \gamma\|v\|_1\|q\|$. We then get

$$b(\pi_h v, q) = b(v, q) \ge \gamma\|v\|_1\|q\| \ge \gamma c^{-1}\|\pi_h v\|_1\|q\|,$$

which is the inf-sup condition at the discrete level.

Now suppose we want to create a stable pair of spaces with $W_h$ the space of piecewise constants. What choice should we make for $V_h$ so that we can construct a Fortin operator and prove the inf-sup condition? In the case of $W_h$ equal piecewise constants, the condition (8.23) comes down to

$$\int_T \operatorname{div} \pi_h v\, dx = \int_T \operatorname{div} v\, dx,$$

for each triangle $T$, or, equivalently,

$$\int_{\partial T} \pi_h v \cdot n\, ds = \int_{\partial T} v \cdot n\, ds.$$

Therefore a sufficient condition is that

$$\int_e \pi_h v \cdot n_e\, ds = \int_e v \cdot n_e\, ds \tag{8.24}$$

for all edges $e$ of the mesh. This suggests that use for $V_h$ a finite element that includes the integrals of the edge normals among the degrees of freedom. In particular, we need at least one DOF per edge. A simple choice for this is the $\mathcal{P}_2$ Lagrange space, which has two DOFs per edge, which can be taken to be the integral of the two components along the edge (and so comprise the integral of the normal component). The other DOFs are the vertex values. This choice, Lagrange $\mathcal{P}_2$ for velocity and $\mathcal{P}_0$ for pressure, was suggested in Fortin's 1972 thesis, and analyzed by Crouzeix and Raviart in 1973. Given $v : \Omega \to \mathbb{R}^2$, we might define $\pi_h v$ triangle-by-triangle, by

$$\pi_h v(x) = v(x) \text{ for all vertices } x, \quad \int_e \pi_h v\, ds = \int_e v\, ds \text{ for all edges } e.$$

These imply (8.24) and so (8.23). However, this operator is not bounded on $H^1$, because it involves vertex values. It can, however, be fixed using a Clément interpolant. Recall that the Clément interpolant $\Pi_h : \mathring{H}^1 \to V_h$ satisfies

$$\|v - \Pi_h v\| \le Ch\|v\|_1, \quad \|\Pi_h v\|_1 \le c\|v\|_1,$$

(among other estimates). Next we define a second map $\tilde{\pi}_h : \mathring{H}^1 \to V_h$ by

$$\tilde{\pi}_T v = 0 \text{ at the vertices of } T, \quad \int_e \pi_T v \, ds = \int_e v \, ds \text{ for all edges } e.$$

Note that $\tilde{\pi}_h$ can be defined triangle by triangle: $(\tilde{\pi}_h v)|_T = \tilde{\pi}_T v|_T$. The map $\tilde{\pi}_T$ is defined on $H^1(T)$, since it only involves integrals on edges of $v$, not the values of $v$ at vertices. Thus, if we consider the unit triangle $\hat{T}$, we have

$$\|\tilde{\pi}_{\hat{T}} \hat{v}\|_{L^2(\hat{T})} \le c\|\hat{v}\|_{H^1(\hat{T})}.$$

The map $\tilde{\pi}_{\hat{T}}$ does not preserve constants, so we cannot use Bramble–Hilbert to reduce to the seminorm on the right hand side. Therefore, when we do the usual scaling to an element $T$ of size $h$ (with a shape regularity constraint), we get, in addition to the usual term $h|v|_{H^1(T)}$ also a term $\|v\|_{L^2(T)}$. That is, scaling gives

$$\|\tilde{\pi}_T v\|_{L^2(T)} \le c(\|v\|_{L^2(T)} + h|v|_{H^1(T)}).$$

Scaling similarly gives us

$$|\tilde{\pi}_T v|_{H^1(T)} \le c(h^{-1}\|v\|_{L^2(T)} + |v|_{H^1(T)}).$$

So, altogether, we get

$$\|\tilde{\pi}_h v\|_1 \le c(h^{-1}\|v\| + |v|_1).$$

Now we are ready to define the Fortin operator $\pi_h$:

$$\pi_h v = \tilde{\pi}_h (I - \Pi_h) v + \Pi_h v.$$

First we check the Fortin property:

$$\int_e \pi_h v \, ds = \int_e (I - \Pi_h) v \, ds + \int_e \Pi_h v \, ds = \int_e v \, ds.$$

Next we check the boundedness. There is no trouble with the Clément interpolant $\Pi_h v$, so we need only bound

$$\|\pi_h (I - \Pi_h) v\|_1 \le ch^{-1}\|(I - \Pi_h)v\|_0 + c\|(I - \Pi_h)v\|_1 \le c\|v\|_1.$$

THEOREM 8.21. *The choice $V_h$ Lagrange $\mathcal{P}_2$, $W_h$ piecewise constant is stable for the Stokes equations.*

It follows immediately that the Galerkin solution satisfies

$$\|u - u_h\|_1 + \|p - p_h\| \le c(\inf_{v \in V_h} \|u - v\|_1 + \inf_{q \in W_h} \|p - q\|),$$

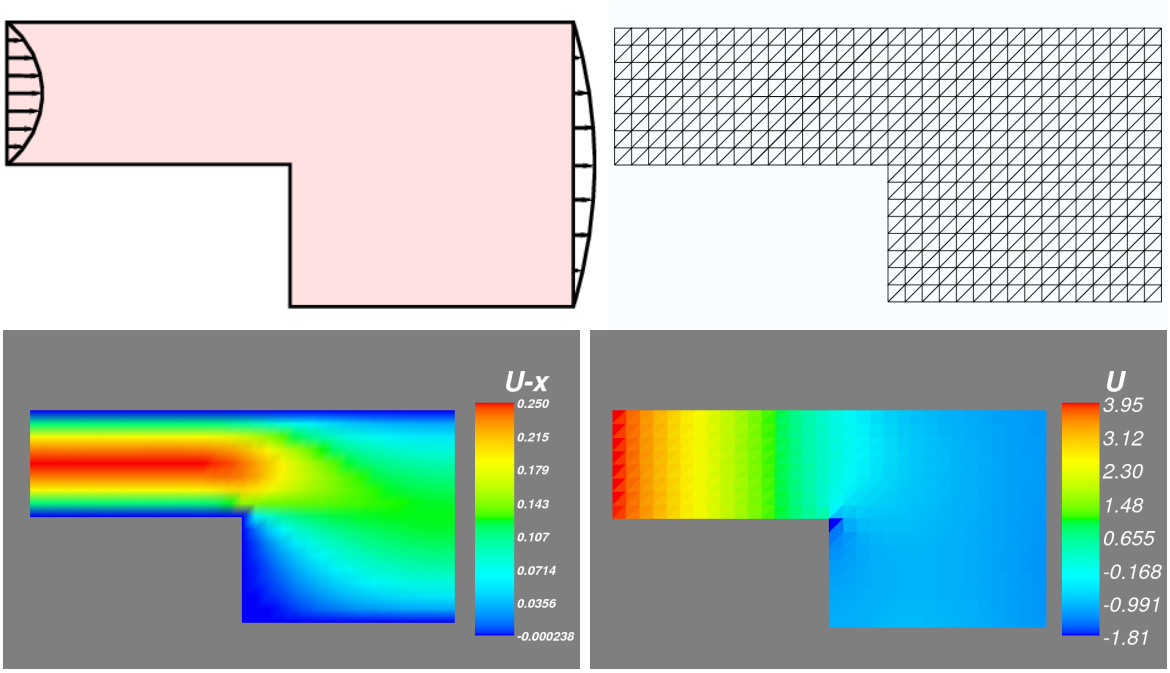
and so

$$\|u - u_h\|_1 + \|p - p_h\| \le ch(\|u\|_2 + \|p\|_1).$$

Notice that the rate of converge is only $O(h)$, the same as we would get for the best approximation using $\mathcal{P}_1$ Lagrange elements. The method in fact does not achieve $\|u - u_h\|_1 = O(h^2)$, because of the low order of pressure approximation.

We now illustrate the performance of the $\mathcal{P}_2$-$\mathcal{P}_0$ with a simple computation coded in FEniCS. The problem we solve is the homogeneous Stokes equations ($f = 0$) with inhomogeneous Dirichlet data for flow over a backward facing step. The problem is illustrated in the first subfigure of Figure 8.5, which shows the domain and the Dirichlet data. The inflow boundary on the left side runs from $x_2 = 0$ to $x_2 = 1$ and the input velocity is $u_1(x_2) = x_2 - x_2^2$, $u_2 = 0$, while at the outflow boundary, which runs from $x_2 = -1$ to $x_2 = 1$, the profile is $u_1 = (1 - x_2^2)/8$, $u_2 = 0$, a parabolic profile of twice the width but half the amplitude. The computational mesh, which has 768 elements, is shown in the second figure, and the computed solution for $u_1$ and $p$ in the final two figures. We note that the computation seems qualitatively reasonable, but artifacts of the discretization are clearly visible. Even though the mesh is quite fine, the accuracy is severely limited arising due to the low order elements (piecewise constants) for the pressure. The problem is greatest in a neighborhood of the reentrant corner where the true pressure has a singularity which the numerical solution is not able to capture at this resolution.

FIGURE 8.5. Flow over a step computed using $P_2$-$P_0$ elements. The quantities plotted are the horizontal component of velocity and the pressure.




**11.2. The mini element.** The mini element, introduced by Arnold, Brezzi, and Fortin in 1985, is the pair $P_1$+bubble for the velocity, and continuous $P_1$ for the pressure. It is the simplest stable element with continuous pressure space, just as the $\mathcal{P}_2$-$\mathcal{P}_0$ is the simplest stable Stokes element with discontinuous pressures. The velocity space, which I described as $\mathcal{P}_1$+bubble is defined as follows. First we define the scalar-valued $\mathcal{P}_1$+bubble $U_h$ with shape functions given by $\mathcal{P}_1(T) + \mathbb{R}b_T$ where $b_T$ is the cubic bubble function on $T$, i.e., the unique (up to nonzero constant multiple) cubic polynomial which vanishes on the boundary of the $T$ and is positive in the interior. It may be written as $\lambda_1\lambda_2\lambda_3$ where the $\lambda_i$ are the

barycentric coordinates of $T$. The DOFs for $U_h$ the vertex values and the integral $u \mapsto \int_T u$. It is easy to check unisolvence.

The mini element then takes $V_h = U_h \times U_h$, while $W_h$ is the usually Lagrange $\mathcal{P}_1$ space.

To prove stability, we again construct a Fortin operator $\pi_h : V \to V_h$, in a very similar manner to that we used for the $\mathcal{P}_2$-$\mathcal{P}_0$ element. To achieve the Fortin property

$$\int_\Omega \operatorname{div} \pi_h v \, q \, dx = \int_\Omega \operatorname{div} v \, q \, dx, \quad q \in W_h, \tag{8.25}$$

we use integration by parts. No boundary terms enter since $q \in H^1$ (thanks to the continuous pressure spaces) and $v$ and $\pi_h v$ vanish of $\partial\Omega$. Now $\operatorname{grad} q$ is a piecewise constant vector field, so it is sufficient that

$$\int_T \operatorname{div} \pi_h v \, dx = \int_T v \, dx.$$

We can accomplish this using the DOFs $v \mapsto \int_T v \, dx$ for the mini space $V_h$. Specifically, we define $\tilde{\pi}_T : L^2(T) \to \mathbb{R} b_T$ by

$$\int_T \tilde{\pi}_T v \, dx = \int_T \tilde{v} \, dx.$$

Notice $\tilde{\pi}_T$ is a bounded operator on $L^2(T)$ into a finite dimensional space. A simple scaling argument gives

$$\|\pi_T v\|_{H^1(T)} \le c h^{-1} \|v\|_{L^2(T)}.$$

We then define $\tilde{\pi}_h : V \to V_h$ by applying $\tilde{\pi}_T$ element-by-element, and define

$$\pi_h = \tilde{\pi}_h (I - \Pi_h) + \Pi_h,$$

where $\Pi_h$ is the Clément interpolant. Just as for the $\mathcal{P}_2$-$\mathcal{P}_0$ element, we easily verify the Fortin property (8.25) and uniform $H^1$ boundedness. Thus we have proven stability for the mini element. The estimate

$$\|u - u_h\|_1 + \|p - p_h\| \le c h (\|u\|_2 + \|p\|_1).$$

We can also use a straightforward Aubin–Nitsche duality argument to get

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

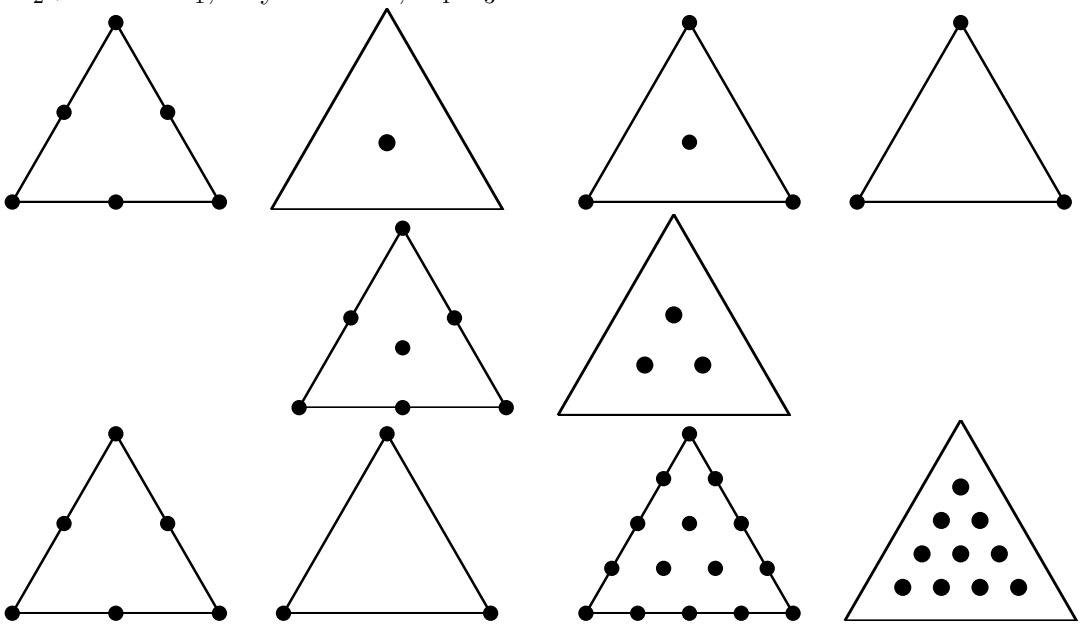
We do *not* get second order convergence for $p$ in $L^2$.

The mini element can be easily generalized to give higher order elements. For example we may use Lagrange $P_2$ elements for the pressure and $\mathcal{P}_2$+quartic bubbles for the velocity (the shape functions are $\mathcal{P}_2(T) + \mathcal{P}_1(T) b_T$. However, this is, in some sense, overkill. The same rates of convergence are achieved by choosing Lagrange $\mathcal{P}_2$ for velocity and Lagrange $\mathcal{P}_1$ for pressure. That simple, popular, element is called the Taylor–Hood element. It is stable, but the proof is far more sophisticated.

**11.3. Stable finite element for the Stokes equation.** We have shown stability for the simplest Stokes element with discontinuous pressures ($\mathcal{P}_2$-$\mathcal{P}_0$) and with continuous pressures (mini). A similar analysis, can be used to to prove the stability of the $\mathcal{P}_2$+bubble–$\mathcal{P}_1$ element (with discontinuous $\mathcal{P}_1$ pressure elements), which, like the $\mathcal{P}_2$-$\mathcal{P}_0$ element was published by Crouzeix and Raviart in their 1973 paper. A more complicated element family is the Taylor–Hood family in which the velocity field is approximated by continuous piecewise

polynomials of degree $r \geq 2$ and the pressure is approximated by continuous piecewise polynomials of degree $r-1$. This method is stable with a very weak restriction on the mesh: it must have at least 3 elements. Even more complicated is the $\mathcal{P}_r$-$\mathcal{P}_{r-1}$ element with discontinuous pressures. For smaller values of $r$ this method is not stable on most meshes. For $r \geq 4$, the method is stable with fairly minor restrictions on the mesh. Specifically, a vertex of the mesh (in the interior or on the boundary) is called *singular* if the edges containing it lie on just two lines. An interior vertex with four incoming edges or a boundary vertex with two or three incoming edges can be nearly singular as measured by the angles between the edges. In 1985 Scott and Vogelius proved that the $\mathcal{P}_r$-$\mathcal{P}_{r-1}$ discontinuous is stable on meshes with no singular or nearly singular vertices (i.e., the inf-sup condition deteriorates as a vertex tends towards singular).

FIGURE 8.6. Stable finite elements for the Stokes equations: $\mathcal{P}_2$-$\mathcal{P}_0$, mini, $\mathcal{P}_2$+bubble-$\mathcal{P}_1$, Taylor-Hood, $\mathcal{P}_4$-$\mathcal{P}_3$.



In 3D, the analogue of the $\mathcal{P}_2$-$\mathcal{P}_0$ element is the $\mathcal{P}_3$-$\mathcal{P}_0$ element, since $\mathcal{P}_3$ Lagrange element has a degree of freedom in each face of a tetrahedron. We may also generalize the $\mathcal{P}_2$+bubble-$\mathcal{P}_1$ element in 2D to $\mathcal{P}_3$+bubble-$\mathcal{P}_1$ in 3D (note that the bubble function has degree 4 in 3D. The mini element has a direct analogue in 3D: $\mathcal{P}_1$+bubble versus continuous $\mathcal{P}_1$. The Taylor–Hood family has also been shown to generalize to 3D (see Boffi 1997, or, for a proof using a Fortin operator, Falk 2008). As far as I know, the analogue of the Scott-Vogelius result in 3D is not understood (and would likely involve very high order elements).

CHAPTER 9

# Finite elements for elasticity

## 1. The boundary value problem of linear elasticity

The equations of elasticity model the deformation of a solid body under the action of imposed forces. Recall that the primary variables used to describe the state of the body are the displacement vector $u : \Omega \to \mathbb{R}^3$ and the stress tensor $\sigma : \Omega \to \mathbb{R}^{3\times 3}$. Here $\Omega \subset \mathbb{R}^3$ describes the body, typically in an undeformed configuration. The meaning of the displacement is that a point $x \in \Omega$ is displaced under the deformation to $x + u(x)$. The stress tensor measures the internal forces generated by the deformation. More precisely, if $S$ is a hypersurface embedded in the body, e.g, a small square embedded in a three-dimensional body, then the force across $S$, or traction, is given by $\int_S \sigma(x) n_S \, ds$. In other words, the traction vector $\sigma(x)n$ is the force per unit area at $x$ across a surface through $x$ with normal $n$. The fact that the traction vector has the form $\sigma n$ for a tensor (matrix) $\sigma$ is known as Cauchy's Theorem. The same theorem shows that, as a consequence of the conservation of angular momentum, the matrix $\sigma$ is symmetric.

The statement that the body is in equilibrium is

$$-\operatorname{div}\sigma = f \text{ in } \Omega, \tag{9.1}$$

where $f$ is the density of imposed forces.

To complete the system, we also need constitutive equations, which describe how internal stresses relate to the the deformation of the body. For an *elastic* material, the stress tensor $\sigma$ at a point depends only the gradient of the displacement at a point. In the linear theory of elasticity, the dependence is of the following form:

$$\sigma = C\,\epsilon(u), \tag{9.2}$$

where $\epsilon(u) = [\operatorname{grad} u + (\operatorname{grad} u)^T]/2$ is the symmetric part of the matrix $\operatorname{grad} u$, $C = C(x) : \mathbb{R}^{n\times n}_{\text{symm}} \to \mathbb{R}^{n\times n}_{\text{symm}}$ is a symmetric positive definite linear operator. (This means that $C\sigma : \tau = C\tau : \sigma$ for all $\sigma, \tau \in \mathbb{R}^{n\times n}_{\text{symm}}$ and there exists $\gamma > 0$ such that $C\tau : \tau \geq \gamma|\tau|^2$ for all $\tau \in \mathbb{R}^{n\times n}_{\text{symm}}$.) The *elasticity tensor* $C$ describes the elastic properties of the material. The material is called homogeneous if $C$ is independent of $x$. The material is called isotropic if its response is invariant under rotations. In this case the elasticity tensor can be written

$$C\tau = 2\mu\tau + \lambda \operatorname{tr}(\tau) I,$$

where $\mu > 0$ and $\lambda \geq 0$ are called the Lamé constants. Instead of the Lamé constants we can use the Young's modulus $E$ and Poisson ratio $\nu$:

$$C\tau = \frac{E}{1+\nu}\left[\tau + \frac{\nu}{1-2\nu}\operatorname{tr}(\tau) I\right]$$

Then $E > 0$ is like a spring constant for the material, the ratio of tensile stress to strain in the same direction (so it has units of stress). The Poisson ratio $\nu$ is dimensionless. It satisfies $0 \le \nu < 1/2$, with the limit $\nu \uparrow 1/2$, or equivalently $\lambda \to +\infty$ being the incompressible limit (nearly attained for some rubbers). For convenience we record the relations between the Lamé constants and the Young's modulus and Poisson ratio:

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E}{1+\nu}\frac{\nu}{1-2\nu}, \quad E = \mu\frac{3\lambda+2\mu}{\lambda+\mu}, \quad \nu = \frac{\lambda}{2(\lambda+\mu)}. \tag{9.3}$$

In order to obtain a well-posed problem, we need to combine the equilibrium equation (9.1) and constitutive equation (9.2) with boundary conditions. Let $\Gamma_D$ and $\Gamma_N$ be disjoint open subsets of $\partial\Omega$ whose closures cover $\partial\Omega$. We assume that $\Gamma_D$ is not empty (it may be all of $\partial\Omega$). On $\Gamma_D$ we impose the displacement

$$u = g \text{ on } \Gamma_D, \tag{9.4}$$

with $g : \Gamma_D \to \mathbb{R}^n$ given. On $\Gamma_N$ we impose the traction:

$$\sigma n = k \text{ on } \Gamma_N, \tag{9.5}$$

with $k : \Gamma_N \to \mathbb{R}^n$ given. The equations (9.2), (9.1), (9.4), and (9.5) constitute a complete boundary value problem for linear elasticity. In particular, we have pure Dirichlet problem

$$-\operatorname{div} C\,\epsilon(u) = f \text{ in } \Omega, \qquad u = g \text{ on } \partial\Omega.$$

We may eliminate the stress and write the elastic boundary value problem in terms of the displacement alone:

$$-\operatorname{div} C\,\epsilon(u) = f \text{ in } \Omega, \tag{9.6}$$

$$u = g \text{ on } \Gamma_D, \quad [C\,\epsilon(u)]n = k \text{ on } \Gamma_N. \tag{9.7}$$

Note that

$$\operatorname{div}\epsilon(u) = \frac{1}{2}\Delta u + \frac{1}{2}\operatorname{grad}\operatorname{div} u,$$

so, in the case of a homogeneous isotropic material, the differential equation can be written

$$-\mu\Delta u - (\mu+\lambda)\operatorname{grad}\operatorname{div} u = f.$$

## 2. The weak formulation

Our next goal is to derive a weak formulation. For this we will need to integrate by parts. By the divergence theorem (applied row-by-row), we have

$$\int_\Omega \operatorname{div}\tau \cdot v\,dx = -\int_\Omega \tau : \operatorname{grad} v\,dx + \int_{\partial\Omega} \tau n \cdot v\,ds$$

for any sufficiently smooth matrix field $\tau$ and vector field $v$. If $\tau$ is a *symmetric* matrix field, then $\tau : \operatorname{grad} v = \tau : \epsilon(v)$ (since $\operatorname{grad} v - \epsilon(v)$ is the skew-symmetric part of $\operatorname{grad} v$, and, at each point, $\tau$ is symmetric, and so orthogonal to all skew-symmetric matrices). Thus for symmetric $\tau$,

$$\int_\Omega \operatorname{div}\tau \cdot v\,dx = -\int_\Omega \tau : \epsilon(v)\,dx + \int_{\partial\Omega} \tau n \cdot v\,ds.$$

It is then straightforward to derive the weak formulation of the elastic boundary value problem (9.6). Let

$$H^1(\Omega;\mathbb{R}^n) = \{\, u = (u_1,\dots,u_n) \,|\, u_i \in H^1(\Omega)\,\},$$
$$H^1_{\Gamma_D,g} = \{\, u \in H^1(\Omega;\mathbb{R}^n) \,|\, u = g \text{ on } \Gamma_D\,\}, \quad H^1_{\Gamma_D} = \{\, u \in H^1(\Omega;\mathbb{R}^n) \,|\, u = 0 \text{ on } \Gamma_D\,\}.$$

The weak formulation seeks $u \in H^1_{\Gamma_D,g}$ such that

$$\int_\Omega C\,\epsilon(u) : \epsilon(v)\,dx = \int_\Omega f\cdot v\,dx + \int_{\Gamma_N} k\cdot v\,ds, \quad v \in H^1_{\Gamma_D}.$$

Defining

$$b : H^1(\Omega;\mathbb{R}^n)\times H^1(\Omega;\mathbb{R}^n) \to \mathbb{R}, \quad b(u,v) = \int C\,\epsilon(u) : \epsilon(v)\,dx,$$
$$F : H^1(\Omega;\mathbb{R}^n) \to \mathbb{R}, \quad F(v) = \int_\Omega f\cdot v\,dx + \int_{\Gamma_N} k\cdot v\,ds,$$

our problem takes the standard form: find $u \in H^1_{\Gamma_D,g}$ such that

$$b(u,v) = F(v), \quad v \in H^1_{\Gamma_D}.$$

As is common, we can reduce to the case where the Dirichlet data $g$ vanishes, by assuming that we can find a function $u_g \in H^1(\Omega;\mathbb{R}^n)$ such that $u_g = g$ on $\Gamma_D$. We can then write $u = u_g + \tilde{u}$ where $\tilde{u} \in H^1_{\Gamma_D}$ satisfies

$$b(\tilde{u},v) = \tilde{F}(v), \quad v \in H^1_{\Gamma_D}.$$

where $\tilde{F}(v) = F(v) - b(\tilde{u},v)$.

The bilinear form $b$ is clearly satisfies $b(v,v) \geq 0$. In fact, since we assumed that $C$ is positive definite on $\mathbb{R}^{n\times n}_{\text{symm}}$, we have

$$b(v,v) \geq \gamma \|\,\epsilon(v)\|^2, \quad v \in H^1(\Omega;\mathbb{R}^n).$$

We now show that the form $b$ is coercive based on *Korn's inequality*. We begin with a simple case, known as Korn's first inequality.

THEOREM 9.1. *Let $\Omega$ be a domain with Lipschitz boundary. Then there exists a constant $c$ such that*

$$\|v\|_1 \leq c\|\,\epsilon(v)\|, \quad u \in \mathring{H}^1(\Omega;\mathbb{R}^n).$$

PROOF.

$$\begin{aligned}
\|\,\epsilon(v)\|^2 &= \frac{1}{4}\int [\operatorname{grad} v + (\operatorname{grad} v)^T] : [\operatorname{grad} v + (\operatorname{grad} v)^T]\,dx \\
&= \frac{1}{4}\|\operatorname{grad} v\|^2 + \frac{1}{4}\|(\operatorname{grad} v)^T\|^2 + \frac{1}{2}\int \operatorname{grad} v : (\operatorname{grad} v)^T\,dx \qquad (9.8)\\
&= \frac{1}{2}\|\operatorname{grad} v\|^2 + \frac{1}{2}\int \operatorname{grad} v : (\operatorname{grad} v)^T\,dx.
\end{aligned}$$

Now if $v \in \mathring{H}^1 \cap H^2$ we can integrate by parts to find that

$$\int \operatorname{grad} v : (\operatorname{grad} v)^T\,dx = -\int v\cdot \operatorname{div}(\operatorname{grad} v)^2\,dx = -\int v\cdot \operatorname{grad}(\operatorname{div} v)\,dx = \int (\operatorname{div} v)^2\,dx,$$

i.e.,

$$\int \operatorname{grad} v : (\operatorname{grad} v)^T \, dx = \| \operatorname{div} v\|^2.$$

By density this holds for all $v \in \mathring{H}^1$, without requiring also $v \in H^2$. Combining with (9.8) gives

$$\| \epsilon(v)\|^2 \geq \frac{1}{2} \| \operatorname{grad} v\|^2, \quad v \in \mathring{H}^1.$$

The proof in completed by invoking Poincaré's inequality $\|v\|_1 \leq c\| \operatorname{grad} v\|$. □

Poincaré inequality holds not just for function in $\mathring{H}^1$, but also for functions which vanish on only an open subset of the boundary. The same is true for Korn's inequality (9.9), although the proof is considerably more difficult.

THEOREM 9.2. *Let $\Omega$ be a domain with a Lipschitz boundary and $\Gamma_D$ a nonempty open subset of $\partial\Omega$. Then there exists a constant $C$ such that*

$$\|v\|_1 \leq c\| \epsilon(v)\|, \quad v \in H^1_{\Gamma_D}(\Omega; \mathbb{R}^n). \tag{9.9}$$

Korn's inequality and the positivity of the elasticity tensor $C$ immediately give coercivity of the bilinear form $b$:

$$b(v, v) \geq \gamma \|v\|_1^2, \quad v \in H^1_{\Gamma_D}(\Omega; \mathbb{R}^n).$$

The well-posedness of the weak formulation of the elastic boundary value problem then follows using the Riesz representation theorem.

THEOREM 9.3. *Let $F : H^1_{\Gamma_D}(\Omega; \mathbb{R}^n) \to \mathbb{R}$ be a bounded linear functional. Then there exists a unique $u \in H^1_{\Gamma_D}(\Omega; \mathbb{R}^n)$ such that*

$$b(u, v) = F(v), \quad v \in H^1_{\Gamma_D}(\Omega; \mathbb{R}^n).$$

*Moreover there is a constant $C$ independent of $F$ such that*

$$\|u\|_1 \leq c\|F\|_{(H^1_{\Gamma_D})^*}.$$

## 3. Displacement finite element methods for elasticity

In view of the coercivity of $b$, we may choose any finite dimensional subspace $V_h \subset H^1_{\Gamma_D}$ and use Galerkin's method to find a unique $u_h \in V_h$ satisfies

$$b(u_h, v) = F(v), \quad v \in V_h.$$

Such a method is called a displacement method since the only quantity taken as an unknown is the displacement (in contrast to mixed methods which we will study below). The quasioptimal error estimate

$$\|u - u_h\|_1 \leq c \inf_{v \in V_h} \|u - v\|_1$$

holds with the constant $c$ depending only on the domain $\Omega$, Dirichlet boundary $\Gamma_D$, and the elasticity tensor $C$. The most common finite element space to use for $V_h$ are the vector Lagrange spaces, i.e., each component is taken to be a continuous piecewise polynomial of degree at most $r$ with respect to a given triangulation. Assuming mesh size $h$ and shape regularity we get the estimate

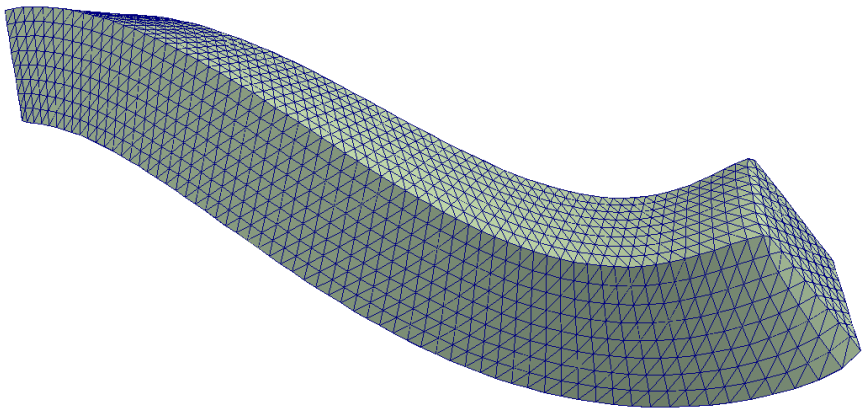
$$\|u - u_h\|_1 \leq ch^r \|u\|_{r+1}.$$

The Aubin-Nitsche duality argument allows us to improve this estimate to

$$\|u - u_h\| \leq ch^{r+1}\|u\|_{r+1}.$$

Next we show some computed examples. In the first example (see the file elas3d.py), we consider a cantilever bar with square cross-section. The domain $\Omega = (0,8) \times (0,1) \times (0,1)$. The left end $x_1 = 0$ is clamped: $u = 0$. On the right end $x_1 = 8$ we impose a displacement which is a rigid motion. On the four rectangular sides we use traction-free boundary conditions $\sigma n = 0$. This was coded in FEniCS using a $64 \times 8 \times 8$ mesh of cubes, each subdivided into 6 tetrahedra, with Lagrange elements of degree 2. See the file elas3d.py. Figure 9.1 shows the bar as deformed by the computed displacement. This is a good way to visualize a displacement vector field, although it should be noted that actual physical displacements for problems for which linear elasticity is a good model would be much smaller, e.g., by a factor of 10 or 100.

FIGURE 9.1. Displacement of elastic bar with left face clamped and a rigid displacement applied to the right face.

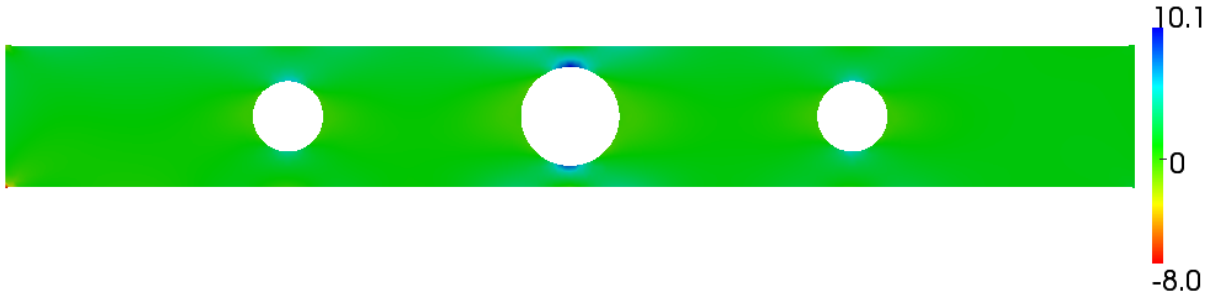


The second example is the analogous problem in two dimensions, except that the domain is the rectangle $(0,8) \times (0,1)$ with three circular cut-outs removed. Figure 9.2 show the stress component $\sigma_{11}$, which gives the tension in the $x_1$ direction (or the compression, if $\sigma_{11} < 0$). This is an important quantity for applications, since if the stress is too large at some point, the structure may fracture or otherwise fail there. Notice the high stress concentrations around the circular cut-outs. For the computations we took $E = 10$, $\nu = .2$, and used Lagrange elements of degree 2. See the program elas2d.py for the code.

## 4. Nearly incompressible elasticity and Poisson locking

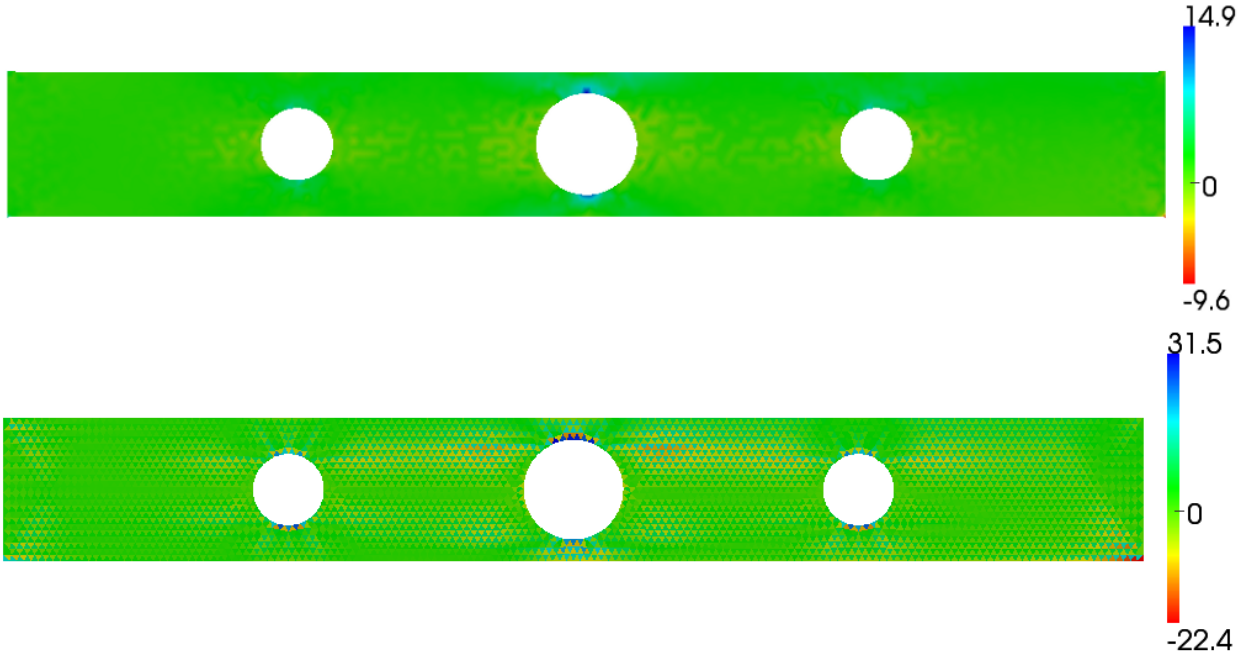
An isotropic elastic material is characterized by the two Lamé coefficients, $\mu > 0$ and $\lambda \geq 0$, or, equivalently, by Young's modulus $E$ and the Poisson ratio $\nu \in [0, 1/2)$. (The relation between these is given in (9.3). As the second Lamé coefficient $\lambda$ increases toward $+\infty$, or, equivalently, as the Poisson ratio $\nu$ increases toward $1/2$, the material becomes nearly incompressible. It turns out that standard displacement finite element methods have difficulty in solving such nearly incompressible problems. To see an example of this, consider the example just computed, with the stress shown in Figure 9.2, but now take the Poisson

FIGURE 9.2. Displacement of 2D elastic bar with cut-outs with left face clamped and a rigid displacement applied to the right face.




ratio equal to 0.499 rather than 0.2 as previously. This gives $\lambda \approx 1664$. The results are show in the first plot of Figure 9.3. Unphysical oscillations in the stress are clearly visible in the first plot, in contrast to the case of $\nu = 0.2$ show in Figure 9.2. Thus the standard displacement finite element method using Lagrange finite elements of degree 2 is not suitable for nearly incompressible materials. The situation is even worse for Lagrange elements of degree 1, show in the second plot of Figure 9.3.

FIGURE 9.3. For a nearly incompressible material, the stress shows unphysical oscillations for quadratic Lagrange elements (top) and, more pronouncedly, for linear Lagrange elements (bottom).




We know that the displacement method gives the error estimate

$$\|u - u_h\|_1 \leq Ch^r \|u\|_{r+1}. \tag{9.10}$$

So why do we not get good results in the nearly incompressible case? The problem is *not* that the exact solution $u$ degenerates. It can be shown that $\|\sigma\|_r$ and $\|u\|_{r+1}$ remain uniformly bounded as $\lambda \to \infty$ (for all values of $r$ if the domain is smooth). So the problem must be the constant $C$ entering the error estimate: it must blow up as $\lambda \to \infty$. In short the accuracy of the finite element method degenerates as $\lambda$ grows, even though the exact solution does not degenerate.

Let us investigate the dependence on$\lambda$ of the constant $C$ in the error bound (9.10). As always, the error is bounded by the stability constant times the consistency error. In this case, the bilinear form

$$b(u,v) = 2\mu \int \epsilon(u) : \epsilon(v)\, dx + \lambda \int (\operatorname{div} u)(\operatorname{div} v)\, dx,$$

so

$$b(v,v) \geq 2\mu \| \epsilon(u)\|^2 \geq \gamma \|u\|_1^2,$$

with the contant $\gamma > 0$ depending only on $\mu$ and the constant in Korn's inequality, but entirely independent of $\lambda$. That is, the bilinear form is coercive *uniformly in* $\lambda$, and so Galerkin's method is stable uniformly in $\lambda$. Thus the difficulties in treating the nearly incompressible cannot be attributed to a degeneration of stability, and we must look to the consistency error.

Recall that the consistency error is bounded by

$$\|b\| \inf_{v \in V_h} \|u - v\|_1$$

where $u$ is the exact solution, $V_h$ is the finite element space, and $\|b\|$ is the norm of the bilinear form (with respect to the $H^1$ norm of its arguments. The infimum is bounded by $ch^r\|u\|_{r+1}$ where $c$ depends on the shape constant of the mesh, but has nothing to do with $\lambda$. But finally we get to the culprit. Since the coefficient $\lambda$ enters the bilinear form $b$, $\|b\|$ tends to $\infty$ with $\lambda$.

## 5. Mixed finite elements for elasticity