MATH 8445–8446, University of Minnesota
Numerical Analysis of Differential Equations

Lecture notes on Numerical Analysis of Partial Differential Equations

– version prepared for 2017–2018 –
Last modified: March 22, 2018

Douglas N. Arnold

Contents

CHAPTER 1

Introduction

Galileo wrote that the great book of nature is written in the language of mathematics. The most precise and concise description of many physical systems is through partial differential equations.

1. Basic examples of PDEs

1.1. Heat flow and the heat equation. We start with a typical physical application of partial differential equations, the modeling of heat flow. Suppose we have a solid body occupying a region $\Omega \subset \mathbb{R}^3$. The temperature distribution in the body can be given by a function $u : \Omega \times J \to \mathbb{R}$ where J is an interval of time we are interested in and $u(x, t)$ is the temperature at a point $x \in \Omega$ at time $t \in J$. The heat content (the amount of thermal energy) in a subbody $D \subset \Omega$ is given by

$$\text{heat content of } D = \int_D cu\,dx$$

where c is the product of the specific heat of the material and the density of the material. Since the temperature may vary with time, so can the heat content of D. The change of heat energy in D from a time t_1 to a time t_2 is given by

$$\begin{aligned}\text{change of heat in } D = \int_D cu(x, t_2)\,dx - \int_D cu(x, t_1)\,dx& \\ = \int_{t_1}^{t_2} \frac{\partial}{\partial t} \int_D cu\,dx\,dt = \int_{t_1}^{t_2} \int_D \frac{\partial(cu)}{\partial t}(x, t)\,dx\,dt,&\end{aligned}$$

where we have used the fundamental theorem of calculus. Now, by conservation of energy, any change of heat in D must be accounted for by heat flowing in or out of D through its boundary or by heat entering from external sources (e.g., if the body were in a microwave oven). The heat flow is measured by a vector field $\sigma(x, t)$ called the *heat flux*, which points in the direction in which heat is flowing with magnitude the rate energy flowing across a unit area per unit time. If we have a surface S embedded in D with normal n, then the heat flowing across S in the direction pointed to by n in unit time is $\int_S \sigma \cdot n\,ds$. Therefore the heat that flows out of D, i.e., across its boundary, in the time interval $[t_1, t_2]$, is given by

$$\text{heat flow out of } D \int_{t_1}^{t_2} \int_{\partial D} \sigma \cdot \boldsymbol{n}\,ds\,dt = \int_{t_1}^{t_2} \int_D \operatorname{div} \sigma\,dx\,dt,$$

where we have used the divergence theorem (the fundamental theorem of calculus in higher dimensions). We denote the heat entering from external sources by $f(x, t)$, given as energy per unit volume per unit time, so that $\int_{t_1}^{t_2}\int_D f(x,t)\,dx\,dt$ gives amount external heat added to D during $[t_1,t_2]$, and so conservation of energy is expressed by the equation

$$\int_{t_1}^{t_2}\int_D \frac{\partial(cu)}{\partial t}(x,t)\,dx\,dt = -\int_{t_1}^{t_2}\int_D \operatorname{div}\sigma\,ds\,dt + \int_{t_1}^{t_2}\int_D f(x,t)\,dx\,dt, \tag{1.1}$$

for all subbodies $D\subset\Omega$ and times t_1, t_2. Thus the quantity

$$\frac{\partial(cu)}{\partial t} + \operatorname{div}\sigma - f$$

must vanish identically, and so we have established the differential equation

$$\frac{\partial(cu)}{\partial t} = -\operatorname{div}\sigma + f, \quad x\in\Omega, \forall t.$$

To complete the description of heat flow, we need a *constitutive equation*, which tells us how the heat flux depends on the temperature. The simplest is Fourier's law of heat conduction, which says that heat flows in the direction opposite the temperature gradient with a rate proportional to the magnitude of the gradient:

$$\sigma = -\lambda\operatorname{grad}u,$$

where the positive quantity λ is called the conductivity of the material. (Usually λ is just a scalar, but if the material is thermally anisotropic, i.e., it has preferred directions of heat flow, as might be a fibrous or laminated material, λ can be a 3×3 positive-definite matrix.) Therefore we have obtained the equation

$$\frac{\partial(cu)}{\partial t} = \operatorname{div}(\lambda\operatorname{grad}u) + f \text{ in } \Omega\times J.$$

The source function f, the material coefficients c and λ and the solution u can all be functions of x and t. If the material is homogeneous (the same everywhere) and not changing with time, then c and λ are constants and the equation simplifies to the *heat equation*,

$$\mu\frac{\partial u}{\partial t} = \Delta u + \tilde{f},$$

where $\mu = c/\lambda$ and we have $\tilde{f} = f/\lambda$. If the material coefficients depend on the temperature u, as may well happen, we get a nonlinear PDE generalizing the heat equation.

The heat equation not only governs heat flow, but all sorts of diffusion processes where some quantity flows from regions of higher to lower concentration. The heat equation is the prototypical *parabolic* differential equation.

Now suppose our body reaches a steady state: the temperature is unchanging. Then the time derivative term drops and we get

$$-\operatorname{div}(\lambda\operatorname{grad}u) = f \text{ in } \Omega, \tag{1.2}$$

where now u and f are functions of f alone. For a homogeneous material, this becomes the Poisson equation

$$-\Delta u = \tilde{f},$$

the prototypical *elliptic* differential equation. For an inhomogeneous material we can leave the steady state heat equation in *divergence form* as in (1.2), or differentiate out to obtain

$$-\lambda\,\Delta u + \operatorname{grad}\lambda\cdot\operatorname{grad}u = f.$$

To determine the steady state temperature distribution in a body we need to know not only the sources and sinks within the body (given by f), but also what is happening at the boundary $\Gamma := \partial\Omega$. For example a common situation is that the boundary is held at a given temperature

$$u = g \text{ on } \Gamma. \tag{1.3}$$

The PDE (1.2) together with the *Dirichlet boundary condition* (1.3) form an elliptic boundary value problem. Under a wide variety of circumstances this problem can be shown to have a unique solution. The following theorem is one example (although the smoothness requirements can be greatly relaxed).

THEOREM 1.1. *Let Ω be a smoothly bounded domain in $\mathbb{R}^n$, and let $\lambda : \bar{\Omega} \to \mathbb{R}_+$, $f : \bar{\Omega} \to \mathbb{R}$, $g : \Gamma \to \mathbb{R}$ be C^∞ functions. Then there exists a unique function $u \in C^2(\bar{\Omega})$ satisfying the differential equation* (1.2) *and the boundary condition* (1.3)*. Moreover u is C^∞.*

Instead of the Dirichlet boundary condition of imposed temperature, we often see the Neumann boundary condition of imposed heat flux (flow across the boundary):

$$\frac{\partial u}{\partial n} = g \text{ on } \Gamma.$$

For example if $g = 0$, this says that the boundary is insulated. We may also have a Dirichlet condition on part of the boundary and a Neumann condition on another.

1.2. Elastic membranes. Consider a taut (homogeneous isotropic) elastic membrane affixed to a flat or nearly flat frame and possibly subject to a transverse force distribution, e.g., a drum head hit by a mallet. We model this with a bounded domain $\Omega \subset \mathbb{R}^2$ which represents the undisturbed position of the membrane if the frame is flat and no force is applied. At any point x of the domain and any time t, the transverse displacement is given by $u(x,t)$. As long as the displacements are small, then u approximately satisfies the membrane equation

$$\rho \frac{\partial^2 u}{\partial t^2} = k\Delta u + f,$$

where ρ is the density of the membrane (mass per unit area), k is the tension (force per unit distance), and f is the imposed transverse force density (force per unit area). This is a second order hyperbolic equation, *the wave equation.* If the membrane is in steady state, the displacement satisfies the Poisson equation

$$-\Delta u = \tilde{f},$$

$f = f/k$.

1.3. Elastic plates. The derivation of the membrane equation depends upon the assumption that the membrane resists stretching (it is under tension), but does not resist bending. If we consider a *plate*, i.e., a thin elastic body made of a material which resists bending as well as stretching, we obtain instead the plate equation

$$\rho \frac{\partial^2 u}{\partial t^2} = -D\Delta^2 u + f,$$

where D is the bending modulus, a constant which takes into account the elasticity of the material and the thickness of the plate ($D = Et^3/[12(1-\nu^2)]$ where E is Young's modulus and ν is Poisson's ratio). Now the steady state equation is the *biharmonic equation*

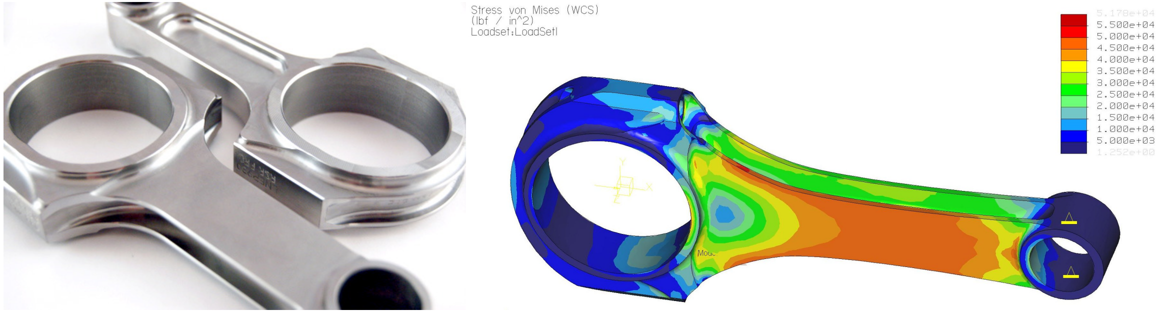
$$\Delta^2 u = \tilde{f}.$$

Later in this course we will study other partial differential equations, including the equations of elasticity, the Stokes and Navier–Stokes equations of fluid flow, and Maxwell's equations of electromagnetics.

2. Some motivations for studying the numerical analysis of PDE

In this course we will study algorithms for obtaining approximate solutions to PDE problems, for example, using the finite element method. Such algorithms are a hugely developed technology (we will, in fact, only skim the surface of what is known in this course), and there are thousands of computer codes implementing them. As an example of the sort of work that is done routinely, here is the result of a simulation using a finite element method to find a certain kind of force distribution, the so-called von Mises stress, engendered in a connecting rod of a Porsche race car when a certain load is applied. The von Mises stress predicts when and where the metal of the rod will deform, and was used to design the shape of the rod.

Figure 1.1. Connector rods designed by LN Engineering for Porsche race cars, and the stress distribution in a rod computed with finite elements.



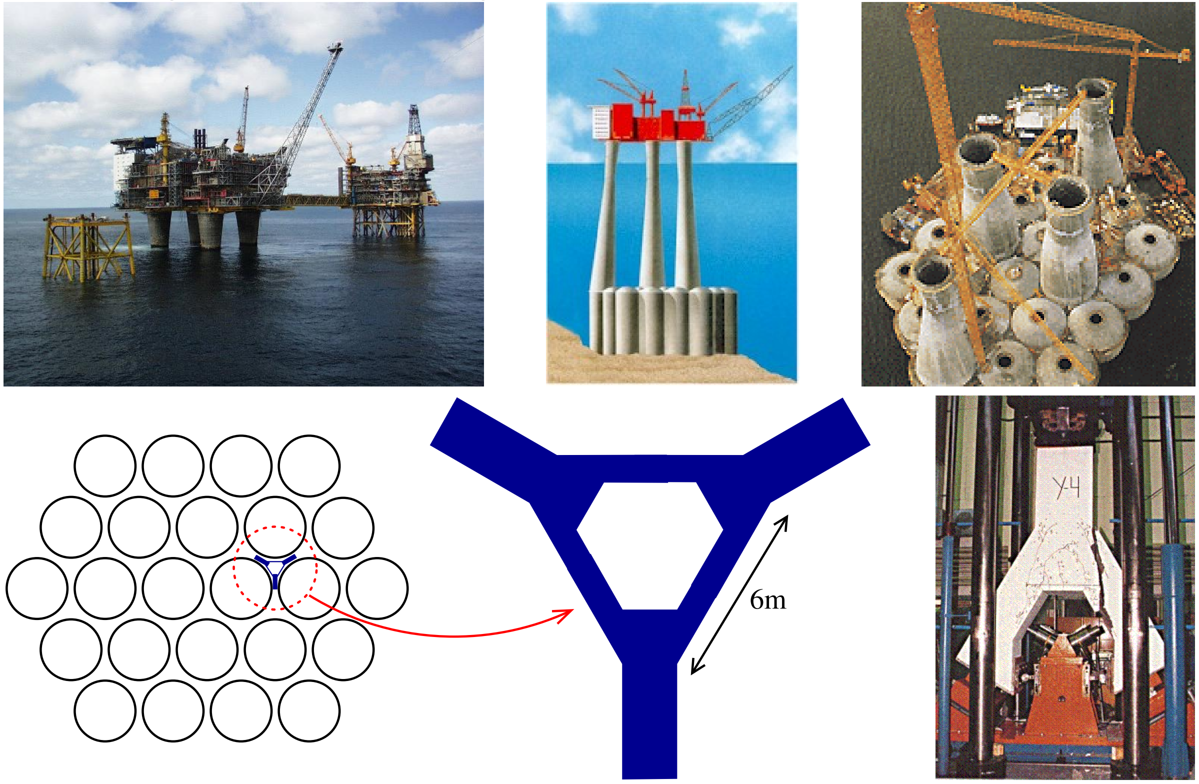
But one should not get the idea that it is straightforward to solve any reasonable PDE problem with finite elements. Not only do challenges constantly arise as practitioners seek to model new systems and solve new equations, but when used with insufficient knowledge and care, even advance numerical software can give disastrous results. A striking example is the sinking of the Sleipner A offshore oil platform in the North Sea in 1991. This occured when the Norwegian oil company, Statoil, was slowly lowering to the sea floor an array of 24 massive concrete tanks, which would support the 57,000 ton platform (which was to accomodate about 200 people and 40,000 tons of drilling equipment). By flooding the tanks in a so-called controlled ballasting operation, they were lowered at the rate of about 5 cm per minute. When they reached a depth of about 65m the tanks imploded and crashed to the sea floor, leaving nothing but a pile of debris at 220 meters of depth. The crash did not result in loss of life, but did cause a seismic event registering 3.0 on the Richter scale, and an economic loss of about $700 million.

An engineering research organization, SINTEF, was appointed to investigate the accident and released a sequence of 16 reports, which they summarized as follows:

> *The conclusion of the investigation was that the loss was caused by a failure in a cell wall, resulting in a serious crack and a leakage that the pumps were not able to cope with. The wall failed as a result of a combination of a serious error in the finite element analysis and insufficient anchorage of the reinforcement in a critical zone.*

A better idea of what was involved can be obtained from this photo and sketch of the platform. The 24 cells and 4 shafts referred to above are shown to the left while at the sea surface. The cells are 12 meters in diameter. The cell wall failure was traced to a tricell, a triangular concrete frame placed where the cells meet, as indicated in the diagram below. To the right of the diagram is pictured a portion of tricell undergoing failure testing.

Figure 1.2. Top row: Offshore platform like the failed Sleipner design, diagram of structure, and concrete cells at sea surface. Bottom row: diagram showing the location and design of a tricell, and tricell failure testing.



The post accident investigation traced the error to inaccurate finite element approximation of one of the most basic PDEs used in engineering, the equations of linear elasticity, which were used to model the tricell (using the popular finite element program NASTRAN). The shear stresses were underestimated by 47%, leading to insufficient design. In particular, certain concrete walls were not thick enough. More careful finite element analysis, made after the accident, predicted that failure would occur with this design at a depth of 62m, which matches well with the actual occurrence at 65m.

CHAPTER 2

The finite difference method for the Laplacian

With the motivation of the previous section, let us consider the numerical solution of the elliptic boundary value problem

$$\Delta u = f \text{ in } \Omega, \quad u = g \text{ on } \Gamma. \tag{2.1}$$

For simplicity we will consider first a very simple domain $\Omega = (0,1) \times (0,1)$, the unit square in $\mathbb{R}^2$. Now this problem is so simplified that we can attack it analytically, e.g., by separation of variables, but it is a very useful *model problem* for studying numerical methods.

1. The 5-point difference operator

Let N be a positive integer and set $h = 1/N$. Consider the *mesh* in $\mathbb{R}^2$

$$\mathbb{R}^2_h := \{\, (mh, nh) \,:\, m, n \in \mathbb{Z} \,\}.$$

Note that each mesh point $x \in \mathbb{R}^2_h$ has four *nearest neighbors* in $\mathbb{R}^2_h$, one each to the left, right, above, and below. We let $\Omega_h = \Omega \cap \mathbb{R}^2_h$, the set of interior mesh points, and we regard this a discretization of the domain Ω. We also define Γ_h as the set of mesh points in $\mathbb{R}^2_h$ which don't belong to Ω_h, but which have a nearest neighbor in Ω_h. We regard Γ_h as a discretization of Γ. We also let $\bar\Omega_h := \Omega_h \cup \Gamma_h$

To discretize (2.1) we shall seek a function $u_h : \bar\Omega_h \to \mathbb{R}$ satisfying

$$\Delta_h u_h = f \text{ on } \Omega_h, \quad u_h = g \text{ on } \Gamma_h. \tag{2.2}$$

Here Δ_h is an operator, to be defined, which takes functions on $\bar\Omega_h$ (*mesh functions*) to functions on Ω_h. It should approximate the true Laplacian in the sense that if v is a smooth function on $\bar\Omega$ and $v_h = v|_{\bar\Omega_h}$ is the associated mesh function, then we want

$$\Delta_h v_h \approx \Delta v|_{\Omega_h}$$

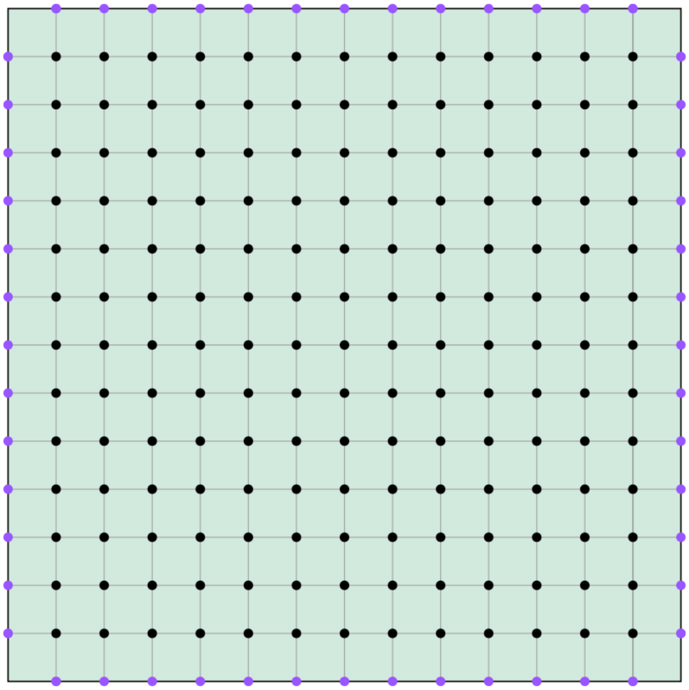
for h small.

Before defining Δ_h, let us turn to the one-dimensional case. That is, given a function v_h defined at the mesh points nh, $n \in \mathbb{Z}$, we want to define a function $D_h^2 v_h$ on the mesh points, so that $D_h^2 v_h \approx v''|_{\mathbb{Z}h}$ if $v_h = v|_{\mathbb{Z}h}$. One natural procedure is to construct the quadratic polynomial p interpolating v_h at three consecutive mesh points $(n-1)h$, nh, $(n+1)h$, and let $D_h^2 v_h(nh)$ be the constant value of p''. This gives the formula

$$D_h^2 v_h(nh) = 2v_h[(n-1)h, nh, (n+1)h] = \frac{v_h\big((n+1)h\big) - 2v_h(nh) + v_h\big((n-1)h\big)}{h^2}.$$

D_h^2 is known as the 3-point difference approximation to d^2/dx^2. We know that if v is C^2 in a neighborhood of nh, then $\lim_{h\to 0} v[x-h, x, x+h] = v''(x)/2$. In fact, it is easy to show

FIGURE 2.1. $\bar{\Omega}_h$ for $h = 1/14$: black: points in Ω_h, purple: points in Γ_h.



by Taylor expansion (do it!), that

$$D_h^2 v(x) = v''(x) + \frac{h^2}{12} v^{(4)}(\xi), \text{ for some } \xi \in \big(x-h, x+h\big),$$

as long as v is C^4 near x. Thus D_h^2 is a second order approximation to d^2/dx^2.

Now returning to the definition of the $\Delta_h \approx \Delta = \partial^2/\partial x^2 + \partial^2/\partial y^2$, we simply use the 3-point approximation to $\partial^2/\partial x^2$ and $\partial^2/\partial y^2$. Writing v_{mn} for $v(mh, nh)$ we then have

$$\begin{aligned}\Delta_h v(mh, nh) &= \frac{v_{m+1,n} - 2v_{mn} + v_{m-1,n}}{h^2} + \frac{v_{m,n+1} - 2v_{mn} + v_{m,n-1}}{h^2} \\ &= \frac{v_{m+1,n} + v_{m-1,n} + v_{m,n+1} + v_{m,n-1} - 4v_{mn}}{h^2}.\end{aligned}$$

From the error estimate in the one-dimensional case we easily get that for $v \in C^4(\bar{\Omega})$,

$$\Delta_h v(mh, nh) - \Delta v(mh, nh) = \frac{h^2}{12}\left[\frac{\partial^4 v}{\partial x^4}(\xi, nh) + \frac{\partial^4 v}{\partial y^4}(mh, \eta)\right],$$

for some ξ, η. Thus:

THEOREM 2.1 (Consistency of Δ_h). *If $v \in C^2(\bar{\Omega})$, then*

$$\lim_{h\to 0} \|\Delta_h v - \Delta v\|_{L^\infty(\Omega_h)} = 0.$$

If $v \in C^4(\bar{\Omega})$, *then*

$$\|\Delta_h v - \Delta v\|_{L^\infty(\Omega_h)} \le \frac{h^2}{6} M_4,$$

where $M_4 = \max(\|\partial^4 v/\partial x^4\|_{L^\infty(\bar{\Omega})}, \|\partial^4 v/\partial y^4\|_{L^\infty(\bar{\Omega})})$.

The discrete PDE $\Delta_h u_h = f$ on Ω_h is a system of $M = (N-1)^2$ linear equations in the unknown values of u_h at the mesh points. Since the values of u_h are given on the boundary mesh points, we may regard (2.2) as a system of M^2 linear equations in M unknowns. For example, in the case $N = 4$, $M = 9$, the system is

$$\begin{pmatrix} -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & -4 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -4 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & -4 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 \end{pmatrix} \begin{pmatrix} u_{1,1} \\ u_{2,1} \\ u_{3,1} \\ u_{1,2} \\ u_{2,2} \\ u_{3,2} \\ u_{1,3} \\ u_{2,3} \\ u_{3,3} \end{pmatrix} = \begin{pmatrix} h^2 f_{1,1} - u_{1,0} - u_{0,1} \\ h^2 f_{2,1} - u_{2,0} \\ h^2 f_{3,1} - u_{3,0} - u_{4,1} \\ h^2 f_{1,2} - u_{0,2} \\ h^2 f_{2,2} \\ h^2 f_{3,2} - u_{4,2} \\ h^2 f_{1,3} - u_{0,3} - u_{1,4} \\ h^2 f_{2,3} - u_{2,4} \\ h^2 f_{3,3} - u_{4,3} - u_{3,4} \end{pmatrix}$$

The matrix may be rewritten as

$$\begin{pmatrix} A & I & O \\ I & A & I \\ O & I & A \end{pmatrix}$$

where I is the 3×3 identity matrix, O is the 3×3 zero matrix, and

$$A = \begin{pmatrix} -4 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & -4 \end{pmatrix}.$$

For general N the matrix can be partitioned into $(N-1) \times (N-1)$ blocks, each in $\mathbb{R}^{(N-1)\times(N-1)}$:

$$\begin{pmatrix} A & I & O & \cdots & O & O \\ I & A & I & \cdots & O & O \\ O & I & A & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & I & A \end{pmatrix},$$

where I and O are the identity and zero matrix in $\mathbb{R}^{(N-1)\times(N-1)}$, respectively, and $A \in \mathbb{R}^{(N-1)\times(N-1)}$ is the tridiagonal matrix with -4 on the diagonal and 1 above and below the diagonal. This assumes the unknowns are ordered

$$u_{1,1}, u_{2,1}, \ldots, u_{N-1,1}, u_{1,2}, \ldots, u_{N-1,N-1},$$

and the equations are ordered similarly.

The matrix can be created as a sparse matrix in python using the SciPy sparse linear algebra package with the following code:

```
I = scipy.sparse.eye(n-1)   # (n-1) x (n-1) identity matrix
e = np.ones(n-1)      # vector (1, 1, ..., 1) of length n-1
e0 = np.ones(n-2)     # vector (1, 1, ..., 1) of length n-2
A = scipy.sparse.diags([e0, -4*e, e0], [-1, 0, 1]) # (n-1) x (n-1) tridiagonal matrix
J = scipy.sparse.diags([e0, e0], [-1, 1])    # same with zeros on diagonal
Lh = scipy.sparse.kronsum(A, J, format='csr')    # the desired matrix
```

Notice that the matrix has many special properties:

- it is sparse with at most 5 elements per row nonzero
- it is block tridiagonal, with tridiagonal and diagonal blocks
- it is symmetric
- it is diagonally dominant
- its diagonal elements are negative, all others nonnegative
- it is negative definite

2. Analysis via a maximum principle

We will now prove that the problem (2.2) has a unique solution and prove an error estimate. The key will be a discrete maximum principle.

THEOREM 2.2 (Discrete Maximum Principle). *Let v be a function on $\bar{\Omega}_h$ satisfying*

$$\Delta_h v \geq 0 \ \text{on}\ \Omega_h.$$

Then $\max_{\Omega_h} v \leq \max_{\Gamma_h} v$. *Equality holds if and only if v is constant.*

PROOF. Suppose $\max_{\Omega_h} v \geq \max_{\Gamma_h} v$. Take $x_0 \in \Omega_h$ where the maximum is achieved. Let x_1, x_2, x_3, and x_4 be the nearest neighbors. Then

$$4v(x_0) = \sum_{i=1}^{4} v(x_i) - h^2 \Delta_h v(x_0) \leq \sum_{i=1}^{4} v(x_i) \leq 4v(x_0),$$

since $v(x_i) \leq v(x_0)$. Thus equality holds throughout and v achieves its maximum at all the nearest neighbors of x_0 as well. Applying the same argument to the neighbors in the interior, and then to their neighbors, etc., we conclude that v is constant. □

REMARKS. 1. The analogous discrete minimum principle, obtained by reversing the inequalities and replacing max by min, holds. 2. This is a discrete analogue of the maximum principle for the Laplace operator.

THEOREM 2.3. *There is a unique solution to the discrete boundary value problem* (2.2).

PROOF. Since we are dealing with a square linear system, it suffices to show nonsingularity, i.e., that if $\Delta_h u_h = 0$ on Ω_h and $u_h = 0$ on Γ_h, then $u_h \equiv 0$. Using the discrete maximum and the discrete minimum principles, we see that in this case u_h is everywhere 0. □

The next result is a statement of maximum norm stability.

THEOREM 2.4. *The solution u_h to* (2.2) *satisfies*

$$\|u_h\|_{L^\infty(\bar\Omega_h)} \le \frac{1}{8}\|f\|_{L^\infty(\Omega_h)} + \|g\|_{L^\infty(\Gamma_h)}. \tag{2.3}$$

This is a stability result in the sense that it states that the mapping $(f,g) \mapsto u_h$ is bounded uniformly with respect to h.

PROOF. We introduce the *comparison function* $\phi(x) = [(x_1 - 1/2)^2 + (x_2 - 1/2)^2]/4$, which satisfies $\Delta_h\phi = 1$ on Ω_h, and $0 \le \phi \le 1/8$ on $\bar\Omega_h$. Set $M = \|f\|_{L^\infty(\Omega_h)}$. Then

$$\Delta_h(u_h + M\phi) = f|_{\Omega_h} + M \ge 0,$$

so

$$\max_{\Omega_h} u_h \le \max_{\Omega_h}(u_h + M\phi) \le \max_{\Gamma_h}(u_h + M\phi) \le \max_{\Gamma_h} g + \frac{1}{8}M.$$

Thus u_h is bounded above by the right-hand side of (2.3). A similar argument applies to $-u_h$ giving the theorem. □

REMARK. A similar argument works with any comparison function satisfying $\Delta_h\phi \ge c_1 > 0$ and $0 \le \phi \le c_2$ (and then the 1/8 in the stability estimate would be replaced with c_2/c_1).

By applying the stability result to the error $u - u_h$ we can bound the error in terms of the *consistency error* $\Delta_h u - \Delta u$.

THEOREM 2.5. *Let u be the solution of the Dirichlet problem* (1.2) *and u_h the solution of the discrete problem* (2.2)*. Then*

$$\|u - u_h\|_{L^\infty(\bar\Omega_h)} \le \frac{1}{8}\|\Delta u - \Delta_h u\|_{L^\infty(\bar\Omega_h)}.$$

PROOF. Since $\Delta_h u_h = f = \Delta u$ on Ω_h, $\Delta_h(u - u_h) = \Delta_h u - \Delta u$. Also, $u - u_h = 0$ on Γ_h. Apply Theorem 2.4 (with u_h replaced by $u - u_h$), we obtain the theorem. □

Combining with Theorem 2.1, we obtain error estimates.

COROLLARY 2.6. *If $u \in C^2(\bar\Omega)$, then*

$$\lim_{h\to 0}\|u - u_h\|_{L^\infty(\bar\Omega_h)} = 0.$$

If $u \in C^4(\bar\Omega)$, then

$$\|u - u_h\|_{L^\infty(\bar\Omega_h)} \le \frac{h^2}{48}M_4,$$

where $M_4 = \max(\|\partial^4 u/\partial x_1^4\|_{L^\infty(\bar\Omega)}, \|\partial^4 u/\partial x_2^4\|_{L^\infty(\bar\Omega)})$.

REMARK. Since $\Delta u = f$, the consistency error can be written $\Delta_h u - f$. Comparing to the difference equations $\Delta_h u_h = f$, we can interpret the consistency error as follows: first we convert the function u into a grid function by restricting it to the grid points and then we plug this grid function into the difference equations and compute the resulting residual. This residual is the consistency error. In the context of finite difference methods it is often called the truncation error or local truncation error.

3. Consistency, stability, and convergence

Now we introduce an abstract framework in which to understand the preceding analysis. It is general enough that it applies, or can be adapted to, a huge variety of numerical methods for PDE. We will keep in mind, as an basic example, the 5-point difference discretization of the Poisson equation with homogeneous boundary conditions, so the PDE problem to be solved is

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

and the numerical method is

$$\Delta_h u_h = f_h \text{ in } \Omega_h, \quad u_h = 0 \text{ on } \Gamma_h.$$

Let X and Y be vector spaces and $L : X \to Y$ a linear operator. Given $f \in Y$, we seek $u \in X$ such that $Lu = f$. This is the problem we are trying to solve. So, for the homogeneous Dirichlet BVP for Poisson's equation, we could take X to be the space of C^2 functions on $\bar{\Omega}$ which vanish on Γ, $Y = C(\bar{\Omega})$, and $L = \Delta$. (Actually, slightly more sophisticated spaces should be taken if we wanted to get a good theory for the Poisson equation, but that won't concern us now.) We shall assume that there is a solution u of the original problem.

Now let X_h and Y_h be finite dimensional normed vector spaces and $L_h : X_h \to Y_h$ a linear operator. Our numerical method, or discretization, is:

$$\text{Given } f_h \in Y_h \text{ find } u_h \in X_h \text{ such that } L_h u_h = f_h.$$

Of course, this is a very minimalistic framework so far. Without some more hypotheses, we do not know if this finite dimensional problem has a solution, or if the solution is unique. And we certainly don't know that u_h is in any sense an approximation of u.

In fact, up until now, there is no way to compare u to u_h, since they belong to different spaces. For this reason, we introduce a *representative* of u, $U_h \in X_h$. We can then talk about the *error* $U_h - u_h$ and its norm $\|U_h - u_h\|_{X_h}$. If this error norm is small, that means that u_h is close to u, or at least close to our representative U_h of u, in the sense of the norm.

In short, we would like the error to be small in norm. To make this precise we do what is always done in numerical analysis: we consider not a single discretization, but a sequence of discretizations. To keep the notation simple, we will now think of $h > 0$ as a parameter tending to 0, and suppose that we have the normed spaces X_h and Y_h and the linear operator $L_h : X_h \to Y_h$ and the element $f_h \in Y_h$ for each h. This family of discretizations is called *convergent* if the norm $\|U_h - u_h\|_{X_h}$ tends to 0 as $h \to 0$.

In our example, we take $X_h = C(\bar{\Omega}_h)$, the space grid functions on $\bar{\Omega}_h$ which vanish on Γ_h, and $Y_h = C(\Omega_h)$, grid functions on the interior domain Ω_h , and equip both with the maximum norm. We also simply define $U_h = u|_{\Omega_h}$. Thus a small error means that u_h is close to the true solution u at all the grid points, which is a desireable result.

Up until this point there is not enough substance to the abstract framework for us to be able to prove a convergence result, because the only connection between the original problem $Lu = f$ and the discrete problems $L_h u_h = f_h$ is that the notations are similar. We surely need some hypotheses. The first of two key hypotheses is *consistency*, which say that, in some sense, the discrete problem is reasonable, in that the solution of the original problem almost satisfies the discrete problem. More precisely, we define the *consistency error* as $L_h U_h - f_h \in Y_h$, a quantity which we can measure using our norm in Y_h. The family of discretizations is called *consistent* if the norm $\|L_h U_h - f_h\|_{Y_h}$ tends to 0 as $h \to 0$.

Not every consistent family of discretizations is convergent (as you can easily convince yourself, since consistency involves the norm in Y_h but not the norm in X_h and for convergence it is the opposite). There is a second key hypothesis, uniform well-posedness of the discrete problems. More precisely, we assume that each discrete problem is uniquely solvable (nonsingular): for every $g_h \in Y_h$ there is a unique $v_h \in X_h$ with $L_h v_h = g_h$. Thus the operator $L_h^{-1} : Y_h \to X_h$ is defined and we call its norm $c_h = \|L_h^{-1}\|_{\mathcal{L}(Y_h,X_h)}$ the *stability constant* of the discretization. The family of discretizations is called *stable* if the stability constants are bounded uniformly in h: $\sup_h c_h < \infty$. Note that stability is a property of the discrete problems and depends on the particular choice of norms, but it does not depend on the true solution u in any way.

With these definition we get a theorem which is trivial to prove, but which captures the underlying structure of many convergence results in numerical PDE.

Theorem 2.7. *Let there be given normed vector spaces X_h and Y_h, an invertible linear operator $L_h : X_h \to Y_h$, an element $f_h \in Y_h$, and a representative $U_h \in X_h$. Define $u_h \in X_h$ by $L_h u_h = f_h$. Then the norm of the error is bounded by the stability constant times the norm of the consistency error. If a family of such discretizations is consistent and stable, then it is convergent.*

Proof. Since $L_h u_h = f_h$,

$$L_h(U_h - u_h) = L_h U_h - f_h.$$

Applying L_h^{-1} we obtain

$$U_h - u_h = L_h^{-1}(L_h U_h - f_h),$$

and taking norms we get

$$\|U_h - u_h\|_{X_h} = \|L_h^{-1}\|_{\mathcal{L}(Y_h,X_h)} \|L_h U_h - f_h\|_{Y_h},$$

which is the desired result. □

Remark. We emphasize that the concepts of convergence, consistency, and stability depend on the choice of norms in X_h, Y_h, and both, respectively. The norm in X_h should be chosen so that the convergence result gives information that is desired. Choosing a weak norm may make the hypotheses easier to verify, but the result of less interest. Similarly, f_h must be chosen in a practical way. We need f_h to compute u_h, so it should be something we know before we solve the problem, typically something easily computed from f. Similarly as well, U_h should be chosen in a reasonable way. For example, choosing $U_h = L_h^{-1} f_h$ would give $U_h = u_h$ so we definitely have a convergent method, but this is cheating: convergence is of no interest with this choice. The one other choice we have at our disposal is the norm on Y_h. This we are free to choose in order to make the hypotheses of consistency and stability possible to verify. Note that weakening the norm on Y_h makes it easier to prove consistency, while strengthening it makes it easier to prove stability.

Returning to our example, we see that the first statement of Theorem 2.1 is just the statement that the method is consistent for any solution $u \in C^2(\bar{\Omega})$, and the second statement says that the consistency error is $O(h^2)$ if $u \in C^4(\bar{\Omega})$. On the other hand, if we apply Theorem 2.4 with $g = 0$, it states that the stability constant $c_h \le 1/8$ for all h, and so the method is stable. We then obtain the convergence result in Corollary 2.6 by the basic result of Theorem 2.7.

4. Fourier analysis

The space of functions $\Omega_h \to \mathbb{R}$, which we denote $C(\Omega_h)$, is isomorphic to $\mathbb{R}^M$, $M = (N-1)^2$. Sometimes we will identify these functions with functions on $\bar{\Omega}_h$ which are zero to Γ_h. With this identification, the discrete Laplacian defines an isomorphism of $C(\Omega_h)$ onto itself. As we just saw, the L^∞ stability constant, $\|\Delta_h^{-1}\|_{\mathcal{L}(L^\infty,L^\infty)} \le 1/8$. In this section we use Fourier analysis to establish a similar L^2 stability result.

First consider the one-dimensional case. With $h = 1/N$ let $I_h = \{h, 2h, \dots, (N-1)h\}$, and let $C(I_h)$ be the space of functions on I_h, which is an $N-1$ dimensional vectorspace. On $C(I_h)$ we define the inner product

$$\langle u, v\rangle_h = h \sum_{k=1}^{N-1} u(kh)v(kh),$$

with the corresponding norm $\|v\|_h$.

The space $C(I_h)$ is a discrete analogue of $L^2(I)$ where I is the unit interval. On this latter space the functions $\sin \pi m x$, $m = 1, 2, \dots$, form an orthogonal basis consisting of eigenfunctions of the operator $-d^2/dx^2$. The corresponding eigenvalues are $\pi^2, 4\pi^2, 9\pi^2, \dots$. We now establish the discrete analogue of this result.

Define $\phi_m \in C(I_h)$ to be the restriction of $\sin \pi m x$ to I_h, i.e., $\phi_m(x) = \sin \pi m x$, $x \in I_h$. It turns out that these mesh functions are precisely the eigenvectors of the operator D_h^2. Indeed

$$D_h^2\phi_m(x) = \frac{\sin \pi m(x+h) - 2\sin \pi m x + \sin \pi m(x-h)}{h^2} = \frac{2}{h^2}(\cos \pi m h - 1)\sin \pi m x.$$

Thus

$$D_h^2\phi_m = -\lambda_m\phi_m, \quad \lambda_m = \frac{2}{h^2}(1 - \cos \pi m h) = \frac{4}{h^2}\sin^2\frac{\pi m h}{2}.$$

Note that

$$0 < \lambda_1 < \lambda_2 < \cdots < \lambda_{N-1} < \frac{4}{h^2}.$$

Note also that for $m \ll N$, $\lambda_m \approx \pi^2 m^2$. In particular $\lambda_1 \approx \pi^2$. To get a strict lower bound we note that $\lambda_1 = 8$ for $N = 2$ and λ_1 increases with N.

Since the operator D_h^2 is symmetric with respect to the inner product on $C(I_h)$, and the eigenvalues λ_m are distinct, it follows that the eigenvectors ϕ_m are mutually orthogonal. (This can also be obtained using trigonometric identities, or by expressing the sin functions in terms of complex exponentials and using the discrete Fourier transform.) Since there are $N-1$ of them, they form a basis of $C(I_h)$.

THEOREM 2.8. *The functions ϕ_m, $m = 1, 2, \dots, N-1$ form an orthogonal basis of $C(I_h)$. Consequently, any function $v \in C(I_h)$ can be expanded as $v = \sum_{m=1}^{N-1} a_m\phi_m$ with $a_m = \langle v, \phi_m\rangle_h/\|\phi_m\|_h^2$, and $\|v\|_h^2 = \sum_{m=1}^{N-1} a_m^2\|\phi_m\|_h^2$.*

From this we obtain immediately a stability result for the one-dimensional Laplacian. If $v \in C(I_h)$ and $D_h^2 v = f$, we expand v in terms of the ϕ_m:

$$v = \sum_{m=1}^{N-1} a_m\phi_m, \quad \|v\|_h^2 = \sum_{m=1}^{N-1} a_m^2\|\phi_m\|_h^2.$$

Then

$$f = -\sum_{m=1}^{N-1} \lambda_m a_m \phi_m, \quad \|f\|_h^2 = \sum_{m=1}^{N-1} \lambda_m^2 a_m^2 \|\phi_m\|_h^2 \geq 8^2 \|v\|_h^2.$$

Thus $\|v\|_h \leq \|f\|_h/8$.

The extension to the two-dimensional case is straightforward. We use the basis $\phi_{mn} = \phi_m \otimes \phi_n$, i.e.,

$$\phi_{mn}(x,y) := \phi_m(x)\phi_n(y), \quad m,n = 1,\ldots,N-1,$$

for $C(\Omega_h)$. It is easy to see that these $(N-1)^2$ functions form an orthogonal basis for $C(\Omega_h)$ equipped with the inner product

$$\langle u, v\rangle_h = h^2 \sum_{m=1}^{N-1}\sum_{n=1}^{N-1} u(mh,nh)v(mh,nh)$$

and corresponding norm $\|\cdot\|_h$. Moreover ϕ_{mn} is an eigenvector of $-\Delta_h$ with eigenvalue $\lambda_{mn} = \lambda_m + \lambda_n \geq 16$. The next theorem follows immediately.

Theorem 2.9. *The operator Δ_h defines an isomorphism from $C(\Omega_h)$ to itself. Moreover $\|\Delta_h^{-1}\| \leq 1/16$ where the operator norm is with respect to the norm $\|\cdot\|_h$ on $C(\Omega_h)$.*

Since the $\|v\|_h \leq \|v\|_{L^\infty(\Omega_h)}$ we also have consistency with respect to the discrete 2-norm. We leave it to the reader to complete the analysis with a convergence result.

5. Analysis via summation by parts

Fourier analysis is not the only approach to get an L^2 stability result. Another uses *summation by parts*, the discrete analogue of integration by parts.

Let v be a mesh function. Define the backward difference operator

$$\partial_x v(mh,nh) = \frac{v(mh,nh) - v((m-1)h,nh)}{h}, \quad 1 \leq m \leq N, \quad 0 \leq n \leq N.$$

In this section we denote

$$\langle v, w\rangle_h = h^2 \sum_{m=1}^{N}\sum_{n=1}^{N} v(mh,nh)w(mh,nh),$$

with the corresponding norm $\|\cdot\|_h$ (this agrees with the notation in the last section for mesh functions which vanish on Γ_h).

Lemma 2.10. *If $v \in C(\Omega_h)$ (the set of mesh functions vanishing on Γ_h), then*

$$\|v\|_h \leq \frac{1}{2}(\|\partial_x v\|_h + \|\partial_y v\|_h).$$

Proof. It is enough to show that $\|v\|_h \leq \|\partial_x v\|_h$. The same will similarly hold for ∂_y as well, and we can average the two results.

For $1 \le m \le N$, $0 \le n \le N$,

$$\begin{aligned}
|v(mh,nh)|^2 &\le \left(\sum_{i=1}^{N} |v(ih,nh) - v((i-1)h,nh)|\right)^2 \\
&= \left(h\sum_{i=1}^{N} |\partial_x v(ih,nh)|\right)^2 \\
&\le \left(h\sum_{i=1}^{N} |\partial_x v(ih,nh)|^2\right)\left(h\sum_{i=1}^{N} 1^2\right) \\
&= h\sum_{i=1}^{N} |\partial_x v(ih,nh)|^2.
\end{aligned}$$

Therefore

$$h\sum_{m=1}^{N} |v(mh,nh)|^2 \le h\sum_{i=1}^{N} |\partial_x v(ih,nh)|^2$$

and

$$h^2\sum_{m=1}^{N}\sum_{n=1}^{N} |v(mh,nh)|^2 \le h^2\sum_{i=1}^{N}\sum_{n=1}^{N} |\partial_x v(ih,nh)|^2,$$

i.e., $\|v\|_h^2 \le \|\partial_x v\|_h^2$, as desired. □

This result is a discrete analogue of Poincaré's inequality, which bounds a function in terms of its gradient as long as the function vanishes on a portion of the boundary. The constant of 1/2 in the bound can be improved. The next result is a discrete analogue of Green's Theorem (essentially, integration by parts).

Lemma 2.11. *If $v, w \in C(\Omega_h)$, then*

$$-\langle \Delta_h v, w\rangle_h = \langle \partial_x v, \partial_x w\rangle_h + \langle \partial_y v, \partial_y w\rangle_h.$$

Proof. Let $v_0, v_1, \ldots, v_N, w_0, w_1, \ldots, w_N \in \mathbb{R}$ with $w_0 = w_N = 0$. Then

$$\begin{aligned}
\sum_{i=1}^{N}(v_i - v_{i-1})(w_i - w_{i-1}) &= \sum_{i=1}^{N} v_i w_i + \sum_{i=1}^{N} v_{i-1}w_{i-1} - \sum_{i=1}^{N} v_{i-1}w_i - \sum_{i=1}^{N} v_i w_{i-1} \\
&= 2\sum_{i=1}^{N-1} v_i w_i - \sum_{i=1}^{N-1} v_{i-1}w_i - \sum_{i=1}^{N-1} v_{i+1}w_i \\
&= -\sum_{i=1}^{N-1}(v_{i+1} - 2v_i + v_{i-1})w_i.
\end{aligned}$$

Hence,

$$-h\sum_{i=1}^{N-1}\frac{v((i+1)h,nh)-2v(ih,nh)+v((i-1)h,nh)}{h^2}w(ih,nh) \\ = h\sum_{i=1}^{N}\partial_x v(ih,nh)\partial_x w(ih,nh),$$

and thus

$$-\langle D_x^2 v, w\rangle_h = \langle \partial_x v, \partial_x w\rangle_h.$$

Similarly, $-\langle D_y^2 v, w\rangle_h = \langle \partial_y v, \partial_y w\rangle_h$, so the lemma follows. □

Combining the discrete Poincaré inequality with the discrete Green's theorem, we immediately get a stability result. If $v \in C(\Omega_h)$, then

$$\|v\|_h^2 \le \frac{1}{2}(\|\partial_x v\|_h^2 + \|\partial_y v\|_h^2) = -\frac{1}{2}\langle \Delta_h v, v\rangle_h \le \frac{1}{2}\|\Delta_h v\|_h \|v\|_h.$$

Thus

$$\|v\|_h \le \|\Delta_h v\|_h, \quad v \in C(\Omega_h),$$

which is a stability result.

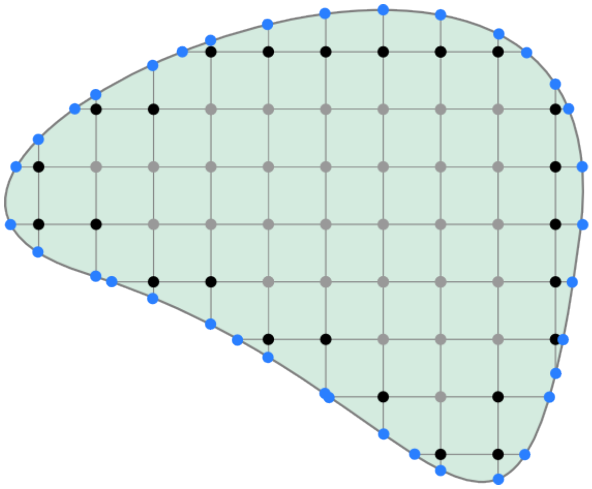
6. Extensions

6.1. Curved boundaries. Thus far we have studied as a model problem the discretization of Poisson's problem on the square. In this subsection we consider a variant which can be used to discretize Poisson's problem on a fairly general domain.

Let Ω be a smoothly bounded open set in $\mathbb{R}^2$ with boundary Γ. We again consider the Dirichlet problem for Poisson's equation, (2.1), and again set $\Omega_h = \Omega \cap \mathbb{R}_h^2$. If $(x, y) \in \Omega_h$ and the segment $(x + sh, y)$, $0 \le s \le 1$ belongs to Γ, then the point $(x + h, y)$, which belongs to Ω_h, is a *neighbor* of (x, y) to the right. If this segment doesn't belong to Ω we define another sort of neighbor to the right, which belongs to Γ. Namely we define the neighbor to be the point $(x + sh, y)$ where $0 < s \le 1$ is the largest value for which $(x + th, y) \in \Omega$ for all $0 \le t < s$. The points of Γ so constructed (as neighbors to the right or left or above or below points in Ω_h) constitute Γ_h. Thus every point in Ω_h has four nearest neighbors all of which belong to $\bar\Omega_h := \Omega_h \cup \Gamma_h$. We also define $\mathring{\Omega}_h$ as those points in Ω_h all four of whose neighbor belong to Ω_h and Ω_h^∂ as those points in Ω_h with at least one neighbor in Γ_h. See Figure 2.2.

In order to discretize the Poisson equation we need to construct a discrete analogue of the Laplacian $\Delta_h v$ for mesh functions v on $\bar\Omega_h$. Of course on $\mathring{\Omega}_h$, $\Delta_h v$ is defined as the usual 5-point Laplacian. For $(x, y) \in \Omega_h^\partial$, let $(x + h_E, y)$, $(x, y + h_N)$, $(x - h_W, y)$, and $(x, y - h_S)$ be the nearest neighbors (with $0 < h_E, h_N, h_W, h_S \le h$), and let v_E, v_N, v_W, and v_S denote the value of v at these four points. Setting $v_0 = v(x, y)$ as well, we will define $\Delta_h v(x, y)$ as a linear combination of these five values of v. In order to derive the formula, we first

FIGURE 2.2. Gray points: $\mathring{\Omega}_h$. Black points: Ω_h^∂. Blue points: Γ_h.



consider approximating $d^2v/dx^2(0)$ by a linear combination of $v(-h_-)$, $v(0)$, and $v(h_+)$, for a function v of one variable. By Taylor's theorem

$$\alpha_- v(-h_-) + \alpha_0 v(0) + \alpha_+ v(h_+) = (\alpha_- + \alpha_0 + \alpha_+)v(0) + (\alpha_+ h_+ - \alpha_- h_-)v'(0) + \frac{1}{2}(\alpha_+ h_+^2 + \alpha_- h_-^2)v''(0) + \frac{1}{6}(\alpha_+ h_+^3 - \alpha_- h_-^3)v'''(0) + \cdots .$$

Thus, to obtain a consistent approximation we must have

$$\alpha_- + \alpha_0 + \alpha_+ = 0, \quad \alpha_+ h_+ - \alpha_- h_- = 0, \quad \frac{1}{2}(\alpha_+ h_+^2 + \alpha_- h_-^2) = 1,$$

which give

$$\alpha_- = \frac{2}{h_-(h_- + h_+)}, \quad \alpha_+ = \frac{2}{h_+(h_- + h_+)}, \quad \alpha_0 = \frac{-2}{h_- h_+}.$$

Note that we have simply recovered the usual divided difference approximation to d^2v/dx^2:

$$\alpha_- v(-h_-) + \alpha_0 v(0) + \alpha_+ v(h_+) = \frac{[v(h_+) - v(0)]/h_+ - [v(0) - v(-h_-)]/h_-}{(h_+ + h_-)/2} = 2v[-h_-, 0, h_+].$$

Returning to the 2-dimensional case, and applying the above considerations to both $\partial^2 v/\partial x^2$ and $\partial^2 v/\partial y^2$ we arrive at the *Shortley–Weller* formula for $\Delta_h v$:

$$\Delta_h v(x,y) = \frac{2}{h_E(h_E + h_W)} v_E + \frac{2}{h_N(h_N + h_S)} v_N + \frac{2}{h_W(h_E + h_W)} v_W + \frac{2}{h_S(h_N + h_S)} v_S - \left(\frac{2}{h_E h_W} + \frac{2}{h_N h_S}\right) v_0.$$

Using Taylor's theorem with remainder we easily calculate that for $v \in C^3(\bar{\Omega})$,

$$\|\Delta v - \Delta_h v\|_{L^\infty(\Omega_h)} \le \frac{2M_3}{3} h,$$

where M_3 is the maximum of the L^∞ norms of the third derivatives of v. Of course at the mesh points in $\mathring{\Omega}_h$, the consistency error is bounded by $M_4h^2/6 = O(h^2)$, as before, but for mesh points neighboring the boundary, it is reduced to $O(h)$.

The approximate solution to (2.1) is $u_h : \bar{\Omega}_h \to \mathbb{R}$ determined again by 2.2. This is a system of linear equations with one unknown for each point of Ω_h. In general the matrix won't be symmetric, but it maintains other good properties from the case of the square:

- it is sparse, with at most five elements per row
- it has negative diagonal elements and non-negative off-diagonal elements
- it is diagonally dominant.

Using these properties we can obtain the discrete maximum principle with virtually the same proof as for Theorem 2.2, and then a stability result as in Theorem 2.4 follows as before. (By contrast, the Fourier analysis approach to stability does not apply when the mesh spacing is not uniform.) In this way we can easily obtain an $O(h)$ convergence result.

We now show how, with a more carefully analysis, we can improve this convergence result. We shall show that, *even though the consistency error is only $O(h)$ at some points, the error is $O(h^2)$ at all mesh points.*

Let X_h denote the space of mesh functions defined on $\bar{\Omega}_h$ and which vanish on the mesh points in Γ_h. On this space we continue to use the maximum norm. Let Y_h denote the space of mesh functions defined on the interior mesh points only, i.e., on Ω_h. On this space we shall use a different norm, namely,

$$\|f\|_{Y_h} := \max\left\{\max_{x\in\mathring{\Omega}_h}|f(x)|, h\max_{x\in\Omega_h^\partial}|f(x)|\right\}. \tag{2.4}$$

Thus we use the maximum norm except with a weight which decreases the emphasis on the points with a neighbor on the boundary. This norm is smaller than the maximum norm, and, measured in this norm, the consistency error is not just $O(h)$ but rather $O(h^2)$:

$$\|\Delta_h u - \Delta u\|_{Y_h} \le \max\left(\frac{M_4}{6}h^2, h\frac{2M_3}{3}h\right) = O(h^2).$$

Now, with respect to a smaller norm in Y_h the condition of stability is more stringent. So the key point is to show that *the Shortley-Weller discrete Laplacian is stable from X_h to Y_h* with this new norm. For the argument we will use the maximum principle with a slightly more sophisticated comparison function.

Before we used as a comparison function $\phi : \bar{\Omega}_h \to \mathbb{R}$ defined by $\phi(x_1, x_2) = [(x_1-1/2)^2 + (x_2-1/2)^2]/4$, where $(1/2, 1/2)$ was chosen as the vertex because it was the center of the square (making $\|\phi\|_{L^\infty}$ as small as possible while satisfying $\Delta_h\phi \equiv 1$). Now, suppose that Ω is contained in the disk of some radius r about some point p. Then we define

$$\phi(x) = \begin{cases} [(x_1-p_1)^2 + (x_2-p_2)^2]/4, & x \in \Omega_h, \\ [(x_1-p_2)^2 + (x_2-p_2)^2]/4 + h, & x \in \Gamma_h \end{cases} \tag{2.5}$$

Thus we perturb the quadratic comparison function by adding h on the boundary. Then ϕ is bounded independent of h ($\|\phi\|_{L^\infty} \le r^2/4 + h \le r^2/4 + 2r$). Moreover $\Delta_h\phi(x) = 1$, if $x \in \mathring{\Omega}_h$, since then ϕ is just the simple quadratic at x and all its neighbors. However, if $x \in \Omega_h^\partial$, then there is an additional term in $\Delta_h\phi(x)$ for each neighbor of x on the boundary (typically one or two). For example, if $(x_1 - h_W, x_2) \in \Gamma_h$ is a neighbor of x and the other neighbors are in $\mathring{\Omega}_h$, then

$$\Delta_h \phi(x) = 1 + \frac{2}{h_W(h_W + h)} h \geq h^{-1},$$

since $h_W \leq h$. Thus we have

$$\Delta_h \phi(x) \geq \begin{cases} 1, & x \in \mathring{\Omega}_h, \\ h^{-1}, & x \in \Omega_h^\partial. \end{cases} \tag{2.6}$$

Now let $v : \bar{\Omega}_h \to \mathbb{R}$ be a mesh function, and set $M = \|\Delta_h v\|_{Y_h}$ (weighted max norm of the Shortley-Weller discrete Laplacian of v). If $x \in \mathring{\Omega}_h$, then $M \geq |\Delta_h v(x)|$ and $\Delta_h \phi(x) = 1$, so

$$\Delta_h (M\phi)(x) \geq |\Delta_h v(x)|.$$

If $x \in \Omega_h^\partial$, then $M \geq h|\Delta_h v(x)|$ and $\Delta_h \phi(x) \geq h^{-1}$, so again

$$\Delta_h (M\phi)(x) \geq |\Delta_h v(x)|.$$

Therefore

$$\Delta_h (v + M\phi) \geq 0 \text{ on } \Omega_h.$$

We can then apply the maximum principle (which easily extends to the Shortley-Weller discrete Laplacian), to get

$$\max_{\bar{\Omega}_h} v \leq \max_{\bar{\Omega}_h} (v + M\phi) \leq \max_{\Gamma_h} (v + M\phi) \leq \max_{\Gamma_h} v + c\|\Delta_h v\|_{Y_h},$$

where $c = \|\phi\|_{L^\infty}$. Of course, we have a similar result for $-v$, so

$$\|v\|_{L^\infty(\bar{\Omega}_h)} \leq \|v\|_{L^\infty(\Gamma_h)} + c\|\Delta_h v\|_{Y_h}.$$

In particular, if v vanishes on Γ_h, then

$$\|v\|_{L^\infty(\bar{\Omega}_h)} \leq c\|\Delta_h v\|_{Y_h}, \quad v \in X_h,$$

which is the desired stability result. As usual, we apply the stability estimate to $v = u - u_h$, and so get the error estimate

$$\|u - u_h\|_{L^\infty(\bar{\Omega}_h)} \leq c\|\Delta_h u - \Delta u\|_{Y_h} = O(h^2).$$

Remark. The perturbation of h on the boundary in the definition (2.5) of the comparison function ϕ, allowed us to place a factor of h in front of the Ω_h^∂ terms in the Y_h norm (2.4) and still obtain stability. For this we needed (2.6) and the fact that the perturbed comparison function ϕ remained bounded independent of h. In fact, we could take a larger perturbation by replacing h with 1 in (2.5). This would lead to a strengthening of (2.6), namely we could replace h^{-1} by h^{-2}, and still have ϕ bounded independently of h. In this way we can prove stability with the same L^∞ norm for X_h and an even weaker norm for Y_h:

$$\|f\|_{Y_h} := \max\left\{ \max_{x\in\mathring{\Omega}_h} |f(x)|, h^2 \max_{x\in\Omega_h^\partial} |f(x)| \right\}.$$

We thus get an even stronger error bound, with $T_h = \Delta_h u - \Delta u$ denoting the consistency error, we get

$$\|u - u_h\|_{L^\infty(\bar{\Omega}_h)} \leq c \max\left\{ \|T_h\|_{L^\infty(\mathring{\Omega}_h)}, h^2 \|T_h\|_{L^\infty(\Omega_h^\partial)} \right\} \leq c \max\left\{ M_4 h^2, M_3 h^3 \right\} = O(h^2).$$

This estimate shows that the points with neighbors on the boundary, despite having the largest consistency error ($O(h)$ rather than $O(h^2)$ for the other grid points), contribute only a small portion of the error ($O(h^3)$ rather than $O(h^2)$).

This example should be another warning to placing too much trust in a naive analysis of a numerical method by just using Taylor's theorem to expand the consistency error. Not only can a method perform worse than this might suggest, because of instability, it can also perform better, because of additional stability properties, as in this example.

6.2. More general PDEs. It is not difficult to extend the method and analysis to more general PDEs. For example, instead of the Poisson equation, we may take

$$\Delta u - a\frac{\partial u}{\partial x_1} - b\frac{\partial u}{\partial x_2} - cu = f,$$

where a, b, and c are continuous coefficient functions on the square $\bar{\Omega}$. The difference method takes the obvious form:

$$\Delta_h u(x) - a(x)\frac{u(x_1+h,x_2)-u(x_1-h,x_2)}{h} - b(x)\frac{u(x_1,x_2+h)-u(x_1,x_2-h)}{h} \\ - c(x)u(x) = f(x), \quad x \in \Omega_h.$$

It is easy to show that the consistency error is $O(h^2)$. As long as the coefficient $c \geq 0$, a version of the discrete maximum principle holds, and one thus obtains stability and convergence.

As an example with a variable coefficient multiplying the highest order term of the PDE, consider the problem

$$\operatorname{div}(a \operatorname{grad} u) = f, \quad \text{i.e.,} \quad \frac{\partial}{\partial x}\big(a\frac{\partial u}{\partial x}\big) + \frac{\partial}{\partial y}\big(a\frac{\partial u}{\partial y}\big) = f,$$

where $a = a(x,y)$ is a scalar function. This is a PDE in *divergence form*, which we saw when we derived the heat equation. It describes the steady-state temperature distribution in a body with a varying, but isotropic, conductivity (the anisotropic case, when a is a matrix-valued function, is more complicated). You might be tempted to differentiate and obtain

$$a\Delta u + \operatorname{grad} a \cdot \operatorname{grad} u = f,$$

and then write down a corresponding finite difference method, but this is to be avoided for several reasons. It leads to a nonsymmetric matrix because of the first order term, and often a matrix for which the maximum principle does not hold. Moreover, it involves $\operatorname{grad} a$ which is not as smooth as a and may not even exist. Thus we derive a difference equation directly from the divergence form. For simplicity, we write it down in the case of one dimension. The extension to two dimensions is then left to the reader.

For the differential equation $(au')' = f$ on an interval with variable coefficient $a(x)$, we use the difference approximation

$$(au')'(x) \approx \frac{1}{h}\left[a(x+h/2)\frac{u(x+h)-u(x)}{h} - a(x-h/2)\frac{u(x)-u(x-h)}{h}\right].$$

It is easy to check that the consistency error for this approximation is $O(h^2)$.

6.3. More general boundary conditions. It is also fairly easy to extend the method to more general boundary conditions, e.g., the Neumann condition $\partial u/\partial n = g$ on all or part of the boundary, although some cleverness is needed to obtain a stable method with consistency error $O(h^2)$ especially on a domain with curved boundary. We will not go into this topic here, but will treat Neumann problems when we consider finite elements.

6.4. Nonlinear problems. Consider, for example, the quasilinear equation

$$\Delta u = F(u, \partial u/\partial x_1, \partial u/\partial x_2),$$

with Dirichlet boundary conditions on the square. Whether this problem has a solution, and whether that solution is unique, or at least locally unique, depends on the nature of the nonlinearity F, and is beyond the scope of these notes. Supposing the problem does have a (locally) unique solution, we may try to compute it with finite differences. A simple scheme is

$$\Delta_h u_h = F(u_h, \partial_{x_1} u_h, \partial_{x_2} u_h), \quad x \in \Omega_h,$$

where we use, e.g., centered differences like

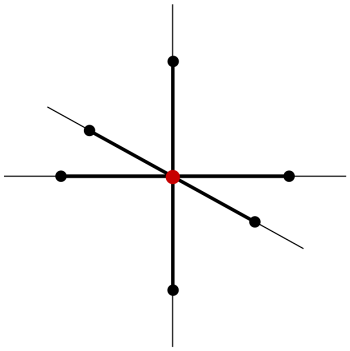
$$\partial_{x_1} u_h(x) = \frac{u(x_1 + h, x_2) - u(x_1 - h, x_2)}{2h}, \quad x \in \Omega_h.$$

Viewing the values of u_h at the M interior mesh points as unknowns, this is a system of M equations in M unknowns. The equations are not linear, but they have the same sparsity pattern as the linear systems we considered earlier: the equation associated to a certain grid point involves at most 5 unknowns, those associated to the grid point and its nearest neighbors.

The nonlinear system is typically solved by an iterative method, very often Newton's method or a variant of it. Issues like solvability, consistency, stability, and convergence can be studied for a variety of particular nonlinear problems. As for nonlinear PDE themselves, many issues arise which vary with the problem under consideration.

6.5. Three dimensions. The 5-point Laplacian on a square grid extends in a straightforward way to a 7-point Laplacian on a cubic grid. Figure 2.3 shows a grid point and its 6 nearest neighbors. The matrix for the 7-point Laplacian is roughly $N^3 \times N^3$ where $N = 1/h$, so much larger for the same h, and solving the matrix equation which arises can be very challenging even for large fast computers.

FIGURE 2.3. A grid point and its six nearest neighbors on a 3-D grid.



CHAPTER 3

Linear algebraic solvers

The finite difference method reduces a boundary value problem for a PDE to a linear algebraic system $Ax = f$, with $A \in \mathbb{R}^{n\times n}$ and $f \in \mathbb{R}^n$. The solution of this system dominates the computation time. (For the 5-point Laplacian on a square with $h = 1/N$, then $n = (N-1)^2$.) The simplest way to solve this is through some variation of Gaussian elimination. Since the matrix A is symmetric positive definite (for the 5-point Laplacian on a square, for instance), we can use the Cholesky decomposition. Cholesky usually requires $O(n^3) = O(N^6)$ floating point additions and multiplications (more precisely $n^3/6 + O(n^2)$, but this is reduced in this case, because of the sparsity of the matrix. Gaussian elimination is not able to exploit the full sparsity of A (since when we factor A as LL^T with L lower triangular, L will be much less sparse that A), but it is able to exploit the fact that A is *banded*: in the natural ordering all the nonzero entries are on the main diagonal or on one of the first $N-1$ sub- or super-diagonals. As a result, the storage count is reduced from $O(n^2) = O(N^4)$ to $O(nN) = O(N^3)$ and the operation count is reduced from $O(N^6)$ to $O(nN^2) = O(N^4)$.

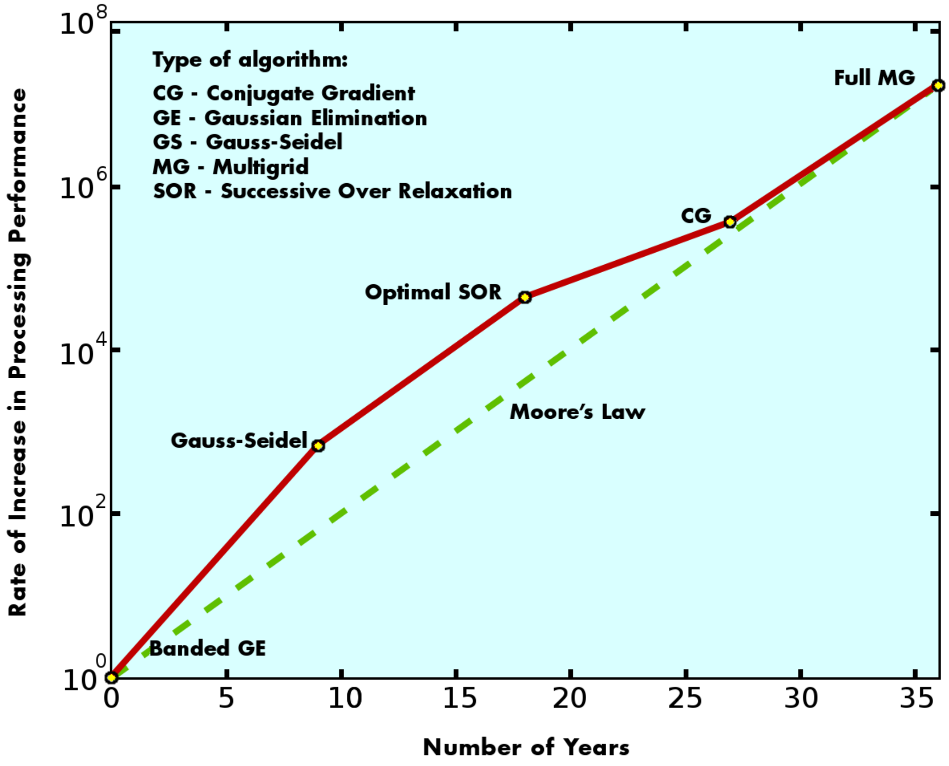
For the 3-dimensional case 7-point Laplacian on a cube, the matrix is $n \times n$ with $n = (N-1)^3$, with the bandwidth $(N-1)^2$. In this case, elimination (e.g., Cholesky) would require storage $O(nN^2) = O(N^5)$ and an operation count of $O(nN^4) = O(N^7)$.

Fortunately, far more efficient ways to solve the equations have been devised, with the best methods requiring essentially $O(N^2)$ operations, i.e., the number being roughly proportional to the number of nonzeros in the matrix. In fact, such algorithmic improvements from the early 1960s to the late 1990s are estimated to account for a speed-up of about 10^7 when solving the 7-point Laplacian or similar problems on a $64\times 64\times 64$ grid, about the same as the speed-up due to improvements in computers over the same period. This is summarized in the following table, taken from Figure 5, page 53 of *Computational Science: Ensuring America's Competitiveness*, a 2005 report to the President of the United States from the President's Information Technology Advisory Committee (PITAC). See also Figure 13, page 32 of the DOE Office of Science report *A science-based case for large-scale simulation*, 2003. Today we solve much larger problems than $64 \times 64 \times 64$ and the speed-up due to algorithms is accordingly much larger, dwarfing that attributable to faster computers.

1. Classical iterations

Gaussian elimination and its variants are called *direct methods*, meaning that they produce the exact solution of the linear system in finite number of steps. (This ignores the effects of round-off error, which is, in fact, a significant issue for some problems.) More efficient methods are *iterative methods*, which start from an initial guess u_0 of the solution of the linear system, and produce a sequence u_1, u_2, …, of iterates which—hopefully—converge to the solution of the linear system. Stopping after a finite number of iterates, gives us an

Figure 3.1. Algorithmic speedup from early 1960s through late 1990s for solving the discrete Laplacian on a cubic mesh of size $64 \times 64 \times 64$. The comparison line labelled "Moore's Law" is based on a speedup by a factor of two every 18 months.



approximate solution to the linear system. This is very reasonable. Since the solution of the linear system is only an approximation of the solution of the PDE problem, there is little point in computing it exactly or nearly exactly. If the numerical discretization provides about 4 significant digits, we should be content if the linear solver provides 5 or 6 digits. Further accuracy in the linear solver serves no purpose.

For an iterative method the goal is, of course, to design an iteration for which

(1) the iteration is efficient, i.e., the amount of work to compute an iteration should not be too large: typically we want it to be proportional to the number n of unknowns;
(2) the rate of convergence of the iterative method is fast, so that not too many iterations are needed.

First we consider some classical iterative methods to solve $Au = f$. One way to motivate such methods is to note that if u_0 is some approximate solution, then the exact solution u may be written $u = u_0 + e$ and the error $e = u - u_0$ is related to the residual $r = f - Au_0$ by the equation $Ae = r$. That is, we can express u as a *residual correction* to u_0: $u = u_0 + A^{-1}(f - Au_0)$. Of course, this merely rephrases the problem, since computing $e = A^{-1}(f - Au_0)$ means solving $Ae = r$ for e, which is as difficult as the original problem of solving $Au = f$ for u. But suppose we can find some matrix B which approximates A^{-1} but is less costly to apply. We are then led to the iteration $u_1 = u_0 + B(f - Au_0)$, which can be repeated to give

$$u_{i+1} = u_i + B(f - Au_i), \quad i = 0, 1, 2, \ldots. \tag{3.1}$$

Of course the effectiveness of such a method will depend on the choice of the *approximate inverse* B. For speed of convergence, we want B to be close to A^{-1}. For efficiency, we want B to be easy to apply. Some typical choices of B are:

- $B = \omega I$ for some $\omega > 0$. As we shall see, this method will converge for symmetric positive definite A if ω is a sufficiently small positive number. This iteration is often called Richardson iteration.
- $B = D^{-1}$ where D is the diagonal matrix with the same diagonal elements as A. This is called the Jacobi method.
- $B = E^{-1}$ where E is the lower triangular matrix with the same diagonal and subdiagonal elements of A. This is the Gauss–Seidel method.

Another way to derive the classical iterative methods, instead of residual correction, is to give a splitting of A as $P + Q$ for two matrices P and Q where P is in some sense close to A but much easier to invert. We then write the equations as $Pu = f - Qu$, which suggests the iteration

$$u_{i+1} = P^{-1}(f - Qu_i).$$

Since $Q = A - P$, this iteration may also be written

$$u_{i+1} = u_i + P^{-1}(f - Au_i).$$

Thus this iteration coincides with (3.1) when $B = P^{-1}$. In particular, the Jacobi method corresponds to the splitting of a matrix into its diagonal and a off-diagonal parts, and the Gauss–Seidel method to the splitting into its lower triangular and the strictly upper triangular parts.

Sometimes the iteration (3.1) is modified to

$$u_{i+1} = (1 - \alpha)u_i + \alpha[u_i + B(f - Au_i)] = u_i + \alpha B(f - Au_i), \quad i = 0, 1, 2, \ldots,$$

for a real parameter α. If $\alpha = 1$, this is the unmodified iteration. For $0 < \alpha < 1$ the iteration has been *damped*, while for $\alpha > 1$ the iteration is *amplified*. The damped Jacobi method will come up below when we study multigrid. The amplified Gauss–Seidel method is known as SOR (successive over-relaxation). This terminology is explained in the next two paragraphs.

Before investigating their convergence, let us particularize the classical iterations to the discrete Laplacian $-\Delta_h^2$ in one or two dimensions. In one dimension, the equations are

$$\frac{-u^{m+1} + 2u^m - u^{m-1}}{h^2} = f^m, \quad m = 1, \ldots, N-1,$$

where $h = 1/N$ and $u^0 = u^N = 0$. The Jacobi iteration is then simply

$$u_{i+1}^m = \frac{u_i^{m-1} + u_i^{m+1}}{2} + \frac{h^2}{2} f^m, \quad m = 1, \ldots, N-1,$$

The error satisfies

$$e_{i+1}^m = \frac{e_i^{m-1} + e_i^{m+1}}{2},$$

so at each iteration the error at a point is set equal to the average of the errors at the neighboring points at the previous iteration. The same holds true for the 5-point Laplacian in two dimensions, except that now there are four neighboring points. In an old terminology, updating the value at a point based on the values at the neighboring points is called *relaxing* the value at the point.

For the Gauss–Seidel method, the corresponding equations are

$$u_{i+1}^m = \frac{u_{i+1}^{m-1} + u_i^{m+1}}{2} + \frac{h^2}{2} f^m, \quad m = 1, \ldots, N-1,$$

and

$$e_{i+1}^m = \frac{e_{i+1}^{m-1} + e_i^{m+1}}{2}, \quad m = 1, \ldots, N-1.$$

We can think of the Jacobi method as updating the value of u at all the mesh points simultaneously based on the old values, while the Gauss–Seidel method updates the values of one point after another always using the previously updated values. For this reason the Jacobi method is sometimes referred to as *simultaneous relaxation* and the Gauss–Seidel method as *successive relaxation* (and amplified Gauss–Seidel as successive overrelaxation). Note that the Gauss–Seidel iteration gives different results if the unknowns are reordered. (In fact, from the point of view of convergence of Gauss–Seidel, there are better orderings than just the naive orderings we have taken so far.) By contrast, the Jacobi iteration is unaffected by reordering of the unknowns. The Jacobi iteration is very naturally a *parallel* algorithm: if we have many processors, each can independently update one or several variables.

Our next goal is to investigate the convergence of (3.1). Before doing so we make some preliminary definition and observations. First we recall that a sequence of vectors or matrices X_i *converges linearly* to a vector or matrix X if there exists a positive number $r < 1$ and a number C such that

$$\|X - X_i\| \le C r^i, \quad i = 1, 2, \ldots. \tag{3.2}$$

In particular this holds (with $C = \|X - X_0\|$) if $\|X - X_{i+1}\| \le r \|X - X_i\|$ $i = 0, 1, \ldots$. For a linearly convergent sequence, the *rate of linear convergence* is the infimum of all r for which there exists a C such that (3.2) holds. In a finite dimensional vector space, both the notion of linear convergence and the rate of linear convergence are independent of a choice of norm. In investigating iterative methods applied to problems with a mesh size parameter h, we will typically find that the rate of linear convergence depends on h. Typical is an estimate like $\|X_i\| \le C r^i$ where all we can say about r is $r \le 1 - ch^p$ for some positive constants c and p. In order to interpret this, suppose that we want the error to be less than some tolerance $\epsilon > 0$. Thus we need to take m iterations with $Cr^m \le \epsilon$, or $r^m \le C^{-1}\epsilon$, or $m \ge |\log(C^{-1}\epsilon)|/|\log r|$ (note that $\log r < 0$ and $\log(C^{-1}\epsilon) < 0$ unless already $\|X - X_0\| \le \epsilon$). Now, for $r = 1 - ch^p$, $|\log r| \approx |ch^p|$, so the number of iterations needed will be about $m = Kh^{-p}$, with $K = c^{-1}|\log(C^{-1}\epsilon)|$. In short, linear convergence with rate $r = 1 - ch^p$ means that the number of iterations required to reduce the error to a given tolerance will be $O(h^{-p})$.

Next we recall that the *spectrum* $\sigma(G)$ of a matrix $G \in \mathbb{R}^{n \times n}$ is its set of eigenvalues, a set of at most n complex numbers. The *spectral radius* $\rho(G) = \max_{\lambda \in \sigma(G)} |\lambda|$. Now consider the L^2-matrix norm $\|G\|_2$ corresponding to the Euclidean norm on $\mathbb{R}^n$. Then

$$\|G\|_2^2 = \sup_{0 \ne x \in \mathbb{R}^n} \frac{(Gx)^T Gx}{x^T x} = \sup_{0 \ne x \in \mathbb{R}^n} \frac{x^T (G^T G) x}{x^T x} = \rho(G^T G),$$

(G^TG is a symmetric positive semidefinite matrix and its spectral radius is the maximum of its Rayleigh quotient). That is, $\|G\|_2 = \sqrt{\rho(G^TG)}$. If G is symmetric, then $G^TG = G^2$, so its eigenvalues are just the squares of the eigenvalues of G, and $\rho(G^TG) = \rho(G^2) = \rho(G)^2$, so $\|G\|_2 = \rho(G)$. Independently of whether G is symmetric or not, for any choice of norm on $\mathbb{R}^n$, the corresponding matrix norm certainly satisfies $\|G\| \geq \rho(G)$. The next theorem shows that equality holds or nearly holds for some choice of the norm.

THEOREM 3.1. *Let $G \in \mathbb{R}^{n\times n}$ be diagonalizable. Then there exists a norm on $\mathbb{R}^n$ such that the corresponding matrix norm satisfies $\|G\| = \rho(G)$. If G is not diagonalizable, then for every $\epsilon > 0$ we can still choose the vector norm so that $\|G\| \leq \rho(G) + \epsilon$.*

PROOF. That G is diagonalizable means that there is an invertible matrix S such that $SGS^{-1} = D$ where D is a diagonal matrix with the eigenvalues of G on its diagonal. We select as the vector norm $\|x\| := \|Sx\|_\infty$. This leads to $\|G\| = \|SGS^{-1}\|_\infty = \|D\|_\infty = \rho(A)$ (since the infinity matrix norm is the maximum of the row sums).

In the nondiagonalizable case we may use the Jordan canonical form to write $SGS^{-1} = J$ where S is an invertible matrix and J has the eigenvalues of G on the diagonal, 0's and ϵ's on the first superdiagonal, and 0's everywhere else. (The usual Jordan canonical form is the case $\epsilon = 1$, but if we conjugate a Jordan block by the matrix $\operatorname{diag}(1, \epsilon, \epsilon^2, \ldots)$ the 1's above the diagonal are changed to ϵ.) Again we select as the vector norm $\|x\| := \|Sx\|_\infty$, so again $\|G\| = \|J\|_\infty$, and this is now bounded by $\rho(A) + \epsilon$. □

An important corollary of this result is a criterion for when the powers of a matrix tend to zero.

THEOREM 3.2. *For $G \in \mathbb{R}^{n\times n}$, $\lim_{i\to\infty} G^i = 0$ if and only if $\rho(G) < 1$, and in this case the convergence is linear with rate $\rho(G)$.*

PROOF. For any choice of vector norm $\|G^n\| \geq \rho(G^n) = \rho(G)^n$, so if $\rho(G) \geq 1$, then G^n does not converge to 0.

Conversely, if $\rho(G) < 1$, then for any $\bar\rho \in (\rho(G), 1)$ we can find an operator norm so that $\|G\| \leq \bar\rho$, and then $\|G^n\| \leq \|G\|^n \leq \bar\rho^n \to 0$. □

We now apply this result to the question of convergence of the iteration (3.1), which we write as

$$u_{i+1} = (I - BA)u_i + Bf = Gu_i + Bf,$$

where the *iteration matrix* $G = I - BA$. The equation $u = Gu + Bf$ is certainly satisfied (where u is the exact solution), and so we have another way to view a classical iteration: it is a one-point iteration for this fixed point equation. The error then satisfies $e_{i+1} = Ge_i$, and the method converges for all starting values $e_0 = u - u_0$ if and only if $\lim_{i\to\infty} G^i = 0$, which, as we have just seen, holds if and only if $\rho(G) < 1$, in which case the convergence is linear with rate of linear convergence $\rho(G)$. Now the condition that the $\rho(G) < 1$ means that all the eigenvalues of $G = I - BA$ lie strictly inside the unit circle in the complex plane, or equivalently that all the eigenvalues of BA lie strictly inside the circle of radius 1 in the complex plane centered at the point 1. If BA has real eigenvalues, then the condition becomes that all the eigenvalues of BA belong to the interval $(0, 2)$. Note that, if A is symmetric positive definite (SPD) and B is symmetric, then BA is symmetric with respect to the inner product $\langle u, v\rangle_A = u^TAv$, so BA does indeed have real eigenvalues in that case.

As a first example, we consider the convergence of the Richardson method for an SPD matrix A. Since the matrix A is SPD, it has a basis of eigenvectors with positive real eigenvalues

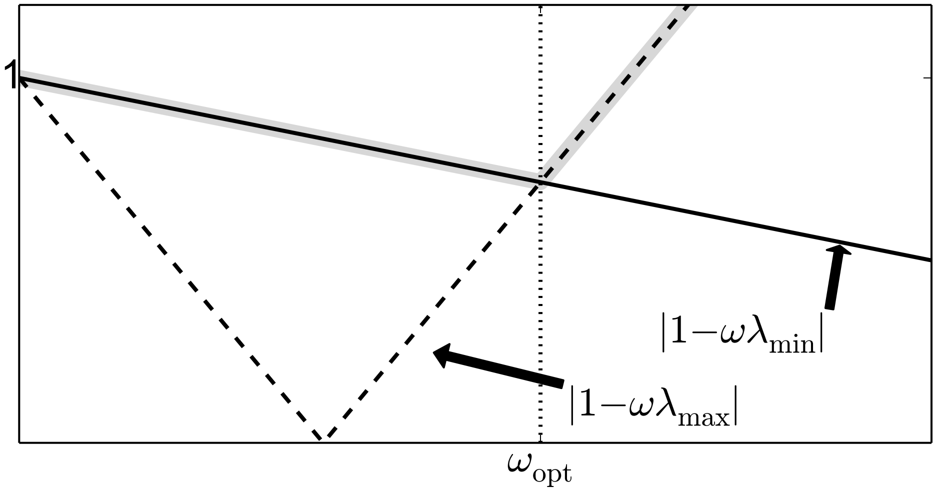
$$0 < \lambda_{\min}(A) = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n = \lambda_{\max}(A) = \rho(A).$$

The eigenvalues of $BA = \omega A$ are then $\omega\lambda_i$, $i = 1, \ldots, n$, and the iteration converges if and only if $0 < \omega < 2/\lambda_{\max}$.

THEOREM 3.3. *Let A be an SPD matrix. Then the Richardson iteration $u_{m+1} = u_m + \omega(f - Au_m)$ is convergent for all choices of u_0 if and only if $0 < \omega < 2/\lambda_{max}(A)$. In this case the rate of convergence is*

$$\max(|1 - \omega\lambda_{max}(A)|, |1 - \omega\lambda_{min}(A)|).$$

FIGURE 3.2. Optimal choice of ω for Richardson iteration.



Note that the optimal choice is given by $\omega\lambda_{\max}(A) - 1 = 1 - \omega\lambda_{\min}(A)$, i.e., $\omega_{\text{opt}} = 2/[\lambda_{\max}(A) + \lambda_{\min}(A)]$, and, with this choice of ω, the rate of convergence is

$$\frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} = \frac{\kappa - 1}{\kappa + 1},$$

where $\kappa = \lambda_{\max}(A)/\lambda_{\min}(A) = \|A\|_2\|A^{-1}\|_2$ is the *spectral condition number* of A. Of course, in practice we do not know the eigenvalues, so we cannot make the optimal choice. But even if we could, we would find very slow convergence when κ is large, as it typically is for discretizations of PDE.

For example, if we consider $A = -D_h^2$, then $\lambda_{\min} \approx \pi^2$, $\lambda_{\max} \approx 4/h^2$, so $\kappa = O(h^{-2})$, and the rate of convergence is like $1 - ch^2$ for some c. Thus the converge is indeed very slow (we will need $O(h^{-2})$ iterations).

Note that for $A = -D_h^2$ the Jacobi method coincides with the Richardson method with $\omega = h^2/2$. Since $\lambda_{\max}(A) < 4/h^2$, we have $\omega < 2/\lambda_{\max}(A)$ and the Jacobi method is convergent. But again convergence is very slow, with a rate of $1 - O(h^2)$. In fact for any $0 < \alpha \le 1$, the damped Jacobi method is convergent, since it coincides with the Richardson method with $\omega = \alpha h^2/2$.

For the Richardson, Jacobi, and damped Jacobi iterations, the approximate inverse B is symmetric, but this is not the case for Gauss–Seidel, in which B is the inverse of the lower triangle of A. Of course we get a similar method if we use B^T, the inverse of the upper triangle of A. If we take two steps of Gauss–Seidel, one with the lower triangle and one with the upper triangle, the iteration matrix is

$$(I - B^T A)(I - BA) = I - (B^T + B - B^T AB)A,$$

so this double iteration is itself a classical iteration with the approximate inverse

$$\bar{B} := B^T + B - B^T AB. \tag{3.3}$$

This iteration is called *symmetric Gauss–Seidel.* Now, from the definition of $\bar{B}$, we get the identity

$$\|v\|_A^2 - \|(I - BA)v\|_A^2 = \langle \bar{B}Av, v\rangle_A. \tag{3.4}$$

It follows that $\langle \bar{B}Av, v\rangle_A \le \|v\|_A^2$, and hence that $\lambda_{\max}(\bar{B}A) \le 1$. Thus the symmetrized Gauss–Seidel iteration is convergent if and only if $\lambda_{\min}(\bar{B}A) > 0$, i.e., if and only if $\bar{B}A$ is SPD with respect to the A inner product. This is easily checked to be equivalent to $\bar{B}$ being SPD with respect to the usual inner product. When this is the case (3.4) implies that $\|(I - BA)v\|_A < \|v\|_A$ for all nonzero v, and hence the original iteration is convergent as well.

In fact the above argument did not use any properties of the original approximate inverse B. So what we have really proved this more general theorem.

Theorem 3.4. *With A an SPD matrix, let $u_{i+1} = u_i + B(f - Au_i)$ be an iterative method in residual correction form, and consider the symmetrized iteration, i.e., $u_{i+1} = u_i + \bar{B}(f - Au_i)$ with $\bar{B}$ given by* (3.3). *Then the symmetrized iteration is convergent if and only if $\bar{B}$ is SPD, and, in that case, the original iteration is convergent as well.*

Returning to Gauss–Seidel, we write $A = L + D + L^T$ where D is diagonal and L strictly lower diagonal, so $B = (L + D)^{-1}$ and

$$\begin{aligned}\bar{B} = B^T + B - B^T AB = B^T(B^{-1} + B^{-T} - A)B \\ = B^T[(L + D) + (L^T + D) - (L + D + L^T)]B = B^T DB,\end{aligned}$$

which is clearly SPD whenever A is. Thus we have proven:

Theorem 3.5. *The Gauss–Seidel and symmetric Gauss–Seidel iterations are convergent for any symmetric positive definite linear system.*

It is worth remarking that the same result is *not* true for the Jacobi iteration: although convergence can be proven for many of the SPD matrices that arise from discretizations of PDE, it is easy to construct an SPD matrix for which Jacobi iteration does not converge. As to the speed of convergence, for Gauss–Seidel applied to the discrete Laplacian, the analysis is much trickier than for Jacobi, but it can again be proven (or convincingly demonstrated via simple numerical experiments) that for $A = -D_h^2$ the rate of convergence is again is about $1 - ch^2$, as for Jacobi, although the constant c is about twice as big for Gauss–Seidel as for Jacobi.

For both of these iterations, applied to the 5-point Laplacian, the cost of an iteration is $O(n) = O(N^2)$, and we need $O(h^{-2}) = O(N^2)$ iterations to achieve a given decrease in the error. Thus the total cost will be $O(N^4)$ operations to achieve a given reduction factor, the same order as for banded Cholesky. In 3 dimensions, the situation is more favorable for the iterative methods. In this case, the cost of an iteration is $O(n) = O(N^3)$, and we will again need $O(N^2)$ iterations, for a total cost of $O(N^5)$, compared to $O(N^7)$ for banded Cholesky.

For SOR, the analysis is more complicated, but can be carried out in a similar way. A careful analysis for Δ_h, which can be found in many texts, shows that there is an optimal value of the relaxation parameter α, and for that value, the spectral radius behaves like $1 - ch$ rather than $1 - ch^2$. This is significantly more efficient, giving $O(N)$ rather than $O(N^2)$ operations. However, in practice it can be difficult or impossible to find the optimal relaxation parameter, and the convergence is quite sensitive to the choice of parameter.

2. The conjugate gradient method

2.1. Line search methods and the method of steepest descents. We now restrict to the case where A is SPD. In this case the solution u of $Au = f$ is also the unique minimizer of the function $F : \mathbb{R}^n \to \mathbb{R}$,

$$F(v) = \frac{1}{2}v^T Av - v^T f$$

This is a quadratic functional with a unique minimum, which can be found by solving the equation $\nabla F(u) = 0$, i.e., $Au = f$. Now, for any $v, w \in \mathbb{R}^n$, we can write

$$\frac{1}{2}v^T Av = \frac{1}{2}[w + (v - w)]^T A[w + (v - w)] = \frac{1}{2}w^T Aw + \frac{1}{2}(v - w)^T A(v - w) + (v - w)^T Aw,$$

so

$$F(v) = F(w) + \frac{1}{2}(v - w)^T A(v - w) + (v - w)^T (Aw - f).$$

If we take $w = u$ the last term vanishes, giving

$$F(v) = F(u) + \frac{1}{2}(v - u)^T A(v - u), \tag{3.5}$$

which again shows that u is the unique minimizer of F, and helps us to visualize the graph of the function $F(v)$. Its graph is an upward opening paraboloid with vertex at the point where $v = F(u)$. At that point F takes its minimum value $F(u) = -\frac{1}{2}u^T Au$.

Now one very general way to try to search for a specific point in a vector space is through a *line search* method:

```
choose initial iterate $u_0$
for $i = 0, 1, \ldots$
    choose $s_i \in \mathbb{R}^n$
    choose $\lambda_i \in \mathbb{R}$
    set $u_{i+1} = u_i + \lambda_i s_i$
end
```

At each step the *search direction* s_i and *step length* λ_i are chosen to, hopefully, get us nearer to the desired point. If the point we are searching for minimizes a function $F : \mathbb{R}^n \to \mathbb{R}$ (quadratic or not), a reasonable choice (but certainly not the only reasonable choice) of search direction is the direction of steepest descent of F at u_i, i.e., $s_i = -\nabla F(u_i)$. In our quadratic case, the steepest descent direction is $s_i = f - Au_i = r_i$, the residual. Thus the Richardson iteration can be viewed as a line search method with steepest descent as search direction, and a fixed step size.

Also for a general minimization problem, for any choice of search direction, there is an obvious choice of stepsize, namely we can do an *exact line search* by minimizing the function of one variable $\lambda \mapsto F(u_i + \lambda s_i)$. Thus we must solve $s_i^T \nabla F(u_i + \lambda s_i) = 0$, which, in the quadratic case, gives

$$\lambda_i = \frac{s_i^T r_i}{s_i^T A s_i}. \tag{3.6}$$

If we choose the steepest descent direction with exact line search, we get $s_i = r_i$, $\lambda_i = r_i^T r_i / r_i^T A r_i$, giving the *method of steepest descents*:

```
choose initial iterate $u_0$
for $i = 0, 1, \ldots$
    set $r_i = f - Au_i$
    set $u_{i+1} = u_i + \frac{r_i^T r_i}{r_i^T A r_i} r_i$
end
```

Thus the method of steepest descents is a variant of the Richardson iteration $u_{i+1} = u_i + \omega(f - Au_i)$ in which the parameter ω depends on i. It does not fit in the category of simple iterations $u_{i+1} = Gu_i + Bf$ with a fixed iteration matrix G which we analyzed in the previous section, so we shall need to analyze it by other means.

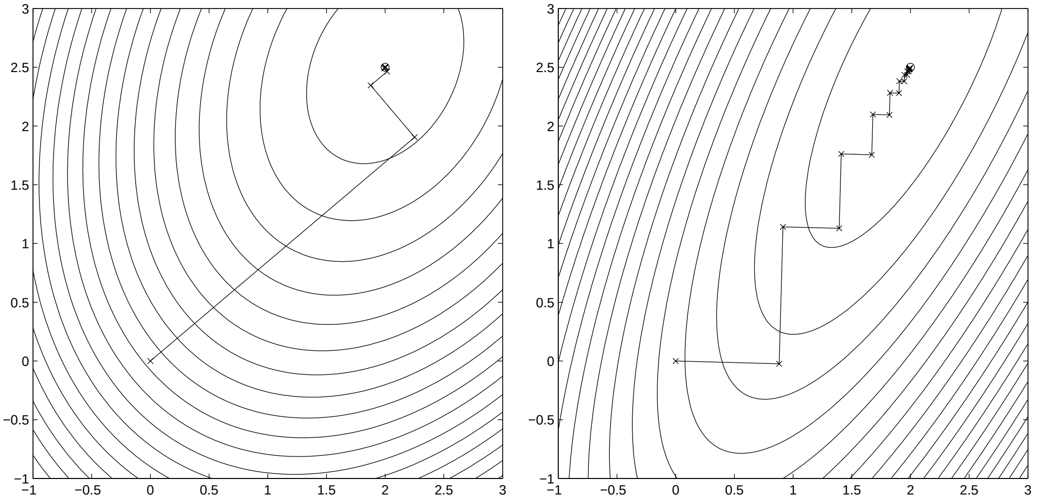
Let us consider the work per iteration of the method of steepest descents. As written above, it appears to require two matrix-vector multiplications per iteration, one to compute Ar_i used in defining the step length, and one to compute Au_i used to compute the residual, and one to compute Ar_i used in defining the step length. However, once we have computed $p_i := Ar_i$ and the step length λ_i we can compute the next residual without an additional matrix-vector multiplication, since $u_{i+1} = u_i + \lambda_i r_i$ implies that $r_{i+1} = r_i - \lambda_i p_i$. Thus we can write the algorithm as

```
choose $u_0$
set $r_0 = f - Au_0$
for $i = 0, 1, \ldots$
    set $p_i = Ar_i$
    set $\lambda_i = \frac{r_i^T r_i}{r_i^T p_i}$
    set $u_{i+1} = u_i + \lambda_i r_i$
    set $r_{i+1} = r_i - \lambda_i p_i$
end
```

Thus, for each iteration we need to compute one matrix-vector multiplication, two Euclidean inner products, and two operations which consist of a scalar-vector multiplication and a vector-vector additions (referred to as a SAXPY operation). The matrix-vector multiplication involves roughly one addition and multiplication for each nonzero in the matrix, while the inner products and SAXPY operations each involve n multiplications and additions. If A is sparse with $O(n)$ nonzero elements, the entire cost per iteration is $O(n)$ operations.

We shall show below that if the matrix A is SPD, the method of steepest descents converges to the solution of $Au = f$ for any initial iterate u_0, and that the convergence is linear with the same rate of convergence as we found for Richardson extrapolation with the optimal parameter, namely $(\kappa - 1)/(\kappa + 1)$ where κ is the spectral condition number of A. This means, again, that the convergence is slow if the condition number is large. This is quite easy to visualize already for 2×2 matrices. See Figure 3.3.

FIGURE 3.3. Convergence of steepest descents with a quadratic cost function. Left: condition number 2; right: condition number: 10.



2.2. The conjugate gradient method. The slow convergence of the method of steepest descents motivates a far superior line search method, the *conjugate gradient method.* CG also uses exact line search to choose the step length, but uses a more sophisticated choice of search direction than steepest descents.

For any line search method with exact line search, $u_1 = u_0 + \lambda_0 s_0$ minimizes F over the 1-dimensional affine space $u_0 + \text{span}[s_0]$, and then $u_2 = u_0 + \lambda_0 s_0 + \lambda_1 s_1$ minimizes F over the 1-dimensional affine space $u_0 + \lambda_0 s_0 + \text{span}[s_1]$. However u_2 does not minimize F over the 2-dimensional affine space $u_0 + \text{span}[s_0, s_1]$. If that were the case, then for 2-dimensional problems we would have $u_2 = u$ and we saw that that was far from the case for steepest descents.

However, it turns out that there is a simple condition on the search directions s_i that ensures that u_2 *is* the minimizer of F over $u_0 + \text{span}[s_0, s_1]$, and more generally that u_i is the minimizer of F over $u_0 + \text{span}[s_0, \dots, s_{i-1}]$. Such a choice of search directions is very favorable. While we only need do 1-dimensional minimizations, after k steps we end up finding the minimizer in an k-dimensional space. In particular, as long as the search directions are linearly independent, this implies that $u_n = u$.

THEOREM 3.6. *Suppose that u_i are defined by exact line search using search directions which are A-orthogonal: $s_i^T A s_j = 0$ for $i \neq j$. Then*

$$F(u_i) = \min\{\, F(v) \mid v \in u_0 + \operatorname{span}[s_0, \ldots, s_{i-1}] \,\}.$$

PROOF. Write W_i for $\operatorname{span}[s_0, \ldots, s_{i-1}]$, so $u_i \in u_0 + W_i$ and we wish to prove that u_i minimizes F over $u_0 + W_i$. This is at least true for $i = 1$, since we use exact line search. The proof is by induction on i, so we assume that it is true and must prove that u_{i+1} minimizes F over $u_0 + W_{i+1}$. Now $u_0 + W_{i+1} = \{\, y + \lambda s_i \mid y \in u_0 + W_i,\ \lambda \in \mathbb{R} \,\}$, so we need to show that

$$F(u_{i+1}) = \min_{\substack{y \in u_0 + W_i \\ \lambda \in \mathbb{R}}} F(y + \lambda s_i).$$

The key point is that the function $(y, \lambda) \mapsto F(y + \lambda s_i)$ decouples into the sum of a function of y which does not depend on λ plus a function of λ which does not depend on y. This is because $u_i \in u_0 + W_i$, so $s_i^T A u_i = s_i^T A u_0 = s_i^T A y$ for any $y \in u_0 + W_i$, thanks to the A-orthogonality of the search directions. Thus

$$\begin{aligned} F(y + \lambda s_i) &= \frac{1}{2} y^T A y + \lambda s_i^T A y + \frac{\lambda^2}{2} s_i^T A s_i - y^T f - \lambda s_i^T f \\ &= F(y) + \left[\frac{\lambda^2}{2} s_i^T A s_i - \lambda s_i^T (f - A u_i) \right]. \end{aligned}$$

Thus the minimum is obtained when $y \in u_0 + W_i$ minimizes $F(y)$, which by the inductive hypothesis occurs when $y = u_i$, and when $\lambda \in \mathbb{R}$ minimizes the term in brackets, which just gives us $\lambda = s_i^T (f - A u_i)/s_i^T A s_i$, the formula for exact line search. Thus the minimizer of F over $u_0 + W_{i+1}$ is indeed $u_i + \lambda_i s_i = u_{i+1}$. □

Now $F(v) = \frac{1}{2}\|v - u\|_A^2 + F(u)$ by (3.5), so u_i minimizes F over some set if and only if it minimizes the function $v \mapsto \|v - u\|_A$ over the same set. Thus we have the following corollary.

COROLLARY 3.7. *If u_i are defined by exact line search using search directions which are A-orthogonal, then u_i minimizes the A-norm of the error over $u_0 + W_i$:*

$$\|u - u_i\|_A = \min\{\, \|u - v\|_A \mid v \in u_0 + W_i \,\}$$

where u is the exact solution and $W_i = \operatorname{span}[s_0, \ldots, s_{i-1}]$.

Any method which uses A-orthogonal (also called "conjugate") search directions has the nice property of the theorem. However it is not so easy to construct such directions. By far the most useful method is the method of conjugate gradients, or the CG method, which defines the search directions by A-orthogonalizing the residuals $r_i = f - A u_i$:

- $s_0 = r_0$
- $s_i = r_i - \displaystyle\sum_{j=0}^{i-1} \frac{s_j^T A r_i}{s_j^T A s_j} s_j.$

This sequence of search directions, together with the exact line search choice of step length (3.6) defines the conjugate gradient. The last formula (which is just the Gram-Schmidt procedure) appears to be quite expensive to implement and to involve a lot of storage, but fortunately we shall see that it may be greatly simplified.

Lemma 3.8. (1) $W_i = \operatorname{span}[s_0, \dots, s_{i-1}] = \operatorname{span}[r_0, \dots, r_{i-1}]$.

(2) *The residuals are l_2-orthogonal: $r_i^T r_j = 0$ for $i \neq j$.*

(3) *There exists $m \le n$ such that $W_1 \subsetneq W_2 \subsetneq \cdots \subsetneq W_m = W_{m+1} = \cdots$ and $u_0 \neq u_1 \neq \cdots \neq u_m = u_{m+1} = \cdots = u$.*

(4) *For $i \le m$, $\{ s_0, \dots, s_{i-1} \}$ is an A-orthogonal basis for W_i and $\{ r_0, \dots, r_{i-1} \}$ is an l_2-orthogonal basis for W_i.*

(5) *$s_i^T r_j = s_i^T r_i = r_i^T r_i$ for $0 \le j \le i$.*

Proof. The first statement comes directly from the definitions. To verify the second statement, note that, for $0 \le j < i$, $F(u_i + t r_j)$ is minimal when $t = 0$, which gives $r_j^T (A u_i - f) = 0$, which is the desired orthogonality. For the third statement, certainly there is a least integer $m \in [1, n]$ so that $W_m = W_{m+1}$. Then $r_m = 0$ since it both belongs to W_m and is orthogonal to W_m. This implies that $u_m = u$ and that $s_m = 0$. Since $s_m = 0$ $u_{m+1} = u_m = u$. Therefore $r_{m+1} = 0$, which implies that $s_{m+1} = 0$, $u_{m+2} = u$, etc.

The fourth statement is an immediate consequence of the preceding ones. For the last statement, we use the orthogonality of the residuals to see that $s_i^T r_i = r_i^T r_i$. But, if $0 \le j \le i$,then

$$s_i^T r_j - s_i^T r_0 = s_i^T A (u_0 - u_j) = 0,$$

since $u_0 - u_j \in W_i$. □

Since $s_i \in W_{i+1}$ and the r_j, $j \le i$ are an orthogonal basis for that space for $i < m$, we have

$$s_i = \sum_{j=0}^{i} \frac{s_i^T r_j}{r_j^T r_j} r_j.$$

In view of part 5 of the lemma, we can simplify

$$s_i = r_i^T r_i \sum_{j=0}^{i} \frac{r_j}{r_j^T r_j} = r_i + r_i^T r_i \sum_{j=0}^{i-1} \frac{r_j}{r_j^T r_j},$$

whence

$$s_i = r_i + \frac{r_i^T r_i}{r_{i-1}^T r_{i-1}} s_{i-1}.$$

This is the formula which is used to compute the search direction. In implementing this formula it is useful to compute the residual from the formula $r_{i+1} = r_i - \lambda_i A s_i$ (since $u_{i+1} = u_i + \lambda_i s_i$). Putting things together we obtain the following implementation of CG:

```
choose initial iterate $u_0$, set $s_0 = r_0 = f - A u_0$
for $i = 0, 1, \dots$
    $\lambda_i = \frac{r_i^T r_i}{s_i^T A s_i}$
    $u_{i+1} = u_i + \lambda_i s_i$
    $r_{i+1} = r_i - \lambda_i A s_i$
    $s_{i+1} = r_{i+1} + \frac{r_{i+1}^T r_{i+1}}{r_i^T r_i} s_i$
end
```

At each step we have to perform one multiplication of a vector by A, two dot-products, and three SAXPYs, very similar to steepest descents (one more SAXPY). Here is the algorithm written out in full in pseudocode:

```
choose initial iterate $u$
$r \leftarrow f - Au$
$r2 \leftarrow r^T r$
$s \leftarrow r$
for $i = 0, 1, \ldots$
    $t \leftarrow As$                      (matrix multiplication)
    $s2 \leftarrow s^T t$                  (dot product)
    $\lambda \leftarrow r2/s2$
    $u \leftarrow u + \lambda s$           (SAXPY)
    $r2old \leftarrow r2$
    $r \leftarrow r - \lambda t$           (SAXPY)
    $r2 \leftarrow r^T r$                  (dot product)
    $s \leftarrow r + (r2/r2old)s$         (SAXPY)
end
```

The conjugate gradient method gives the exact solution in n iterations, but it is most commonly used as an iterative method and terminated with far fewer operations. A typical stopping criterion would be to test if $r2$ is below a given tolerance. To justify this, we shall show that the method is linearly convergence and we shall establish the rate of convergence. For analytical purposes, it is most convenient to use the vector norm $\|u\|_A := (u^T Au)^{1/2}$, and its associated matrix norm.

We start with a third characterization of $W_i = \operatorname{span}[s_0, \ldots, s_{i-1}] = \operatorname{span}[r_0, \ldots, r_{i-1}]$.

Lemma 3.9. *$W_i = \operatorname{span}[r_0, Ar_0, \ldots, A^{i-1}r_0]$ for $i = 1, 2, \ldots, m$.*

Proof. Since $\dim W_i = i$, it is enough to show that $W_i \subset \operatorname{span}[r_0, Ar_0, \ldots, A^{i-1}r_0]$, which we do by induction. This is certainly true for $i = 1$. Assume it holds for some i. Then, since $u_i \in u_0 + W_i$,

$$r_i = f - Au_i = (f - Au_0) + A(u_0 - u_i) \in r_0 + AW_i \in \operatorname{span}[r_0, Ar_0, \ldots, A^i r_0],$$

and therefore W_{i+1}, which is spanned by W_i and r_i belongs to $\operatorname{span}[r_0, Ar_0, \ldots, A^i r_0]$, which completes the induction. □

The space $\operatorname{span}[r_0, Ar_0, \ldots, A^{i-1}r_0]$ is called the *Krylov space* generated by the matrix A and the vector r_0. Note that we have as well

(3.7)
$$W_i = \operatorname{span}[r_0, Ar_0, \ldots, A^{i-1}r_0] = \{\, p(A)r_0 \,|\, p \in \mathcal{P}_{i-1} \,\} = \{\, q(A)(u - u_0) \,|\, q \in \mathcal{P}_i, q(0) = 0 \,\}.$$

Here $\mathcal{P}_i$ denotes the space of polynomials of degree at most i. Since r_i is l_2-orthogonal to W_i, $u - u_i$ is A-orthogonal to W_i. Now $u - u_i \in u - u_0 + W_i$ and for any $w \in W_i$,

$$u - u_0 + w = (u - u_i) + (u_i - u_0 + w),$$

where the first term on the right hand side is A-orthogonal to W_i and the second term belongs to W_i. By the Pythagorean theorem $\|u - u_0 + w\|_A \ge \|u - u_i\|_A$. Thus

$$\|u - u_i\|_A = \inf_{w\in W_i} \|u - u_0 + w\|_A.$$

Combining with the characterization of W_i in (3.7), we get

$$\|u - u_i\|_A = \inf_{\substack{q\in\mathcal{P}_i\\ q(0)=0}} \|u - u_0 + q(A)(u - u_0)\|_A = \inf_{\substack{p\in\mathcal{P}_i\\ p(0)=1}} \|p(A)(u - u_0)\|_A.$$

Applying the obvious bound $\|p(A)(u-u_0)\|_A \le \|p(A)\|_A\|u-u_0\|_A$ we see that we can obtain an error estimate for the conjugate gradient method by estimating

$$K = \inf_{\substack{p\in\mathcal{P}_i\\ p(0)=1}} \|p(A)\|_A.$$

Now if $0 < \rho_1 < \cdots < \rho_n$ are the eigenvalues of A, then the eigenvalues of $p(A)$ are $p(\rho_j)$, $j = 1, \ldots, n$, and $\|p(A)\|_A = \max_j |p(\rho_j)|$. Thus[1]

$$K = \inf_{\substack{p\in\mathcal{P}_i\\ p(0)=1}} \max_j |p(\rho_j)| \le \inf_{\substack{p\in\mathcal{P}_i\\ p(0)=1}} \max_{\rho_1\le\rho\le\rho_n} |p(\rho)|.$$

The final infimum is a property of polynomials of one variable, and it can be calculated explicitly, as will be explained below. Namely, for any $0 < a < b$, and integer $n > 0$,

$$\min_{\substack{p\in\mathcal{P}_n\\ p(0)=1}} \max_{x\in[a,b]} |p(x)| = \frac{2}{\left(\frac{\sqrt{b/a}+1}{\sqrt{b/a}-1}\right)^n + \left(\frac{\sqrt{b/a}-1}{\sqrt{b/a}+1}\right)^n}. \tag{3.8}$$

This gives

$$K \le \frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^i + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i} \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i,$$

where $\kappa = \rho_n/\rho_1$ is the condition number of A. (To get the right-hand side, we suppressed the second term in the denominator of the left-hand side, which is less than 1 and tends to zero with i, and kept only the first term, which is greater than 1 and tends to infinity with i.) We have thus proven that

$$\|u - u_i\|_A \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i \|u - u_0\|_A,$$

which is linear convergence with rate

$$r = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

Note that $r \sim 1 - 2/\sqrt{\kappa}$ for large κ. So the convergence deteriorates when the condition number is large. However, this is still a notable improvement over the classical iterations.

[1] Here we bound $\max_j |p(\rho_j)|$ by $\max_{\rho_1\le\rho\le\rho_n} |p(\rho)|$ simply because we can minimize the latter quantity explicitly. However this does not necessarily lead to the best possible estimate, and the conjugate gradient method is often observed to converge faster than the result derived here. Better bounds can sometimes be obtained by taking into account the distribution of the spectrum of A, rather than just its minimum and maximum.

For the discrete Laplacian, where $\kappa = O(h^{-2})$, the convergence rate is bounded by $1 - ch$, not $1 - ch^2$.

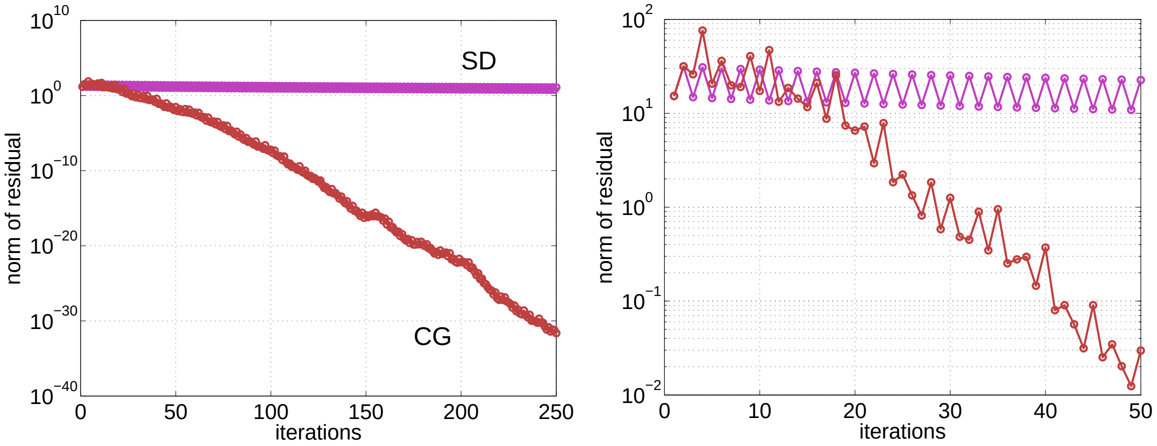
The above analysis yields a convergence estimate for the method of steepest descent as well. Indeed, the first step of conjugate gradients coincides with steepest descents, and so, for steepest descents,

$$\|u - u_1\|_A \le \frac{2}{\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}} \|u - u_0\|_A = \frac{\kappa - 1}{\kappa + 1} \|u - u_0\|_A.$$

Of course, the same result holds if we replace u_0 by u_i and u_1 by u_{i+1}. Thus steepest descents converges linearly, with rate $(\kappa - 1)/(\kappa + 1)$ (just like Richardson iteration with the optimal parameter). Notice that the estimates indicate that a large value of κ will slow the convergence of both steepest descents and conjugate gradients, but, since the dependence for conjugate gradients is on $\sqrt{\kappa}$ rather than κ, the convergence of conjugate gradients will usually be much faster.

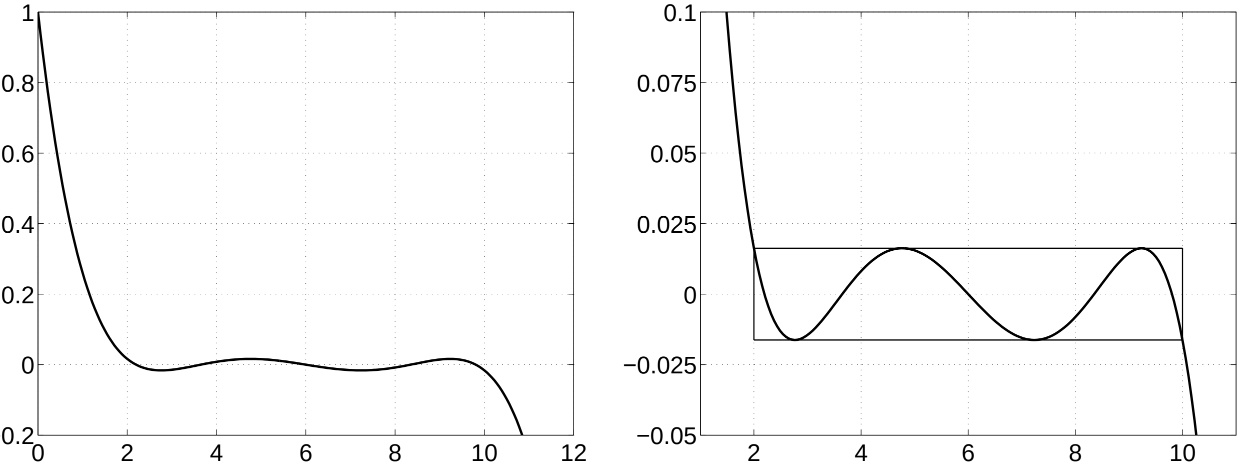
The figure shows a plot of the norm of the residual versus the number of iterations for the conjugate gradient method and the method of steepest descents applied to a matrix of size 233 arising from a finite element simulation. The matrix is irregular, but sparse (averaging about 6 nonzero elements per row), and has a condition number of about 1,400. A logarithmic scale is used on the y-axis so the near linearity of the graph reflects linear convergence behavior. For conjugate gradients, the observed rate of linear convergence is between .7 and .8, and it takes 80 iterations to reduce the initial residual by a factor of about 10^6. The convergence of steepest descents is too slow to be useful: in 400 iterations the residual is not even reduced by a factor of 2.

Figure 3.4. Convergence of conjugate gradients for solving a finite element system of size 233. On the left 300 iterations are shown, on the right the first 50. Steepest descents is shown for comparison.



Remark. There are a variety of conjugate-gradient-like iterative methods that apply to matrix problems $Au = f$ where A is either indefinite, non-symmetric, or both. Many share the idea of approximation of the solution in a Krylov space.

FIGURE 3.5. The quintic polynomial equal to 1 at 0 with the smallest L^∞ norm on $[2, 10]$. This is a scaled Chebyshev polynomial, and so the norm can be computed exactly.



Our analysis of conjugate gradients and steepest descents depended on the explicit solution of the minimization problem given in (3.8). Here we outline the proof of this result, leaving the details as an exercise.

The Chebyshev polynomials are defined by the recursion

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \text{ for } n = 1, 2, \ldots,$$

so T_n is a polynomial of degree n. From this follows two explicit formulas for T_n:

$$T_n(x) = \cos(n \arccos x), \quad T_n(x) = \frac{1}{2}[(x + \sqrt{x^2-1})^n + (x - \sqrt{x^2-1})^n],$$

with the first equation valid for $|x| \le 1$ and the second valid for $|x| \ge 1$.

The polynomial T_n satisfies $|T_n(x)| \le 1$ on $[-1, 1]$ with equality holding for $n+1$ distinct numbers in $[-1, 1]$. This can be used to establish the following: for any $\alpha < -1$, there does not exist any polynomial $q \in \mathcal{P}_n$ with $q(\alpha) = T_n(\alpha)$ and $|q(x)| < 1$ on $[-1, 1]$. In other words, T_n minimizes of $\max_{x\in[-1,1]} |p(x)|$ over all polynomials in $\mathcal{P}_n$ which take the value $T_n(\alpha)$ at α.

Scaling this result we find that

$$p(x) = \left[T_n\left(-\frac{b+a}{b-a}\right)\right]^{-1} T_n\left(\frac{2x-b-a}{b-a}\right)$$

solves the minimization problem (3.8) and gives the minimum value claimed. This polynomial is plotted for $n = 5$, $a = 2$, $b = 10$ in Figure 3.5.

2.3. Preconditioning. The idea is we choose an SPD matrix $M \approx A$ such that the system $Mz = c$ is relatively easy to solve. We then consider the *preconditioned system* $M^{-1}Ax = M^{-1}b$. The new matrix $M^{-1}A$ is SPD with respect to the M inner product, and we solve the preconditioned system using conjugate gradients but using the M-inner product in place of the l_2-inner product. Thus to obtain the preconditioned conjugate gradient algorithm, or PCG, we replace A and f by $M^{-1}A$ and $M^{-1}f$ everywhere and change expressions of the form $x^T y$ into $x^T My$. Note that the A-inner product $x^T Ay$ remains invariant under these two changes. Thus we obtain the algorithm:

```
choose initial iterate $u_0$, set $s_0 = \bar{r}_0 = M^{-1}f - M^{-1}Au_0$
**for** $i = 0, 1, \dots$
    $\lambda_i = \dfrac{\bar{r}_i^T M\bar{r}_i}{s_i^T As_i}$
    $u_{i+1} = u_i + \lambda_i s_i$
    $\bar{r}_{i+1} = \bar{r}_i - \lambda_i M^{-1}As_i$
    $s_{i+1} = \bar{r}_{i+1} + \dfrac{\bar{r}_{i+1}^T M\bar{r}_{i+1}}{\bar{r}_i^T M\bar{r}_i} s_i$
**end**
```

Note that term $s_i^T As_i$ arises as the M-inner product of s_i with $M^{-1}As_i$. The quantity $\bar{r}_i$ is the residual in the preconditioned equation, which is related to the regular residual, $r_i = f - Au_i$ by $r_i = M\bar{r}_i$. Writing PCG in terms of r_i rather than $\bar{r}_i$ we get

```
choose initial iterate $u_0$, set $r_0 = f - Au_0$, $s_0 = M^{-1}r_0$
**for** $i = 0, 1, \dots$
    $\lambda_i = \dfrac{r_i^T M^{-1}r_i}{s_i^T As_i}$
    $u_{i+1} = u_i + \lambda_i s_i$
    $r_{i+1} = r_i - \lambda_i As_i$
    $s_{i+1} = M^{-1}r_{i+1} + \dfrac{r_{i+1}^T M^{-1}r_{i+1}}{r_i^T M^{-1}r_i} s_i$
**end**
```

Thus we need to compute $M^{-1}r_i$ at each iteration. Otherwise the work is essentially the same as for ordinary conjugate gradients. Since the algorithm is just conjugate gradients for the preconditioned equation we immediately have an error estimate:

$$\|u_i - u\|_A \le 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^i \|u_0 - u\|_A,$$

where κ now is the ratio of the largest to the least eigenvalue of $M^{-1}A$. To the extent that M approximates A, this ratio will be close to 1 and so the algorithm will converge quickly.

The matrix M is called the *preconditioner*. A good preconditioner should have two properties. First, it must be substantially easier to solve systems with the matrix M than with the original matrix A, since we will have to solve such a system at each step of the preconditioned conjugate gradient algorithm. Second, the matrix $M^{-1}A$ should be substantially better conditioned than A, so that PCG converges faster than ordinary CG. In short, M should be near A, but much easier to invert. Note that these conditions are similar to those we look for in defining a classical iteration via residual correction. If $u_{i+1} = u_i + B(f - Au_i)$ is an iterative method for which B is SPD, then we might use $M = B^{-1}$ as a preconditioner.

For example, the Jacobi method suggests taking M to be the diagonal matrix with the same diagonal entries as A. When we compute $M^{-1}r_i$ in the preconditioned conjugate gradient algorithm, we are simply applying one Jacobi iteration. Similarly we could use symmetric Gauss-Seidel to get a preconditioner.

In fact, we can show that conjugate gradients preconditioned by some SPD approximate inverse always converges faster than the corresponding classical iterative method. For if λ is an eigenvalue of BA, then $-\rho \le 1-\lambda \le \rho$ where ρ is the spectral radius of $I - BA$, and so

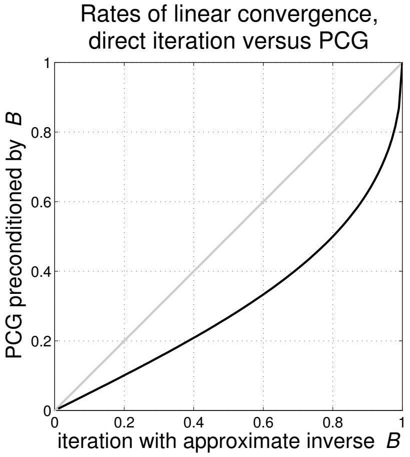
$$\lambda_{\min}(BA) \ge 1-\rho, \quad \lambda_{\max}(BA) \le 1+\rho, \quad \kappa(BA) \le \frac{1+\rho}{1-\rho}.$$

Thus the rate of convergence for the PCG method is at most

$$\frac{\sqrt{\kappa(BA)}-1}{\sqrt{\kappa(BA)}+1} \le \frac{\sqrt{\frac{1+\rho}{1-\rho}}-1}{\sqrt{\frac{1+\rho}{1-\rho}}+1} = \frac{1-\sqrt{1-\rho^2}}{\rho}.$$

The last quantity is strictly less than ρ for all $\rho \in (0,1)$; see Figure 3.6. (For ρ small it is about $\rho/2$, while for the important case of $\rho \approx 1-\epsilon$ with ϵ small, it is approximately $1-\sqrt{2\epsilon}$.) Thus the rate of convergence of PCG with B as a preconditioner is better than that of the classical iteration with B as approximate inverse.

FIGURE 3.6. If an iteration achieves a rate of linear convergence $\rho < 1$, then the rate of convergence of conjugate gradients using the iteration as a preconditioner is bounded by $(1-\sqrt{1-\rho^2})/\rho$, which is always smaller.



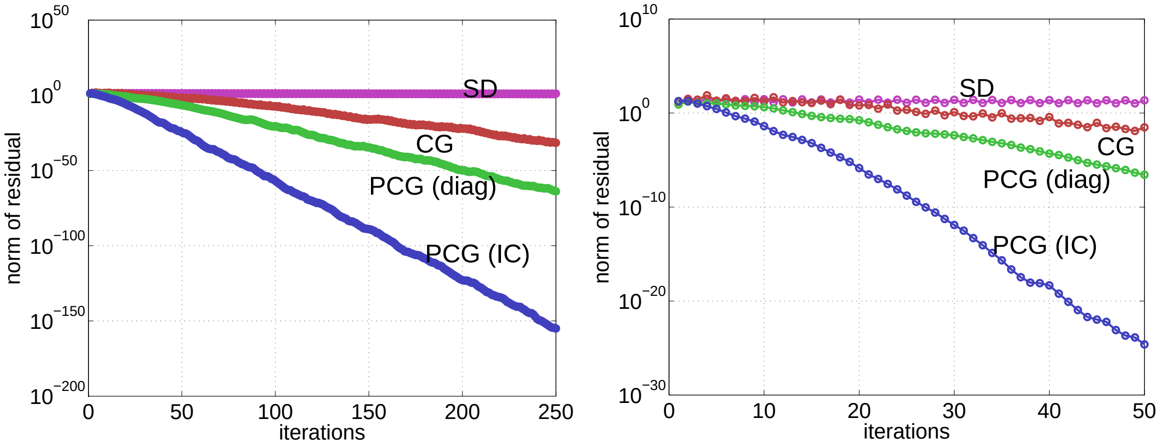
Diagonal (Jacobi) preconditioning is often inadequate (in the case of the 5-point Laplacian it accomplishes nothing, since the diagonal is constant). Symmetric Gauss-Seidel is somewhat better, but often insufficient as well. A third possibility which is often applied when A is sparse is to determine M via the *incomplete Cholesky factorization*. This means that a triangular matrix L is computed by the Cholesky algorithm applied to A, except that no fill-in is allowed: only the non-zero elements of A are altered, and the zero elements left untouched. One then takes $M = LL^T$, and, so M^{-1} is easy to apply. Yet, other preconditioners take into account the source of the matrix problem. For example, if a matrix arises from the discretization of a complex partial differential equation, we might precondition it by the discretization matrix for a simpler related differential equation (if that lead to a linear systems which is easier to solve). In fact the derivation of good preconditioners for important classes of linear systems remain a very active research area.

We close with numerical results for preconditioned conjugate gradients with both the diagonal preconditioner and incomplete Cholesky factorization as preconditioner. In Figure 3.7 we reproduce the results shown in Figure 3.4, together with these preconditioned iterations. By fitting the log of the norm of the residual to a linear polynomial, we can compute the observed rates of linear convergence. They are:

steepest descents	0.997	PCG (diag.)	0.529
conjugate gradients	0.725	PCG (IC)	0.228

The preconditioned methods are much more effective. Diagonal preconditioning reduces the number of iterations needed by conjugate gradients to reduce the initial error by a factor of 10^{-6} from 80 to 44. Incomplete Cholesky preconditioning reduces further to 18 iterations.

FIGURE 3.7. Convergence of conjugate gradients for solving a finite element system of size 233, unpreconditioned, diagonally preconditioned, and preconditioned by incomplete Cholesky factorization. Steepest descents is shown as well. On the left 300 iterations are shown, on the right the first 50.

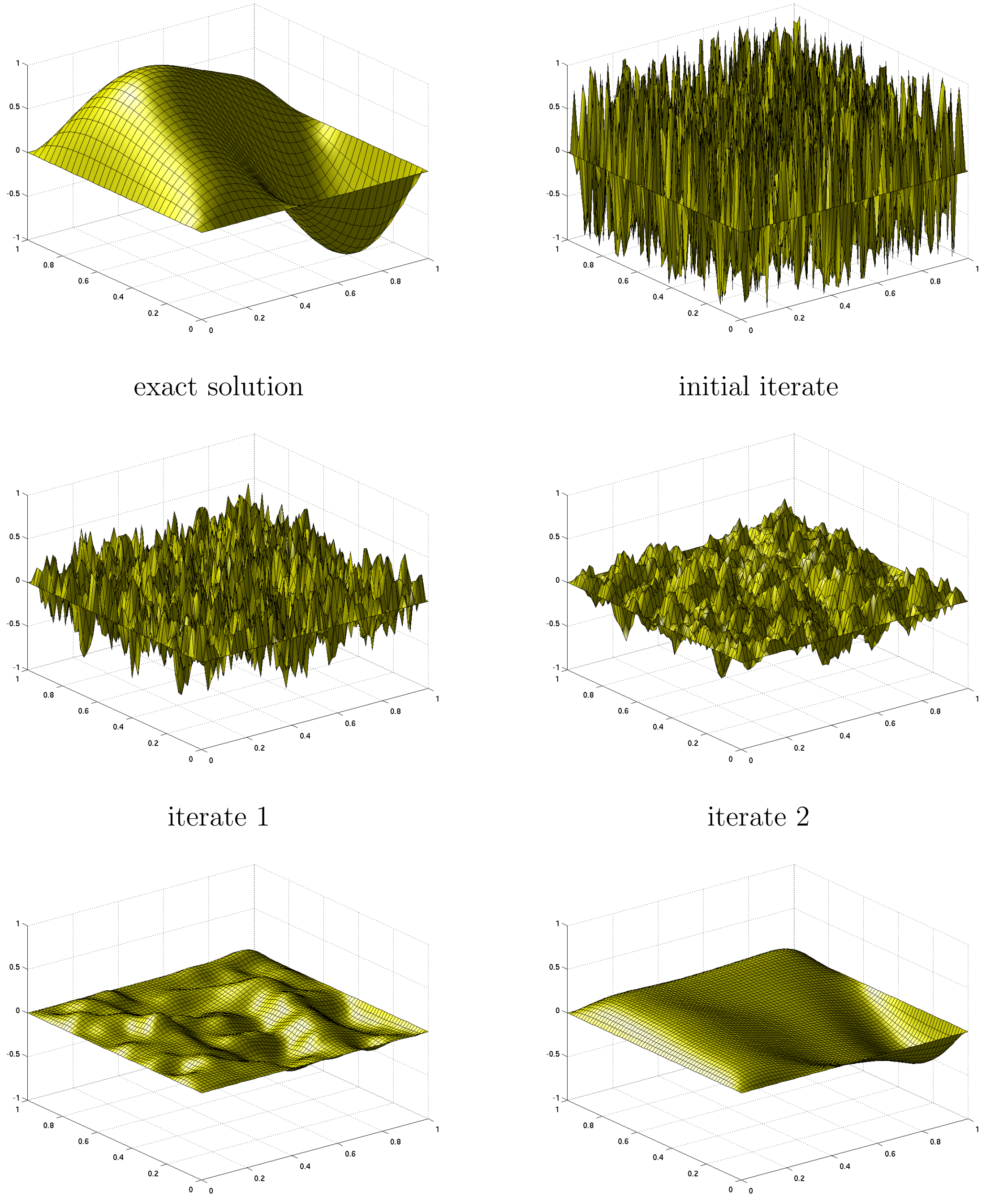


3. Multigrid methods

Figure 3.8 shows the result of solving a discrete system of the form $-\Delta_h u_h = f$ using the Gauss–Seidel iteration. We have take $h = 64$, and chosen a smooth right-hand side vector f which results in the vector u_h which is shown in the first plot. The initial iterate u_0, which is shown in the second plot, was chosen at random, and then the iterates u_1, u_2, u_{10}, u_{50}, and u_{500} are shown in the subsequent plots. In Figure 3.9, the maximum norm error $\|u_h - u_i\|/\|u_h\|$ is plotted for $i = 0, 1, \ldots, 50$.

These numerical experiments illustrate the following qualitative properties, which are typical of the Gauss–Seidel iteration applied to matrices arising from the discretization of elliptic PDEs.

FIGURE 3.8. Iterative solution to $-\Delta_h u_h = f$, $h = 1/64$, using Gauss–Seidel. The random initial iterate is rapidly smoothed, but approaches the solution u_h only very slowly.



exact solution

initial iterate

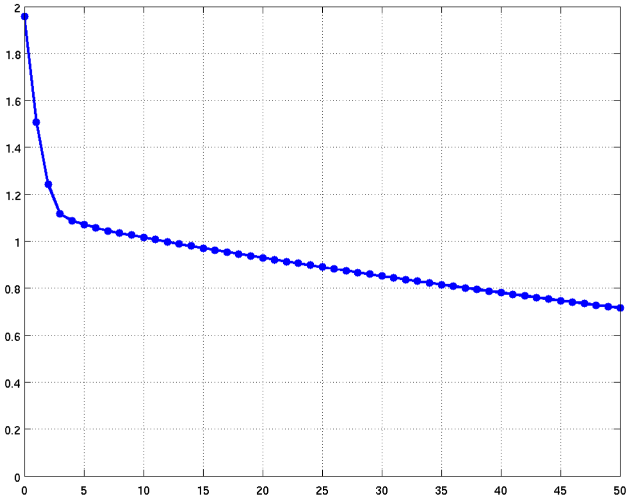
iterate 1

iterate 2

iterate 10

iterate 50

FIGURE 3.9. Error in the Gauss–Seidel iterates 0 through 50 in l^∞ (●).



- If we start with a random error, the norm of the error will be reduced fairly quickly for the first few iterations, but the error reduction occurs much more slowly after that.
- After several iterations the error is much smoother, but not much smaller, than initially. Otherwise put, the highly oscillatory modes of the error are suppressed much more quickly by the iteration than the low frequency modes.

The first observation is valid for all the methods we have studied: Richardson, Jacobi, damped Jacobi, and Gauss–Seidel. The second obervation—that Gauss–Seidel iteration *smooths* the error—is shared by the Jacobi method with a damping parameter $\alpha < 1$, but not by the standard undampd Jacobi method.

If we take the Richardson method with $\omega = 1/\lambda_{\max}(A)$ for the operator $A = -D_h^2$, it is very easy to see how the smoothing property comes about. The initial error can be expanded in terms of the eigenfunctions of A: $e_0 = \sum_{m=1}^n c_i \sin m\pi x$. The mth component in this expansion is multiplied by $1 - \lambda_m/\lambda_{\max} = 1 - \lambda_m/\lambda_n$ at each iteration. Thus the high frequency components, $m \approx n$, are multiplied by something near to 0 at each iteration, and so are damped very quickly. Even the intermediate eigenvalues, $\lambda_m \approx \lambda_n/2$ are damped reasonably quickly (by a factor of about 1/2 at each iteration). But the low frequency modes, for which $\lambda_m \ll \lambda_n$, decrease very slowly.

This also explains the first observation, that the norm of the error decreases quickly at first, and then more slowly. The norm of the error has contributions from all modes present in the initial error. Those associated to the higher frequency modes disappear in a few iterations, bringing the error down by a significant fraction. But after that the error is dominated by the low frequency modes, and so decays very slowly.

The same analysis applies to damped Jacobi with positive damping, and shows that undamped Jacobi doesn't have the smoothing property: the mth mode is multiplied by about $1 - 2\lambda_m/\lambda_n$, and so convergence is very slow for low frequency modes and also the highest frequency modes $\lambda_m \approx \lambda_n$. For the intermediate modes, $\lambda_m \approx \lambda_n/2$, convergence is very fast.

Establishing the smoothing property for Gauss–Seidel is more complicated, since the eigenfunctions of the Gauss–Seidel iteration don't coincide with those of A even for $A = -D_h^2$. However both numerical study and careful analysis show that Gauss–Seidel does indeed have the smoothing property for discretized elliptic operators.

The idea behind the multigrid method is to create an iterative method which reduces all components of the residual quickly by putting together two steps. First it applies the approximate inverse from Gauss–Seidel or another classical iterative method with the smoothing property to the residual. This greatly reduces the high frequency components of the residual, but barely reduces the low frequency components. The new residual, being relatively smooth, can then be accurately approximated on a coarser mesh. So, for the second step, the residual is (somehow) transferred to a coarser mesh, and the equation solved there, thus reducing the low frequency components. On the coarser mesh, it is of course less expensive to solve. For simplicity, we assume for now that an exact solver is used on the coarse mesh. Finally this coarse mesh solution to the residual problem is somehow transferred back to the fine mesh where it can be added back to our smoothed approximation.

Thus we have motivated the following rough outline of an algorithm:

(1) Starting from an initial guess u_0 apply a fine mesh smoothing iteration to get an improved approximation $\bar{u}$.
(2) Transfer the residual in $\bar{u}$ to a coarser mesh, solve a coarse mesh version of the problem there, transfer the solution back to the fine mesh, and add it back to $\bar{u}$ to get $\bar{\bar{u}}$.

Taking $\bar{\bar{u}}$ for u_1 and thus have described an iteration to get from u_0 to u_1 (which we can then apply again to get from u_1 to u_2, and so on). In fact it is much more common to also apply a fine mesh smoothing at the end of the iteration, i.e., to add a third step:

(3) Starting from $\bar{\bar{u}}$ apply the smoothing iteration to get an improved approximation $\bar{\bar{\bar{u}}}$.

The point of including the third step is that it leads to a multigrid iteration which is symmetric, which is often advantageous (e.g., the iteration can be used as a preconditioner for conjugate gradients). If the approximation inverse B used for the first smoothing step is not symmetric, we need to apply B^T (which is also an approximate inverse, since A is symmetric) to obtain a symmetric iteration.

We have just described a *two-grid* iteration. The true multigrid method will involve not just the original mesh and one coarser mesh, but a whole sequence of meshes. However, once we understand the two-grid iteration, the multigrid iteration will follow easily.

To make the two-grid method more precise we need to explain step 2 more fully, namely (a) how do we transfer the residual from the fine mesh to the coarse mesh?; (b) what problem do we solve on the coarse mesh?; and (c) how do we transfer the solution of that problem from the coarse mesh to the fine mesh? For simplicity, we suppose that $N = 1/h$ is even and that we are interested in solving $A_h u = f$ where $A = -D_h^2$. Let $H = 2h = (N/2)^{-1}$. We will use the mesh of size H as our coarse mesh. The first step of our multigrid iteration is then just

$$\bar{u} = u_0 + B_h(f - A_h u_0),$$

where B_h is just the approximate inverse of A_h from Gauss–Seidel or some other smoothing iteration. The resulting residual is $f - A_h\bar{u}$. This is a function on the fine mesh points $h, 2h, \dots, (N-1)h$, and a natural way to transfer it to the coarse mesh is restrict it to the even grid points $2h, 4h, \dots, (N-2)h = H, 2H, \dots, (N/2-1)H$, which are exactly the coarse mesh grid points. Denoting this *restriction operator* from fine grid to coarse grid functions (i.e., from $\mathbb{R}^{N-1} \to \mathbb{R}^{N/2-1}$) by P_H, we then solve $A_H e_H = P_H(f - A_h\bar{u}_h)$ where, of course, $A_H = -D_H^2$ is the 3-point difference operator on the coarse mesh. To transfer the solution e_H, a coarse grid function, to the fine grid, we need a *prolongation operator* $Q_H : \mathbb{R}^{N/2-1} \to \mathbb{R}^{N-1}$. It is natural to set $Q_H e_H(jh) = e_H(jh)$ if j is even. But what about when j is odd: how should we define $Q_H e_H$ at the midpoint of two adjacent coarse mesh points? A natural choice, which is simple to implement, is $Q_H e_H(jh) = [e_H((j-1)h) + e((j+1)h)]/2$. With these two operators second step is

$$\bar{\bar{u}} = \bar{u} + Q_H A_H^{-1} P_H (f - A_h\bar{u}).$$

And then final post-smoothing step is

$$\bar{\bar{\bar{u}}} = \bar{\bar{u}} + B_h^T(f - A_h\bar{\bar{u}}).$$

Actually this does not give a symmetric iteration. To obtain symmetry we need $Q_h = cP_H^T$ and that is not the case for the grid transfer operators we defined. We have

$$Q_H = \begin{pmatrix} 1/2 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & 0 & \cdots & 0 \\ 1/2 & 1/2 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1/2 & 1/2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1/2 \end{pmatrix}, \tag{3.9}$$

but P_H as we described it, consists only of 0's and 1's. Therefore one commonly takes a different choice for P_H, namely $P_H = (1/2)Q_H^T$. This means that the transferred coarse grid function doesn't just take the value of the corresponding fine grid function at the coarse grid point, but rather uses a weighted average of the fine grid function's values at the point in question and the fine grid points to the left and right (with weights 1/4, 1/2, 1/4). With this choice, $Q_H A_h P_H$ is symmetric; in fact, $Q_H A_h P_H = A_H$. This is a useful formula. For operators other than the $A_h = -D_h^2$, we can use the same intergrid transfer operators, namely Q_H given by (3.9) and $P_H = (1/2)Q_H^T$, and then define the coarse grid operator by $A_H = Q_H A_h P_H$.

Remark. In a finite element context, the situation is simpler. If the fine mesh is a refinement of the coarse mesh, then a coarse mesh function is already a fine mesh function. Therefore, the operator Q_H can be taken simply to be the inclusion operator of the coarse mesh space into the fine mesh space. The residual in $u_0 \in S_h$ is most naturally viewed as a functional on S_h: $v \mapsto (f, v) - B(u_0, v)$. It is then natural to transfer the residual to the coarse mesh simply by restricting the test function v to S_H. This operation $S_h^T \to S_H^T$ is exactly the adjoint of the inclusion operator $S_H \to S_h$. Thus the second step, solving the coarse mesh problem for the restricted residual is obvious in the finite element case: we find

$e_H \in S_H$ such that

$$B(e_H, v) = (f, v) - B(\bar{u}, v), \quad v \in S_H,$$

and then we set $\bar{\bar{u}} = \bar{u} + e_H \in S_h$.

Returning to the case of finite differences we have arrived at the following two-grid iterative method to solve $A_h u_h = f_h$.

```
$u_h = \text{twogrid}(h, A_h, f_h, u_0)$
  input:  $h$, mesh size ($h = 1/n$ with $n$ even)
          $A_h$, operator on mesh functions
          $f_h$, mesh function (right-hand side)
          $u_0$, mesh function (initial iterate)
  output: $u_h$, mesh function (approximate solution)

for $i = 0, 1, \dots$ until satisfied
  1. presmoothing: $\bar{u} = u_i + B_h(f_h - A_h u_i)$
  2. coarse grid correction:
    2.1. residual computation: $r_h = f_h - A_h \bar{u}$
    2.2. restriction: $H = 2h$, $r_H = P_H r_h$, $A_H = P_H A_h Q_H$
    2.3. coarse mesh solve: solve $A_H e_H = r_H$
    2.4. prolongation: $e_h = Q_H e_H$
    2.5. correction: $\bar{\bar{u}} = \bar{u} + e_h$
  3. postsmoothing: $u_h \leftarrow u_{i+1} = \bar{\bar{u}} + B_h^T(f_h - A_h \bar{\bar{u}})$
end
```

Algorithm 3.1: Two-grid iteration for approximately solving $A_h u_h = f_h$.

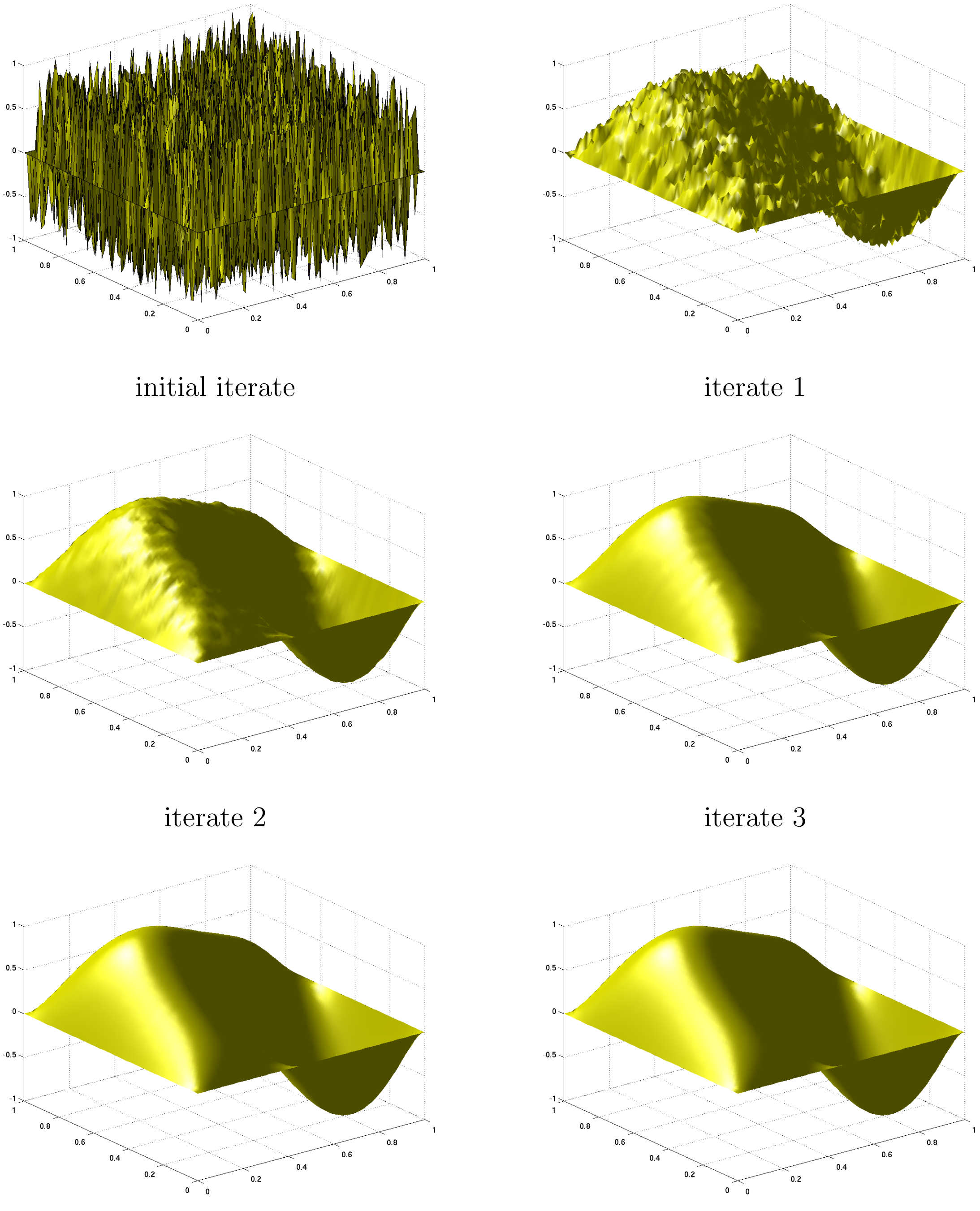
In the smoothing steps, the matrix B_h could be, for example, $(D - L)^{-1}$ where D is diagonal, L strictly lower triangular, and $A_h = D - L - L^T$. This would be a Gauss–Seidel smoother, but there are other possibilities as well. Besides these steps, the major work is in the coarse mesh solve. To obtain a more efficient algorithm, we may also solve on the coarse mesh using a two-grid iteration, and so involving an even coarser grid. In the following multigrid algorithm, we apply this idea recursively, using multigrid to solve at each mesh level, until we get to a sufficiently coarse mesh, $h = 1/2$, at which point we do an exact solve (with a 1×1 matrix!).

```
$u_h = \text{multigrid}(h, A_h, f_h, u_0)$
  *input:*  $h$, mesh size ($h = 1/n$ with $n$ a power of 2)
            $A_h$, operator on mesh functions
            $f_h$, mesh function (right-hand side)
            $u_0$, mesh function (initial iterate)
  *output*: $u_h$, mesh function (approximate solution)

if $h = 1/2$ then
  $u_h = A_h^{-1} f_h$
else
  for $i = 0, 1, \ldots$ until satisfied
    1. presmoothing: $\bar{u} = u_i + B_h(f - A_h u_i)$
    2. coarse grid correction:
      2.1. residual computation: $r_h = f_h - A_h \bar{u}$
      2.2. restriction: $H = 2h$, $r_H = P_H r_h$, $A_H = P_H A_h Q_H$
      2.3. coarse mesh solve: $e_H = \text{multigrid}(H, A_H, r_H, 0)$
      2.4. prolongation: $e_h = Q_H e_H$
      2.5. correction: $\bar{\bar{u}} = \bar{u} + e_h$
    3. postsmoothing: $u_h \leftarrow u_{i+1} = \bar{\bar{u}} + B_h^T(f - A_h \bar{\bar{u}})$
  end
end if
```

Algorithm 3.2: Multigrid iteration for approximately solving $A_h u_h = f$.

Figure 3.10 shows 5 iterations of this multigrid algorithm for solving the system $-\Delta_h u_h = f$, $h = 1/64$, considered at the beginning of this section, starting from a random initial guess (we would get even better results starting from a zero initial guess). Compare with Figure 3.8. The fast convergence of the multigrid algorithm is remarkable. Indeed, for the multigrid method discussed here it is possible to show that the iteration is linearly convergent with a rate independent of the mesh size (in this example, it is roughly 0.2). This means that the number of iterations needed to obtain a desired accuracy remains bounded independent of h. It is also easy to count the number of operations per iteration. Each iteration involves two applications of the smoothing iteration, plus computation of the residual, restriction, prolongation, and correction on the finest mesh level. All those procedures cost $O(n)$ operations. But then, during the coarse grid solve, the same procedures are applied on the grid of size $2h$, incurring an additional cost of $O(n/2)$. Via the recursion the work will be incurred for each mesh size $h, 2h, 4h, \ldots$. Thus the total work per iteration will be $O(n + n/2 + n/4 + \ldots + 1) = O(n)$ (since the geometric series sums to $2n$). Thus the total work to obtain the solution of the discrete system to any desired accuracy is itself $O(n)$, i.e., optimal.

FIGURE 3.10. Iterative solution to $-\Delta_h u_h = f$, $h = 1/64$, using multigrid.



initial iterate

iterate 1

iterate 2

iterate 3

iterate 4

iterate 5

CHAPTER 4

Finite element methods for elliptic equations

1. Weak and variational formulations

Model PDE: $-\operatorname{div} a \operatorname{grad} u + cu = f$ in Ω

Here Ω is a bounded domain in $\mathbb{R}^n$; $0 < \underline{a} \le a(x) \le \bar{a}$, $0 \le c(x) \le \bar{c}$

First consider the homogeneous Dirichlet BC: $u = 0$ on $\partial\Omega$.

Assuming that $a \in C^1(\bar{\Omega})$, $c \in C(\bar{\Omega})$, $u \in C^2(\bar{\Omega})$ satisfies the PDE and BC (a *strong solution*), then it also satisfies the *weak formulation*:

Find $u \in \mathring{H}^1(\Omega)$ such that

$$\int (a \operatorname{grad} u \cdot \operatorname{grad} v + cuv) = \int fv, \qquad v \in \mathring{H}^1(\Omega),$$

A solution of the weak formulation need not belong to $C^2(\bar{\Omega})$, but if it does, then it is a strong solution.

The *variational formulation* is completely equivalent to the weak formulation

$$u = \operatorname*{argmin}_{v \in \mathring{H}^1(\Omega)} \int_\Omega [\frac{1}{2}(a \operatorname{grad} v \cdot \operatorname{grad} v + cv^2) - fv]$$

Extensions: Neumann BC, Robin BC, mixed BC, inhomogeneous Dirichlet BC. First order term to the PDE (then the problem is not symmetric and there is no variational formulation, but weak formulation is fine).

All these problems can be put in the weak form: Find $u \in V$ such that

$$b(u, v) = F(v), \quad v \in V, \tag{4.1}$$

where V is a Hilbert space (H^1 or $\mathring{H}^1$), $b : V \times V \to \mathbb{R}$ is a bilinear form, $F : V \to \mathbb{R}$ is a linear form. (The inhomogeneous Dirichlet problem takes this form if we solve for $u - u_g$ where u_g is a function satisfying the inhomogeneous Dirichlet BC $u_g = g$ on $\partial\Omega$.) For symmetric problems (no first order term), the bilinear form b is symmetric, and the weak form is equivalent to the variational form:

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

2. Galerkin method and finite elements

Let V_h be a finite dimensional subspace of V. If we replace the V in the weak formulation with V_h we get a discrete problem: Find $u_h \in V_h$ such that

$$b(u_h, v) = F(v), \quad v \in V_h. \tag{4.2}$$

This is called the *Galerkin method.* For symmetric problems it is equivalent to the *Rayleigh–Ritz* method, which replaces V by V_h in the variational formulation:

$$u_h = \operatorname*{argmin}_{v \in V_h}[\frac{1}{2}b(v,v) - F(v)].$$

The Galerkin solution can be reduced to a set of n linear equations in n unknowns where $n = \dim V_h$ by choosing a basis. Adopting terminology from elasticity, the matrix is called the *stiffness matrix* and the right hand side is the *load vector.*

Comparing (4.1) and (4.2), we find that the error in the Galerkin method $u - u_h$ satisfies

$$b(u - u_h, v) = 0, \quad v \in V_h. \tag{4.3}$$

This relation, known as *Galerkin orthogonality*, is key to the analysis of Galerkin methods.

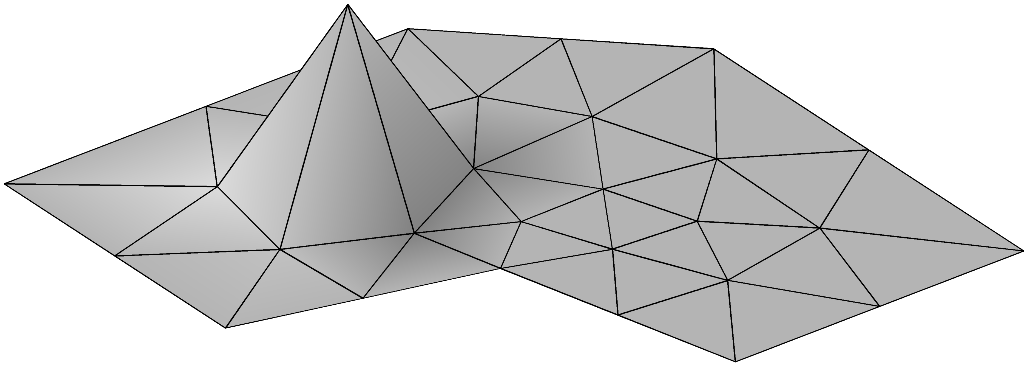
To define a simple finite element method, we suppose that Ω is a polygon in $\mathbb{R}^2$ and let $\mathcal{T}_h$ be a simplicial decomposition of Ω (covering of $\bar{\Omega}$ by closed triangles so that the intersection of any two distinct elements of $\mathcal{T}_h$ is either empty or a common edge or vertex. Let

$$M_0^1(\mathcal{T}_h) = \{\, v \in C(\Omega) \,|\, V|_T \in \mathcal{P}_1(T) \forall T \in \mathcal{T}_h \,\} = \{\, v \in H^1(\Omega) \,|\, V|_T \in \mathcal{P}_1(T) \forall T \in \mathcal{T}_h \,\},$$

and $\mathring{M}_0^1(\mathcal{T}_h) = \mathring{H}^1(\Omega) \cap M_0^1(\mathcal{T}_h)$. The $\mathcal{P}_1$ finite element method for the Dirichlet problem is the Galerkin method with $V_h = \mathring{M}_0^1(\mathcal{T}_h)$.

We can use the Lagrange (hat function) basis for V_h to ensure that (1) the matrix is sparse, and (2) the integrals entering into the stiffness matrix and load vector are easy to compute.

FIGURE 4.1. A hat function basis element for $M_0^1(\mathcal{T}_h)$.



In the special case where Ω is the unit square, and $\mathcal{T}_h$ is obtained from a uniform $m \times m$ partition into subsquares, each bissected by its SW-NE diagonal (so $n = (m-1)^2$), the resulting stiffness matrix is exactly the same as the matrix of the 5-point Laplacian.

3. Lagrange finite elements

This section is written mostly for 2D, although extending to n dimensions is straightforward.

A finite element space is a space of piecewise polynomials with respect to a given triangulation (simplicial decomposition) $\mathcal{T}_h$, but not just any space of piecewise polynomials. It is constructed by specifying the following things for each $T \in \mathcal{T}_h$:

- *Shape functions*: a finite dimensional space $V(T)$ consisting of polynomial functions on T.
- *Degrees of freedom*: a finite set of linear functionals $V(T) \to \mathbb{R}$ which are *unisolvent* on $V(T)$. This means that real values can be assigned arbitrarily to each DOF, and these determine one and only one element of $V(T)$. In other words, the DOF form a basis for the dual space of $V(T)$.

We further assume that each degree of freedom on T is associated to a subsimplex of T, i.e., to a vertex, an edge, or T itself (in 2D). Moreover, if a subsimplex is shared by two different triangles in T_1 and T_2 in $\mathcal{T}_h$, the DOFs for T_1 and T_2 associated to the subsimplex are in 1-to-1 correspondence.

When all this is specified, the *assembled finite element space* is defined as all functions $v \in L^2(\Omega)$ such that

- $v|_T \in V(T)$ for all $T \in \mathcal{T}_h$
- The DOFs are single-valued in the sense that whenever q is a subsimplex shared by T_1 and T_2, then the corresponding DOFs on applied to $v|_{T_1}$ and $v|_{T_2}$ take on the same value.

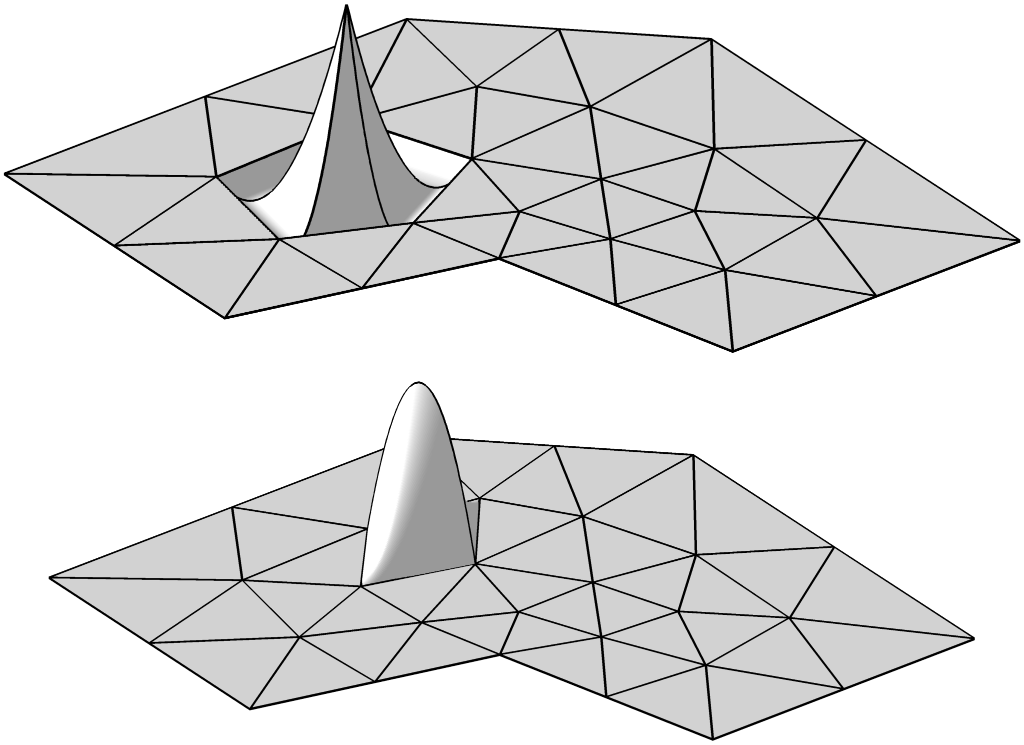
Note that we do not specify the interelement continuity explicitly. It is determined by the fact that the shared DOFs are single-valued.

The reason for this definition is that it is easy to construct and compute with piecewise polynomial spaces defined in this way. First of all, we immediately obtain a set of *global degrees of freedom*, by considering all the degrees of freedom associated with all the subsimplices of the triangulation. An element of the FE space is uniquely determined by an arbitrary assignment of values to the global degrees of freedom. Thus the dimension of the FE space is the sum over the subsimplices of the number of degrees of freedom associated to the subsimplex. A basis for the FE space is obtained by setting one of the global DOFs to 1 and all the rest to zero. The resulting basis function is supported in the union of the triangles which contain the subsimplex. Thus we have a *local basis* (small supports), and will obtain a sparse stiffness matrix.

The simplest example is the $\mathcal{P}_1$ element, or Lagrange element of degree 1, discussed above. Then the shape functions are simply the linear polynomials: $V(T) = \mathcal{P}_1(T)$ (dimension equals 3 is 2D). The degrees of freedom on T are the evaluation functionals associated to the 3 vertices. These DOFs are certainly unisolvent: a linear function in 2D is determined by its value at any 3 non-colinear points. Clearly any continuous piecewise linear function belongs to the FE space, since it can be specified by assigning its vertex values. Conversely, if v is an element of the assembed FE space and two triangles T_1 and T_2 share a common edge e, then $v|_{T_1}$ and $v|_{T_2}$ must agree on all of e, since on e they are both linear functions, and they agree at the two end points of e (a linear function in 1D is determined by its value at any 2 distinct points). This shows that the assembled FE space consists precisely of the continuous piecewise linears. The global degrees of freedom are the vertex values, and the corresponding local basis consists of the hat functions.

For the Lagrange element of degree 2, the shape functions are $V(T) = \mathcal{P}_2(T)$. There is one DOF associated to each vertex (evaluation at the vertex), and one associated to each edge (evaluation at the midpoint of the edge). Let us check unisolvence. Since $\dim V(T) = 6$ and there are 6 DOFs on T, we must show that if all 6 DOFs vanish for some $v \in V(T)$, then $v \equiv 0$. For this choose a numbering of the vertices of T, and let $\lambda_i \in \mathcal{P}_1(\mathbb{R}^2)$, $i = 1, 2, 3$, denote the linear function that is 1 on the ith vertex and which vanishes on the opposite edge, e of T. Thus the equation $\lambda_i = 0$ is the equation of the line through the edge e. Since v vanishes at the two endpoints and the midpoint of e, $v|_e$ vanishes (a quadratic in 1D which vanishes at 3 points is zero). Therefore v is divisible by the linear polynomial λ_i, for each i. Thus v is a multiple of $\lambda_1\lambda_2\lambda_3$. But v is quadratic and this product is cubic, so the only possibility is that $v \equiv 0$. It is easy to check that the assembled space is exactly the space $M_0^2(\mathcal{T}_h)$ of continuous piecewise quadratic functions. There is one basis function associated to each vertex and one to each edge.

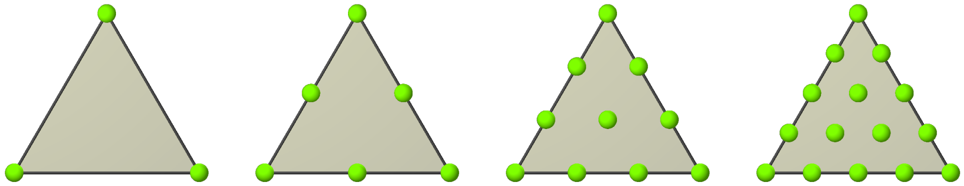
Figure 4.2. Basis functions for $M_0^2(\mathcal{T}_h)$.



Note: the linear functions λ_i are the barycentric coordinate functions on T. They satisfy $\lambda_1 + \lambda_2 + \lambda_3 \equiv 1$.

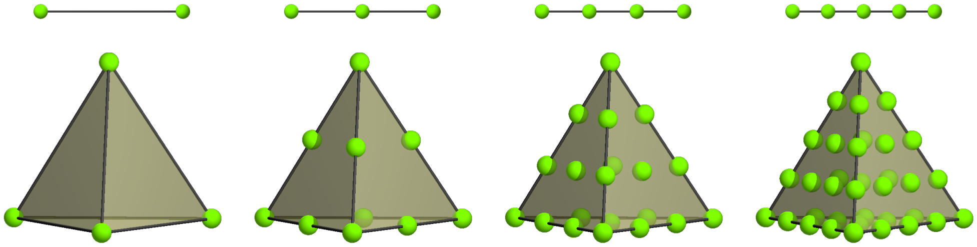
Higher degree Lagrange elements are defined similarly. $V(T) = \mathcal{P}_r(T)$. $\dim V(T) = (r+1)(r+2)/2$. There is 1 DOF at each vertex, $r-1$ on each edge, and $(r-2)(r-1)/2$ in each triangle. Note that $3 \times 1 + 3 \times (r-1) + (r-2)(r-1)/2 = (r+1)(r+2)/2$. The DOFs are the evaluation functionals at the points with barycentric coordinates all of the form i/r with $0 \le i \le r$ and integer. See Figure 4.3.

FIGURE 4.3. Lagrange finite elements of degree 1, 2, and 3.



Lagrange elements can be defined in a similar way in n-dimensions.

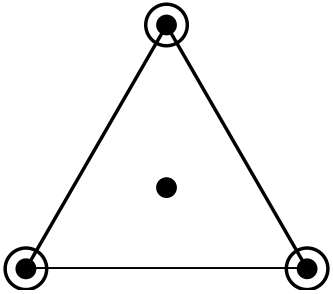
FIGURE 4.4. Lagrange finite elements of degree 1, 2, and 3 in 1-D and 3-D.



The restriction of a Lagrange element of degree r to a face or an edge is a Lagrange element of degree r on the face or edge. This leads to an inductive proof of unisolvence valid for all dimensions.

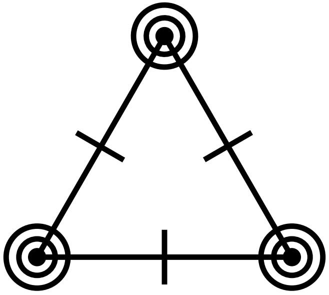
Other finite element spaces. Cubic Hermite finite elements. Shape functions are $\mathcal{P}_3(T)$ for a triangle T. Three DOFs for each vertex v: $u \mapsto u(v)$, $u \mapsto (\partial u/\partial x)(v)$, $u \mapsto (\partial u/\partial y)(v)$; and one DOF associated to the interior $u \mapsto \int_T v$ (alternatively, evaluation at the barycenter). Proof of unisolvence. Note that the assembled finite element space belongs to C^0 and H^1, but not to C^1 and H^2.

FIGURE 4.5. Cubic Hermite element.



Quintic Hermite finite elements (Argyris elements). Shape functions are $\mathcal{P}_5(T)$ for a triangle T. Six DOFs for each vertex v: evaluation of u, both 1st partials, and all three 2nd partials of u at v. One DOF for each edge: evaluation of $\partial u/\partial n$ at midpoint. Unisolvent, H^2.

FIGURE 4.6. Quintic Hermite element.



4. Finite element assembly

Now we consider how to efficiently compute the stiffness matrix in a finite element computation. We consider the Neumann problem

$$-\operatorname{div} a \operatorname{grad} u + b \cdot \operatorname{grad} u\, v + cuv = f \text{ in } \Omega, \quad a\frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega.$$

so the finite element method seeks $u_h \in V_h$ such that

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

where

$$b(u,v) := \int_\Omega (a \operatorname{grad} u \cdot \operatorname{grad} v + b \cdot \operatorname{grad} uv + cuv)\, dx, \quad F(v) = \int_\Omega fv\, dx.$$

For simplicity suppose we use $V_h = M_0^1(\mathcal{T}_h)$, the Lagrange finite element space of degree 1 for some triangulation in 2-D. Let $p_1, \ldots, p_{n_{\text{vert}}}$ denote the vertices of the triangulation $\mathcal{T}_h$ (where n_{vert} is the number of vertices, which is equal to $\operatorname{div} V_h$), and let $\phi_1, \ldots, \phi_{n_{\text{vert}}}$ denote the corresponding hat function basis functions (such as the one pictured in Figure 4.1). The stiffness matrix, which we need to compute, is given by

$$A_{ij} = b(\phi_j, \phi_i), \quad i, j = 1, \ldots, n_{\text{vert}}.$$

This might suggest as an algorithm

```
for $i = 1, \ldots, n_{\text{vert}}$
  for $j = 1, \ldots, n_{\text{vert}}$
    compute $A_{ij} = b(\phi_j, \phi_i)$
  end
end
```

but, in fact, such an algorithm is very inefficient and *should never be used.*

To define an efficient algorithm, we introduce the local vertices, basis functions, and stiffness matrix. For each triangle T, let p_k^T, $k = 1, 2, 3$ denote the three vertices of T, and let ϕ_k^T denote the three local basis functions on T. The local stiffness matrix on T is the 3×3 matrix

$$A_{ij}^T = b^T(\phi_j^T, \phi_i^T), \quad i, j = 1, 2, 3,$$

where

$$b^T(v,w) = \int_T (a \operatorname{grad} v \cdot \operatorname{grad} w + b \cdot \operatorname{grad} v w + cvw)\,dx.$$

Note that the integral is only over the triangle T. We need to compute the quantity $A^T_{ij} = b^T(\phi^T_j, \phi^T_i)$ in order to compute the (global) stiffness matrix. It will be part of exactly one element of the stiffness matrix. Fortunately, this quantity is easily computable. The three functions ϕ^T_j and their gradients can be easily expressed analytically in terms of the coordinates of the vertices p^T_1, p^T_2, p^T_3 of T. Therefore, if the coefficients a, b, and c are constant on the element T, it is straightforward to give an arithmetic expression for A^T_{ij}. If the coefficents are variable, one commonly evaluates them at one or a few quadrature points in T and computes $b^T(\phi^T_j, \phi^T_i)$ through a quadrature rule.

To relate the local quantities and the global quantities, we define I^T_k as the global vertex number of the kth vertex of T:

$$p_{I^T_k} = p^T_k, \quad k = 1, 2, 3.$$

The values of I^T_k can be stored in an integer table with one row for each triangle and 3 columns. This the (j,k) entry of the table is the global vertex number of the kth vertex of the jth triangle of the mesh. This is called the *connectivity table* of the mesh.

The finite element assembly algorithm to compute the stiffness matrix organizes the computation as a loop over the elements, in each element we: (1) compute the local stiffness matrix, (2) add the resulting elements into the appropriate elements on the global stiffness matrix.

```
Initialize $A$ to 0
**for** $T \in \mathcal{T}_h$
  compute $A^T_{ij} = b^T(\phi^T_j, \phi^T_i), \quad i,j = 1,2,3$
  $A_{I^T_i I^T_j} = A_{I^T_i I^T_j} + A^T_{ij}, \quad i,j = 1,2,3$
**end**
```

This is how a finite element stiffness matrix is computed in practice. Note that the computation is organized as a single loop over elements, rather than as a double loop over vertices.

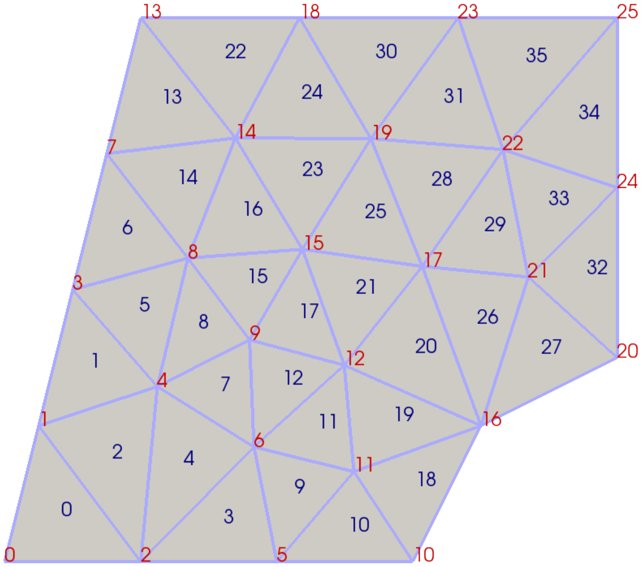
Thus to compute the stiffness matrix, we need two tables which describe the mesh: a real table of size $n_{\text{vert}} \times 2$ which gives the coordinates of the vertices, and the integer connectivity table of size $n_{\text{elt}} \times 3$. These tables are created when the mesh is generated. Figure 4.7 shows a mesh of $n_{\text{vert}} = 26$ triangles (numbered, in red, from 0 to 25 rather than 1 to 26), and $n_{\text{elt}} = 36$ elements (numbered in blue from 0 to 35), and the corresponding mesh data tables.

5. Coercivity, inf-sup condition, and well-posedness

First we consider this in an abstract framework, in which we are given a Hilbert space V, a bounded bilinear form $b : V \times V \to \mathbb{R}$ and a bounded linear form $F : V \to \mathbb{R}$, and the problem we wish to solve is

$$b(u,v) = F(v), \quad v \in V. \tag{4.4}$$

FIGURE 4.7. Finite element mesh data structure.



Vertex coordinates

0	0.00000	0.00000
1	0.50000	2.00000
2	2.00000	0.00000
3	1.00000	4.00000
4	2.24443	2.57632
⋮	⋮	⋮

Connectivity

0	0	1	2
1	1	3	4
2	1	2	4
3	2	5	6
4	2	4	6
⋮	⋮	⋮	⋮

We will analyze the Galerkin method using the ideas of stability and consistency we introduced in Chapter 2, § 1.2. Recall that there we studied a problem of the form find $u \in X$ such that $Lu = f$ where L mapped the vector space X where the solution was sought to the vector space Y where the data f belonged. For the problem (4.4), the solution space $X = V$, and the data space $Y = V^*$, the dual space of V. The linear operator L is simply given by

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

So the problem (4.4) is simply: given $F \in V^*$ find $u \in V$ such that $Lu = F$.

First of all, before turning to discretization, we consider the well-posedness of this problem. We have already assumed that the bilinear form b is bounded, i.e., there exists a constant M such that

$$|b(w, v)| \le M\|w\|\|v\|, \quad w, v \in V,$$

where, of course, the norm is the V norm. This implies that the operator $L : V \to V^*$ is a bounded linear operator (with the same bound).

We will now consider hypotheses on the bilinear form b that ensure that the problem (4.4) is well-posed. We will consider three cases, in increasing generality.

5.1. The symmetric coercive case. First we assume that the bilinear form b is symmetric ($b(w, v) = b(v, w)$) and *coercive*, which means that there exists a constant $\gamma > 0$, such that

$$b(v, v) \ge \gamma\|v\|^2, \quad v \in V.$$

In this case $b(w, v)$ is an inner product on V, and the associated norm is equivalent to the V norm:

$$\gamma\|v\|_V^2 \le b(v, v) \le M\|v\|_V^2.$$

Thus V is a Hilbert space when endowed with the b inner product, and the Riesz Representation Theorem gives that for all $F \in V^*$ there exists a unique $u \in V$ such that

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

i.e., (4.4) has a unique solution for any data, and $L^{-1} : V^* \to V$ is well-defined. Taking $v = u$ and using the coercivity condition we immediately have $\gamma \|u\|_V^2 \le F(u) \le \|F\|_{V^*}\|u\|_V$, so $\|u\|_V \le \gamma^{-1}\|F\|_V$, i.e., $\|L^{-1}\|_{\mathcal{L}(V^*,V)} \le \gamma^{-1}$.

In short, well-posedness is an immediate consequence of the Riesz Representation Theorem in the symmetric coercive case.

As a simple example of the utility of this result, let us consider the Neumann problem

$$-\operatorname{div} a \operatorname{grad} u + cu = f \text{ in } \Omega, \quad a\frac{\partial u}{\partial n} = 0 \text{ on } \partial\Omega$$

where $0 < \underline{a} \le a(x) \le \bar{a}$, $0 < \underline{c} \le c(x) \le \bar{c}$. The weak formulation is: Find $u \in V$ such that

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

where $V = H^1$,

$$b(w,v) = \int_\Omega (a \operatorname{grad} w \cdot \operatorname{grad} v + cuv), \quad F(v) = \int_\Omega fv.$$

Clearly the bilinear form b is bounded with $M = \max(\bar{a}, \bar{c})$, and is coercive with $\gamma = \min(\underline{a}, \underline{c})$. It follows that the weak formulation is well-posed. It admits a unique solution and $\|u\|_{H^1} \le \gamma^{-1}\|F\|_{(H^1)'} \le \gamma^{-1}\|f\|_{L^2}$.

5.2. The coercive case. Even if we dispense with the assumption of symmetry, coercivity implies well-posedness. From coercivity we have $\gamma\|w\|^2 \le b(w,w) = Lw(w) \le \|Lw\|_{V^*}\|w\|$, so

$$\|w\| \le \gamma^{-1}\|Lw\|_{V^*}, \quad w \in V. \tag{4.5}$$

This immediately leads to three conclusions:

- L is one-to-one.
- If L is also onto, so L^{-1} is well-defined, then $\|L^{-1}\|_{\mathcal{L}(V^*,V)} \le \gamma^{-1}$.
- The range $W = L(V)$ is closed in V^*.

The first two points are immediate. For the third, suppose that for some $u_1, u_2, \ldots \in V$, Lu_n converges to some $G \in V^*$. We must show that $G = Lu$ for some $u \in V$. Since Lu_n converges in V^* it forms a Cauchy sequence. Using (4.5) we conclude that u_n forms a Cauchy sequence in V, and so converge to some u in V. Since L is bounded $Lu_n \to Lu$ in V^*, so $Lu = G$, showing that indeed W is closed in V^*.

It remains to show that L is onto, i.e., the closed subspace $W = L(V)$ is the whole of V^*. If W were a strict closed subspace of V^* then there would exist a nonzero element $v \in V$ such that $G(v) = 0$ for all $G \in W$, i.e., $b(w,v) = Lw(v) = 0$ for all $w \in V$ and this particular v. But, taking $w = v$ and using coercivity we get a contradiction.

Thus we have shown that for a bounded coercive bilinear form, symmetric or not, the abstract weak formulation (4.4) is well-posed. This result is known as the *Lax-Milgram theorem.*

5.3. The inf-sup condition. It turns out to be very useful to consider a much more general case. Suppose that, instead of coercivity, we assume that

(1) (inf-sup condition) There exists $\gamma > 0$ such that for all $0 \neq w \in V$ there exists $0 \neq v \in V$ such that
$$b(w,v) \geq \gamma \|w\| \|v\|.$$
(2) (dense range condition) For all $0 \neq v \in V$ there exists a $w \in V$ such that $b(w,v) \neq 0$.

We shall see that it is easy to adapt the proof of the Lax-Milgram theorem to this case.

Note that the inf-sup condition can be written
$$\inf_{0\neq w\in V} \sup_{0\neq v\in V} \frac{b(w,v)}{\|w\|\|v\|} > 0,$$
which explains its name. The dense range condition is equivalent to the condition that the range $W = L(V)$ is dense in V^*. Clearly coercivity implies both these conditions (take $v = w$ for the first and $w = v$ for the second). In the symmetric case the second condition follows from the first. In any case, using these two conditions it is easy to carry out the above argument, as we do now.

The bound (4.5) follows directly from the inf-sup condition. Again this implies L is 1-to-1 and that $W = L(V)$ is *closed* in V^*, and furnishes a bound on $\|L^{-1}\|$ if L is onto. Since L has dense range, by the second condition, and closed range, it is indeed onto.

This version is in some sense the most general possible. If (4.4) is well-posed, so $L^{-1} : V^* \to V$ exists, then it is a simple matter of untangling the definitions to see that
$$\inf_{0\neq w\in V} \sup_{0\neq v\in V} \frac{b(w,v)}{\|w\|\|v\|} = \|L^{-1}\|^{-1}_{\mathcal{L}(V^*,V)},$$
and so the inf-sup condition holds with $\gamma = 1/\|L^{-1}\|^{-1}_{\mathcal{L}(V^*,V)}$. Thus the inf-sup condition and dense range condition are equivalent to well-posedness.

6. Stability, consistency, and convergence

Now we turn to discretization, again using the framework of Chapter 2, § 1.2. First we consider the coercive (but not necessarily symmetric) case. Thus we suppose again that $b : V \times V \to \mathbb{R}$ is a bounded, coercive bilinear form, with constants M and γ. Consider V_h a finite dimensional subspace of V. Restricting the bilinear form b to $V_h \times V_h$ defines an operator $L_h : V_h \to V_h^*$, and restricting F to V_h gives a linear form $F_h : V_h \to \mathbb{R}$. Galerkin's method is just $L_h u_h = F_h$. We will show that this method is consistent and stable, and so convergent.

Stability just means that L_h is invertible and the stability constant given by $\|L_h^{-1}\|_{\mathcal{L}(V_h^*,V_h)}$. Since b is coercive over all of V it is certainly coercive over V_h, and so the last section implies stability with stability constant γ^{-1}. In short, *if the bilinear form is coercive, then for any choice of subspace the Galerkin method is stable with the stability constant bounded by the inverse of the coercivity constant.*

To talk about consistency, as in Chapter 2, we need to define a representative U_h of the solution u in V_h. A natural choice, which we shall make, is that U_h is the orthogonal projection of u into V_h, so that
$$\|u - U_h\| = \inf_{v\in V_h} \|u - v\|.$$

The consistency error is

$$\|L_hU_h - F_h\|_{V_h^*} = \sup_{0\neq v\in V_h} \frac{|(L_hU_h - F_h)(v)|}{\|v\|}.$$

But $(L_hU_h - F_h)(v) = b(U_h, v) - F(v) = b(U_h - u, v)$, so $|(L_hU_h - F_h)(v)| \le M\|u - U_h\|\|v\|$. Therefore the consistency error is bounded by

$$M \inf_{v\in V_h} \|u - v\|.$$

We therefore obtain the convergence estimate

$$\|U_h - u_h\| \le M\gamma^{-1} \inf_{v\in V_h} \|u - v\|.$$

We can then apply the triangle inequality to deduce

$$\|u - u_h\| \le (1 + M\gamma^{-1}) \inf_{v\in V_h} \|u - v\|.$$

This is the fundamental estimate for finite elements. It shows that finite elements are *quasioptimal*, i.e., that the error in the finite element solution is no more than a constant multiple of the error in the best possible approximation from the subspace. The constant can be taken to be 1 plus the bound of the bilinear form times the stability constant.

Remark. We obtained stability using coercivity. From the last section we know that we could obtain stability as well if we had instead of coercivity, a discrete inf-sup condition: There exists $\gamma > 0$ such that for all $0 \neq w \in V_h$ there exists $0 \neq v \in V_h$ such that

$$b(w, v) \ge \gamma\|w\|\|v\|.$$

(In the finite dimensional case the dense range condition follows from the inf-sup condition since an operator from V_h to V_h^* which is 1-to-1 is automatically onto) The big difficulty however is that *the fact that b satisfies the inf-sup condition over V does not by any means imply that it satisfies the inf-sup condition over a finite dimensional subspace* V_h. In short, for coercive problems stability is automatic, but for more general well-posed problems Galerkin methods may or may not be stable (depending on the choice of subspace), and proving stability can be difficult.

7. Finite element approximation theory

In this section we turn to the question of finite element approximation theory, that is of estimating

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

where V_h is a finite element space. For simplicity, we first consider the case where $V_h = M_0^1(\mathcal{T}_h)$, the Lagrange space of continuous piecewise linear functions on a given mesh $\mathcal{T}_h$ where $\mathcal{T}_h$ is a simplicial decomposition of Ω with mesh size $h = \max_{T\in\mathcal{T}_h} \operatorname{diam} T$. (Note: we are overloading the symbol h. If we were being more careful we would consider a sequence of meshes T_i with mesh size h_i tending to zero. But the common practice of using h as both the index and the mesh size saves writing subscripts and does not lead to confusion.)

First we need some preliminary results on Sobolev spaces: density of smooth functions, Poincaré inequality, Sobolev embedding $H^s(\Omega) \subset C(\bar{\Omega})$ if $s > n/2$.

THEOREM 4.1 (Poincaré inequality). *Let Ω be a bounded domain in $\mathbb{R}^n$ with Lipschitz boundary (e.g., smooth boundary or a polygon). Then there exists a constant c, depending only on Ω, such that*

$$\|u\|_{L^2(\Omega)} \le c\|\operatorname{grad} u\|_{L^2(\Omega)}, \quad u \in H^1(\Omega) \text{ such that } \int_\Omega u = 0.$$

The same inequality holds for all $u \in \mathring{H}^1(\Omega)$, or even for all $u \in H^1(\Omega)$ which vanish on a non-empty open subset of the boundary.

The result for u of mean value zero is sometimes called the Poincaré–Neumann inequality. In one dimension it is called Wirtinger's inequality. The result for $u \in \mathring{H}^1(\Omega)$ is sometimes called the Poincaré–Friedrichs inequality or just the Friedrichs inequality.

One proof of this result is based on Rellich's theorem that $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$. Other proofs are more explicit. Here we give a very simple proof of the Poincaré–Neumann inequality in one dimension.

If $u \in C^1(\bar{I})$ where I is an interval of length L, and $\int u = 0$, then there exists a point $x_0 \in I$ such that $u(x_0) = 0$. Therefore

$$|u(x)|^2 = |\int_{x_0}^x u'(s)\,ds|^2 \le |\int_I |u'(s)|\,ds|^2 \le L\int_I |u'(s)|^2\,ds.$$

Integrating, we get $\|u\| \le L\|u'\|$. This can be extended to $u \in H^1$ using density of C^∞ in H^1.

An alternative proof uses Fourier cosine series: $u(x) = \sum_{n=1}^\infty a_n \cos n\pi x/L$ (where the sum starts at $n = 1$, since $\int u = 0$). This gives the result $\|u\| \le L/\pi\|u'\|$, in which the constant is sharp (achieved by $u(x) = \cos \pi x/L$). In fact the result can be proved with the constant d/π, d =diameter of Ω for any *convex* domain in n-dimensions (Payne and Weinberger, Arch. Rat. Mech. Anal. 5 1960, pp. 286–292). The dependence of the constant on the domain is more complicated for non-convex domains.

Multi-index notation: In n-dimensions a multi-index α is an n-tuple $(\alpha_1, \dots, \alpha_n)$ with the α_i non-negative integers. We write $|\alpha| = \alpha_1 + \cdots + \alpha_n$ for the degree of the multi-index, $\alpha! = \alpha_1! \cdots \alpha_n!$,

$$|\alpha| = \alpha_1 + \cdots + \alpha_n, \quad \alpha! = \alpha_1! \cdots \alpha_n!, \quad x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}, \quad D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Thus a general element of $\mathcal{P}_r(\mathbb{R}^n)$ is $p(x) = \sum_{|\alpha| \le r} a_\alpha x^\alpha$, and a general constant-coefficient linear partial differential operator of degree r is $Lu = \sum_{|\alpha| \le r} a_\alpha D^\alpha u$. Taylor's theorem for a smooth function defined in a neighborhood of a point $x_0 \in \mathbb{R}^n$ is

$$u(x) = \sum_{|\alpha| \le m} \frac{1}{\alpha!} D^\alpha u(x_0)(x - x_0)^\alpha + O(|x - x_0|^{m+1})$$

We write $\alpha \le \beta \iff \alpha_i \le \beta_i$, $i = 1, \dots, n$. We have

$$D^\alpha x^\beta = \begin{cases} \frac{\beta!}{(\beta-\alpha)!} x^{\beta-\alpha}, & \alpha \le \beta, \\ 0, & \text{otherwise.} \end{cases}$$

In particular $D^\alpha x^\alpha = \alpha!$.

Let Ω be a bounded domain in $\mathbb{R}^n$ with Lipschitz boundary (for our applications, it will be a triangle). It is easy to see that the DOFs

$$u \mapsto \int_\Omega D^\alpha u, \quad |\alpha| \le r,$$

are unisolvent on $\mathcal{P}_r(\Omega)$. Therefore we can define $P_r : H^r(\Omega) \to \mathcal{P}_r(\Omega)$ by

$$\int_\Omega D^\alpha P_r u(x)\,dx = \int_\Omega D^\alpha u(x)\,dx, \quad |\alpha| \le r.$$

It follows immediately from this definition that $D^\beta P_r u = P_{r-|\beta|} D^\beta u$ for $|\beta| \le r$.

Remark. The rth Taylor polynomial of u at x_0 is $T_r u$ given by

$$D^\alpha T_r u(x_0) = D^\alpha u(x_0), \quad |\alpha| \le r.$$

So $P_r u$ is a sort of averaged Taylor polynomial of u.

Let $u \in H^{r+1}(\Omega)$. Then $u - P_r u$ has integral zero on Ω, so the Poincaré inequality gives

$$\|u - P_r u\| \le c_1 \sum_{|\alpha|=1} \|D^\alpha (u - P_r u)\| = c_1 \sum_{|\alpha|=1} \|D^\alpha u - P_{r-1}(D^\alpha u)\|,$$

for some constant c_1 depending only on Ω (where we use the $L^2(\Omega)$ norm). Applying the same reasoning to $D^\alpha u - P_{r-1}(D^\alpha u)$, we have $\|D^\alpha u - P_{r-1}(D^\alpha u)\|$ is bounded by the sum of the norms of second partial derivatives, so

$$\|u - P_r u\| \le c_2 \sum_{|\alpha|=2} \|D^\alpha u - P_{r-2}(D^\alpha u)\|.$$

Continuing in this way we get

$$\|u - P_r u\| \le c_r \sum_{|\alpha|=r} \|D^\alpha u - P_0(D^\alpha u)\| \le C \sum_{|\alpha|=r+1} \|D^\alpha u\|.$$

For any $|\beta| \le r$ may also apply this result to $D^\beta u \in H^{r+1-|\beta|}$ to get

$$\|D^\beta u - P_{r-|\beta|} D^\beta u\| \le C \sum_{|\gamma| \le r-|\beta|+1} \|D^\gamma D^\beta u\|$$

so

$$\|D^\beta (u - P_r u)\| \le C \sum_{|\alpha|=r+1} \|D^\alpha u\|.$$

Since this holds for all $|\beta| \le r$, we

$$\|u - P_r u\|_{H^r} \le c|u|_{H^{r+1}}, \quad u \in H^{r+1}(\Omega).$$

We have thus given a constructive proof of the follow important result.

Theorem 4.2 (Bramble–Hilbert lemma). *Let Ω be a Lipschitz domain and $r \ge 0$. Then there exists a constant c only depending on the domain Ω and on r such that*

$$\inf_{p \in \mathcal{P}_r} \|u - p\|_{H^r} \le c|u|_{H^{r+1}}, \quad u \in H^{r+1}(\Omega).$$

REMARK. This proof of the Bramble–Hilbert lemma, based on the Poincaré inequality, is due to Verfürth (*A note on polynomial approximation in Sobolev spaces,* M2AN 33, 1999). The method is constructive in that it exhibits a specific polynomial p satisfying the estimate (namely $P_r u$). Based on classical work of Payne and Weinberger on the dependence of the Poincaré constant on the domain mentioned above, it leads to good explicit bounds on the constant in the Bramble–Hilbert lemma. A much older constructive proof is due to Dupont and Scott and taught in the textbook of Brenner and Scott. However that method is both more complicated and it leads a worse bound on the contant. Many texts (e.g., Braess) give a non-constructive proof of the Bramble–Hilbert lemma based on Rellich's compactness theorem.

Now we derive an corollary of the Bramble–Hilbert lemma (which is of such importance that sometimes the corollary is itself referred to as the Bramble–Hilbert lemma).

COROLLARY 4.3. *Let Ω be a Lipschitz domain and $r \ge 0$, and $\pi : H^{r+1}(\Omega) \to \mathcal{P}_r(\Omega)$ be a bounded linear projection onto $\mathcal{P}_r(\Omega)$. Then there exists a constant c which only depends on the domain Ω, r, and the norm of π such that*

$$\|u - \pi u\|_{H^r} \le c|u|_{H^{r+1}}.$$

Note: the hypothesis means that π is a bounded linear operator mapping $H^{r+1}(\Omega)$ into $\mathcal{P}_r(\Omega)$ such that $\pi u = u$ if $u \in \mathcal{P}_r(\Omega)$. Bounded means that $\|\pi\|_{\mathcal{L}(H^{r+1},H^r)} < \infty$. It doesn't matter what norm we choose on $\mathcal{P}_r$, since it is a finite dimensional space.

PROOF.

$$\begin{aligned}\|u - \pi u\|_{H^r} &= \inf_{p \in \mathcal{P}_r} \|(u-p) - \pi(u-p)\|_{H^r} \\ &\le (1 + \|\pi\|_{\mathcal{L}(H^{r+1},H^r)}) \inf_{p \in \mathcal{P}_r} \|u - p\|_{H^{r+1}} = c(1 + \|\pi\|)|u|_{H^{r+1}}.\end{aligned}$$

□

We will be applying this Bramble–Hilbert corollary on the individual triangles T of the finite element mesh. However, if we apply the corollary with $\Omega = T$, the unknown constant c which arises in the corollary will depend on the individual triangle T, and we will not be able to control it. So instead we will apply the corollary on one fixed reference triangle, and then scale the result from the reference triangle to an arbitrary triangle T, and determine how the constant is effected. Thus, we let $\Omega = \hat{T}$ be the unit triangle with vertices $\hat{v}_0 = (0,0)$, $\hat{v}_1 = (1,0)$, and $\hat{v}_2 = (0,1)$, $r = 1$, and let $\pi = I_{\hat{T}}$ the linear interpolant: $I_{\hat{T}} u \in \mathcal{P}_1(\hat{T})$ and $I_{\hat{T}} u(\hat{v}_i) = u(\hat{v}_i)$, $i = 0,1,2$. The $I_{\hat{T}} u$ is defined for all $u \in C(\bar{\hat{T}})$ and $\|I_{\hat{T}} u\|_{L^\infty} \le \|u\|_{L^\infty} \le c\|u\|_{H^2}$, where we use the Sobolev embedding theorem in the last step (and c is some absolute constant). From the corollary we get

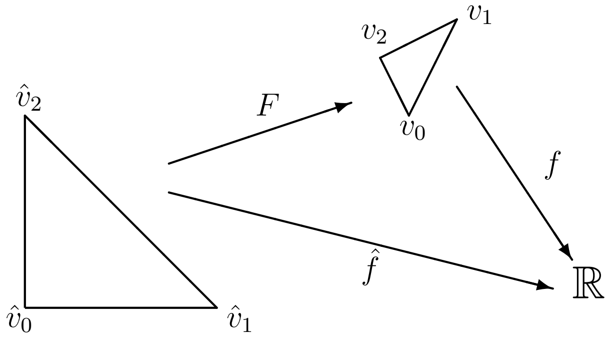
$$\|u - I_{\hat{T}} u\|_{H^1(\hat{T})} \le c|u|_{H^2(\hat{T})}. \tag{4.6}$$

This result will turn out to be a key step in analyzing piecewise linear interpolation.

The next step is to scale this result from the unit triangle to an arbitrary triangle T. Suppose the vertices of T are v_0, v_1, and v_2. There exists a unique affine map F taking $\hat{v}_i$ to v_i, $i = 0,1,2$. Indeed,

$$x = F\hat{x} = v_0 + B\hat{x}, \quad B = (v_1 - v_0 | v_2 - v_0),$$

Figure 4.8. Mapping between the reference triangle and an arbitrary triangle.



where the last notation means that B is the 2×2 matrix whose columns are the vectors $v_1 - v_0$ and $v_2 - v_0$. The map F takes $\hat T$ 1-to-1 onto T. Since the columns of B are both vectors of length at most h_T, certainly the four components b_{ij} of B are bounded by h_T, and so, in any convenient norm, $\|B\| \le ch_T$ (with c depending on the norm chosen). Moreover, $\det B = 2|T|$, the ratio of the area of T to the area of $\hat T$, $|\hat T| = 1/2$.

Now to any function f on T we may associate the pulled-back function $\hat f$ on $\hat T$ where

$$\hat f(\hat x) = f(x) \quad \text{with } x = F\hat x.$$

I.e., $\hat f = f \circ F$. See Figure 4.8.

Next we relate derivatives and norms of a function f with its pull-back $\hat f$. For the derivative we simply use the chain rule:

$$\frac{\partial \hat f}{\partial \hat x_j}(\hat x) = \sum_{i=1}^{2} \frac{\partial f}{\partial x_i}(x)\frac{\partial x_i}{\partial \hat x_j} = \sum_{i=1}^{2} b_{ij}\frac{\partial f}{\partial x_i}(x).$$

Similarly,

$$\frac{\partial^2 \hat f}{\partial \hat x_j \partial \hat x_l}(\hat x) = \sum_{i=1}^{2}\sum_{k=1}^{2} b_{ij}b_{kl}\frac{\partial^2 f}{\partial x_i \partial x_k}(x),$$

etc. Thus we have

$$\sum_{|\alpha|=r} |D^\alpha \hat f(\hat x)| \le c\|B\|^r \sum_{|\beta|=r} |D^\beta f(x)|. \tag{4.7}$$

In the same way we have

$$\sum_{|\beta|=r} |D^\beta f(x)| \le c\|B^{-1}\|^r \sum_{|\alpha|=r} |D^\alpha \hat f(\hat x)|. \tag{4.8}$$

In (4.7) we may bound $\|B\|$ by ch_T. To bound $\|B^{-1}\|$ in (4.8), we introduce another geometric quantity, namely the diameter ρ_T of the inscribed disk in T. Then any vector of length ρ_T is the difference of two points in T (two opposite points on the inscribed circle), and these are mapped by B^{-1} to two points in $\hat T$, which are at most $\sqrt{2}$ apart. Thus, using the Euclidean norm $\|B^{-1}\| \le \sqrt{2}/\rho_T$, i.e., $\|B^{-1}\| = O(\rho_T^{-1})$. We have thus shown

$$\sum_{|\alpha|=r} |D^\alpha \hat{f}(\hat{x})| \le c h_T^r \sum_{|\beta|=r} |D^\beta f(x)|, \quad \sum_{|\beta|=r} |D^\beta f(x)| \le c\rho_T^{-r} \sum_{|\alpha|=r} |D^\alpha \hat{f}(\hat{x})|.$$

Now let us consider how norms map under pull-back. First we consider the L^2 norm. Let $|T|$ denote the area of T. Changing variables from $\hat{x}$ to $x = F\hat{x}$, we have

$$\|f\|_{L^2(T)}^2 = \int_T |f(x)|^2\, dx = 2|T| \int_{\hat{T}} |\hat{f}(\hat{x})|^2\, d\hat{x} = 2|T| \|\hat{f}\|_{L^2(\hat{T})}^2.$$

That is, $\|f\|_{L^2(T)} = \sqrt{2|T|}\|\hat{f}\|_{L^2(\hat{T})}$. Next consider the H^r seminorm:

$$\begin{aligned} |f|_{H^r(T)}^2 = \int_T \sum_{|\beta|=r} |D^\beta f(x)|^2\, dx \le & c\rho_T^{-2r} \int_T \sum_{|\alpha|=r} |D^\alpha \hat{f}(\hat{x})|^2\, dx \\ = & 2c|T|\rho_T^{-2r} \int_{\hat{T}} \sum_{|\alpha|=r} |D^\alpha \hat{f}(\hat{x})|^2\, d\hat{x}, \end{aligned}$$

so

$$|f|_{H^r(T)} \le c\sqrt{|T|}\rho_T^{-r} |\hat{f}|_{H^r(\hat{T})}.$$

Similarly,

$$|\hat{f}|_{H^r(\hat{T})} \le c\frac{1}{\sqrt{|T|}} h_T^r |f|_{H^r(T)}.$$

Now let $u \in H^2(T)$, and let $\hat{u} \in H^2(\hat{T})$ be the corresponding function. We saw in (4.6) that

$$\|\hat{u} - I_{\hat{T}}\hat{u}\|_{H^1(\hat{T})} \le c|\hat{u}|_{H^2(\hat{T})}.$$

Now it is easy to see that the pull-back of $I_T u$ is $I_{\hat{T}}\hat{u}$ (both are linear functions which equal $u(v_i)$ at the vertex $\hat{v}_i$). Therefore $\hat{u} - I_{\hat{T}}\hat{u} = \widehat{u - I_T u}$. We then have

$$\|u - I_T u\|_{L^2(T)} \le c\sqrt{|T|}\|\hat{u} - I_{\hat{T}}\hat{u}\|_{L^2(\hat{T})} \le c\sqrt{|T|}|\hat{u}|_{H^2(\hat{T})} \le c h_T^2 |u|_{H^2(T)},$$

and

$$|u - I_T u|_{H^1(T)} \le c\sqrt{|T|}\rho_T^{-1}|\hat{u} - I_{\hat{T}}\hat{u}|_{H^1(\hat{T})} \le c\sqrt{|T|}\rho_T^{-1}|\hat{u}|_{H^2(\hat{T})} \le c h_T^2/\rho_T |u|_{H^2(T)}.$$

If the triangle T is not too distorted, then ρ_T is not much smaller than h_T. Let us define $\sigma_T = h_T/\rho_T$, the *shape constant* of T. We have proved:

THEOREM 4.4. *Let T be a triangle with diameter h_T, and let I_T be the linear interpolant at the vertices of T. Then there exists an absolute constant c such that*

$$\|u - I_T u\|_{L^2(T)} \le c h_T^2 |u|_{H^2(T)}, \quad u \in H^2(T).$$

Moreover there exists a constant c' depending only on the shape constant for T such that

$$|u - I_T u|_{H^1(T)} \le c' h_T |u|_{H^2(T)}, \quad u \in H^2(T).$$

Now we have analyzed linear interpolation on a single but arbitrary triangle, we can just add over the triangles to analyze piecewise linear interpolation.

THEOREM 4.5. *Suppose we have a sequence of triangulations $\mathcal{T}_h$ with mesh size h tending to 0. For u a continuous function on $\bar{\Omega}$, let $I_h u$ denote the continuous piecewise linear interpolant of u on the mesh $\mathcal{T}_h$. Then there exists an absolute constant c such that*

$$\|u - I_h u\|_{L^2(\Omega)} \le c h^2 |u|_{H^2(\Omega)}, \quad u \in H^2(\Omega).$$

If the mesh sequence is shape regular (i.e., the shape constant is uniformly bounded), then there exists a constant c' depending only on a bound for the shape constant such that

$$|u - I_h u|_{H^1(\Omega)} \le c' h |u|_{H^2(\Omega)}, \quad u \in H^2(\Omega).$$

In a similar fashion, for the space of Lagrange finite elements of degree r we can analyze the interpolant defined via the degrees of freedom.

THEOREM 4.6. *Suppose we have a sequence of triangulations $\mathcal{T}_h$ with mesh size h tending to 0. Let V_h be the space of Lagrange finite elements of degree r with respect to the mesh, and for u a continuous function on $\bar{\Omega}$, let $I_h u$ denote the interpolant of u into V_h defined through the degrees of freedom. Then there exists an absolute constant c such that*

$$\|u - I_h u\|_{L^2(\Omega)} \le c h^s |u|_{H^s(\Omega)}, \quad u \in H^s(\Omega), \quad 2 \le s \le r+1.$$

If the mesh sequence is shape regular (i.e., the shape constant is uniformly bounded), then there exists a constant c' depending only on a bound for the shape constant such that

$$|u - I_h u|_{H^1(\Omega)} \le c' h^{s-1} |u|_{H^s(\Omega)}, \quad u \in H^s(\Omega), \quad 2 \le s \le r+1.$$

Thus for smooth u (more precisely, $u \in H^{r+1}$), we obtain the rate of convergence $O(h^{r+1})$ in L^2 and $O(h^r)$ in H^1 when we approximation with Lagrange elements of degree r.

The proof of this result is just the Bramble–Hilbert lemma and scaling. Note that we must assume $s \ge 2$ so that $u \in H^s$ is continuous and the interpolant is defined. On the other hand we are limited to a rate of $O(h^{r+1})$ in L^2 and $O(h^r)$ in H^1, since the interpolant is exact on polynomials of degree r, but not higher degree polynomials.

8. Error estimates for finite elements

8.1. Estimate in H^1.

To be concrete, consider the Dirichlet problem

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

with a coefficient a bounded above and below by positive constants of Ω and $f \in L^2$. The weak formulation is: find $u \in V = \mathring{H}^1(\Omega)$ such that

$$b(u, v) = F(v), \quad v \in V, \tag{4.9}$$

where $b(u, v) = \int_\Omega \operatorname{grad} u \cdot \operatorname{grad} v \, dx$, $F(v) = \int_\Omega f v \, dx$. Clearly b is bounded: $|b(w, v)| \le M\|w\|_1 \|v\|_1$, (with $M = \sup a$). By Poincaré's inequality, Theorem 4.1, b is coercive: $b(v, v) \ge \gamma \|v\|_1^2$.

Now suppose that Ω is a polygon and let V_h be the space of Lagrange finite elements of degree r vanishing on the boundary with respect to a mesh of Ω of mesh size h, and define u_h to be the finite element solution: $u_h \in V_h$,

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

By the fundamental estimate for finite element methods, proven in Section 6,

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

(where $c = 1 + M\gamma^{-1}$). Then we may apply the finite element approximation theory summarized in Theorem 4.6, and conclude that

$$\|u-u_h\|_1 \le ch^r \|u\|_{r+1} \tag{4.10}$$

as long as the solution u belongs to H^{r+1}. If u is less smooth, the rate of convergence will be decreased accordingly.

In short, one proves the error estimate in H^1 by using quasi-optimality in H^1 (which comes from coercivity), and then finite element approximation theory.

8.2. Estimate in L^2. Now let $g \in L^2(\Omega)$ be a given function, and consider the computation of $G(u) := \int_\Omega ug\,dx$, which is a functional of the solution u of our Dirichlet problem. We ask how accurately $G(u_h)$ approximates $G(u)$. To answer this, we define an auxiliary function $\phi \in V$ by

$$b(w,\phi) = G(w), \quad w \in V.$$

This is simply a weak formulation of the Dirichlet problem

$$-\operatorname{div} a \operatorname{grad} \phi = g \text{ in } \Omega, \quad \phi = 0 \text{ on } \partial\Omega.$$

We will assume that this problem satisfies H^2 regularity, i.e., the solution $\phi \in H^2(\Omega)$ and satisfies

$$\|\phi\|_2 \le c\|g\|_0.$$

This is true, for example, if Ω is either a convex Lipschitz domain or a smooth domain and a is a smooth coefficient.

Remark. Note that we write $b(w,\phi)$ with the trial function ϕ second and the test function w first, the opposite as for the original problem (4.9). Since the bilinear form we are considering is symmetric, this is not a true distinction. But if we started with an nonsymmetric bilinear form, we would still define the auxiliary function ϕ in this way. In short ϕ satisfies a boundary value problem for the *adjoint* differential equation.

Now consider the error in $G(u)$:

$$G(u) - G(u_h) = \int_\Omega (u-u_h)g\,dx = b(u-u_h,\phi) = b(u-u_h,\phi - v)$$

for any $v \in V_h$, where the second equality comes from the definition of the auxiliary function ϕ and the third from Galerkin orthogonality (4.3). Therefore

$$|G(u)-G(u_h)| \le M\|u-u_h\|_1 \inf_{v\in V_h} \|\phi - v\|_1.$$

Now finite element approximation theory and 2-regularity tell us

$$\inf_{v\in V_h} \|\phi - v\|_1 \le ch\|\phi\|_2 \le ch\|g\|_0.$$

Thus

$$|G(u)-G(u_h)| \le ch\|u-u_h\|_1\|g\|_0 \le ch^{r+1}\|u\|_{r+1}\|g\|_0.$$

In short, if $g \in L^2(\Omega)$, the error in $G(u) = \int_\Omega ug\,dx$ is $O(h^{r+1})$, one power of h higher than the H^1 error.

A very important special case is when $g = u - u_h$. Then $G(u) - G(u_h) = \|u - u_h\|_0^2$, so we have

$$\|u - u_h\|_0^2 \le ch\|u - u_h\|_1 \|u - u_h\|_0,$$

or

$$\|u - u_h\|_0 \le ch\|u - u_h\|_1 \le ch^{r+1}\|u\|_{r+1}.$$

That is, the L^2 error in the finite element method is one power of h higher than the H^1 error.

Remark. The idea of introducing an auxiliary function ϕ, so we can express $G(u - u_h)$ or $\|u - u_h\|_0^2$ as $b(u - u_h, \phi)$ and estimate it using Galerkin orthogonality is the Aubin–Nitsche duality method. If we use it to estimate $G(u - u_h)$ where g is smoother than L^2 and we have higher order elliptic regularity, we can get even higher order estimates, so called negative-norm estimates.

9. A posteriori error estimates and adaptivity

The error estimate (4.10) is a typical *a priori* error estimate for the finite element method. It indicates that, as long as we know a priori that the unknown solution of our problem belongs to H^{r+1}, then the error $\|u - u_h\|_1$ will converge to zero as $O(h^r)$. By contrast an *a posteriori* error estimate attempts to bound the error in terms of u_h, allowing the error in the finite element solution to be approximated once the finite element solution itself has been calculated. One important use of a posteriori error estimates is in estimating how accurate the computed solution is. Another relates to the fact that the some a posteriori error estimates give a way of attributing the error to the different elements of the mesh. Therefore they suggest how the mesh might be refined to most effectively decrease the error (basically by subdividing the elements which are contributing a lot to the error). This is the basic idea of *adaptivity*, which we shall discuss below.

9.1. The Clément interpolant. First we need a new tool from finite element approximation theory. Suppose we are given a polygonal domain Ω and a mesh of mesh size h. Let V_h be the usual Lagrange finite element space of degree r. Given a continuous function u on $\bar{\Omega}$, we may define the interpolant $I_h u$ of u into V_h through the usual degrees of freedom. Then we have the error estimate

$$\|u - I_h u\|_t \le ch^{s-t}\|u\|_s, \quad u \in H^s(\Omega),$$

valid for integers $0 \le t \le 1$, $2 \le s \le r+1$. See Theorem 4.6 (the constant c here depends only on r and the shape regularity of the mesh). We proved this result element-by-element, using the Bramble–Hilbert lemma and scaling. Of course this result implies that

$$\inf_{v \in V_h} \|u - v\|_t \le ch^{s-t}\|u\|_s, \quad u \in H^s(\Omega),$$

for the same ranges of t and s. The restriction $t \le 1$ is needed, since otherwise the functions in V_h, being continuous but not generally C^1, do not belong to $H^t(\Omega)$. Here we are concerned with weakening the restriction $s \ge 2$, so we can say something about the approximation by piecewise polynomials of a function u that does not belong to $H^2(\Omega)$. We might hope for example that

$$\inf_{v\in V_h} \|u-v\|_0 \le ch\|u\|_1, \quad u \in H^1(\Omega).$$

In fact, this estimate is true, and is important to the development of a posteriori error estimates and adaptivity. However it can not be proven using the usual interpolant $I_h u$, because $I_h u$ is not defined unless the function u has well-defined point values at the node points of the mesh, and this is not true for a general function $u \in H^1$. (In 2- and 3-dimensions the Sobolev embedding theorem implies the existence of point values for function in H^2, but not in H^1.)

The way around this is through a different operator than I_h, called the Clément interpolant, or quasi-interpolant. For each polynomial degree $r \ge 1$ and each mesh, the Clément interpolant $\Pi_h : L^2(\Omega) \to V_h$ is a bounded linear operator. Its approximation properties are summarized in the following theorem.

Theorem 4.7 (Clément interpolant). *Let Ω be a domain in $\mathbb{R}^n$ furnished with a simplicial triangulation with shape constant γ and maximum element diameter h, let r be a positive integer, and let V_h denote the Lagrange finite element space of continuous piecewise polynomials of degree r. Then there exists a bounded linear operator $\Pi_h : L^2(\Omega) \to V_h$ and a constant c depending only on γ and r such that*

$$\|u - \Pi_h u\|_t \le ch^{s-t}\|u\|_s, \quad u \in H^s(\Omega),$$

for all $0 \le t \le s \le r+1$, $t \le 1$.

Now we define the Clément interpolant. Let $\mu_i : C(\bar\Omega) \to \mathbb{R}$, $i = 1, \dots, N$, be the usual DOFs for V_h and ϕ_i the corresponding basis functions. Thus the usual interpolant is

$$I_h u = \sum_i \mu_i(u)\phi_i, \quad u \in C(\bar\Omega).$$

To define the Clément interpolant we let S_i denote the support of ϕ_i, i.e., the union of triangles where ϕ_i is not identically zero (if μ_i is a vertex degree of freedom this is the union of the elements with that vertex, if an edge degree of freedom, the union of the triangles with that edge, etc.). Denote by $P_i : L^2(S_i) \to \mathcal{P}_r(S_i)$ the L^2-projection. Then we set

$$\Pi_h u = \sum_i \mu_i(P_i u)\phi_i, \quad u \in L^2(\Omega).$$

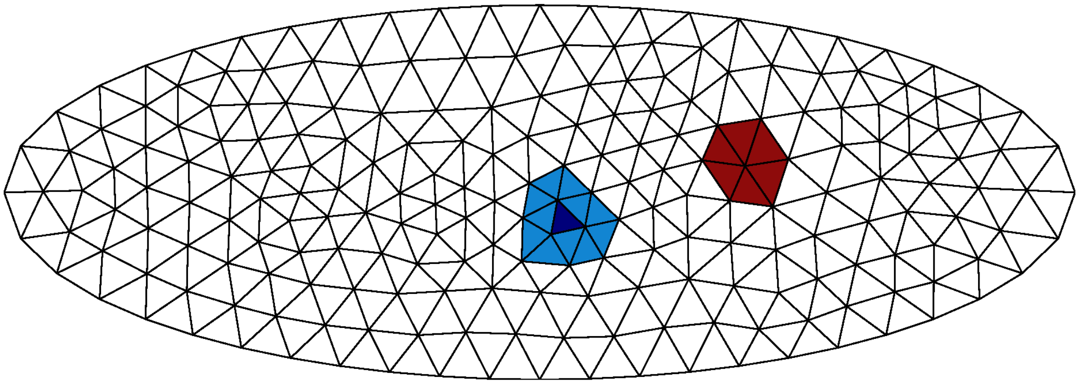
The usual interpolant $I_h u$ is completely local in the sense that if u vanishes on a particular triangle T, then $I_h u$ also vanishes on T. The Clément interpolation operator is not quite so local, but is nearly local in the following sense. If u vanishes on the set $\tilde T$, defined to be the union of the triangles that share at least one vertex with T (see Figure 4.9), then $\Pi_h u$ vanishes on T. In fact, for any $0 \le t \le s \le r+1$,

$$\|u - \Pi_h u\|_{H^t(T)} \le ch_T^{s-t}\|u\|_{H^s(\tilde T)}, \quad u \in H^s(\tilde T), \tag{4.11}$$

where the constant depends only on the shape regularity of the mesh and r. From (4.11), using the shape regularity, the estimate of Theorem 3.7 easily follows.

To avoid too much technicality, we shall prove (4.11) in the case of linear elements, $r = 1$. Thus we are interested in the case $t = 0$ or 1 and $t \le s \le 2$. Let T be a particular triangle

FIGURE 4.9. Shown in brown S_z for a vertex z and in blue $\tilde{T}$ for a triangle T.



and let z_i, μ_i, ϕ_i, S_i denote its vertices and corresponding DOFs, basis functions, and their supports, for $i = 1, 2, 3$. Note that it is easy to see that

$$\|\phi_i\|_{L^2(T)} \le |T|^{1/2} \le ch_T, \quad \|\operatorname{grad}\phi_i\|_{L^2(T)} \le ch_T^{-1}|T|^{1/2} \le c, \tag{4.12}$$

where the constants may depend on the shape constant for T. Next, using the Bramble–Hilbert lemma and scaling on S_i, we have for $0 \le t \le s \le 2$,

$$\|u - P_i u\|_{H^t(S_i)} \le ch_T^{s-t}\|u\|_{H^s(S_i)}. \tag{4.13}$$

In performing the scaling, we map the whole set S_i to the corresponding set $\hat{S}_i$ using the affine map F which takes one of the triangles T in S_i to the unit triangle $\hat{T}$. The Bramble–Hilbert lemma is applied on the scaled domain $\hat{S}_i$. Although there is not just a single domain S_i—it depends on the number of triangles meeting at the vertex z_i and their shapes—the constant which arises when applying the Bramble–Hilbert lemma on the scaled domain can bounded on the scaled domain in terms only of the shape constant for the triangulation (this can be established using compactness).

We also need one other bound. Let i and j denote the indices of two different vertices of T. For $u \in L^2(\tilde{T})$, both $P_i u$ and $P_j u$ are defined on T. If $\hat{u}$ denotes the corresponding function on the scaled domain $\hat{\tilde{T}}$, then we have

$$\begin{aligned} \|P_i u - P_j u\|_{L^\infty(T)} &= \|\hat{P}_i\hat{u} - \hat{P}_j\hat{u}\|_{L^\infty(\hat{T})} \le c\|\hat{P}_i\hat{u} - \hat{P}_j\hat{u}\|_{L^2(\hat{T})} \\ &\le c(\|\hat{P}_i\hat{u} - \hat{u}\|_{L^2(\hat{T})} + \|\hat{P}_j\hat{u} - \hat{u}\|_{L^2(\hat{T})}) \\ &\le c(\|\hat{P}_i\hat{u} - \hat{u}\|_{L^2(\hat{S}_i)} + \|\hat{P}_j\hat{u} - \hat{u}\|_{L^2(\hat{S}_j)}) \end{aligned}$$

where the first inequality comes from equivalence of norms on the finite dimensional space $\mathcal{P}_1(\hat{T})$ and the second from the triangle inequality. Both of the terms on the right-hand side can be bounded using the Bramble–Hilbert lemma and scaled back to S_i, just as for (4.13). In this way we obtain the estimate

$$\|P_i u - P_j u\|_{L^\infty(T)} \le ch_T^s|T|^{-1/2}\|u\|_{H^s(\tilde{T})}. \tag{4.14}$$

Therefore also

$$|\mu_i(P_i u - P_j u)| \le ch_T^s|T|^{-1/2}\|u\|_{H^s(\tilde{T})}. \tag{4.15}$$

Now on the triangle T with vertices numbered z_1, z_2, z_3 for simplicity,

$$\Pi_h u = \sum_{i=1}^{3} \mu_i(P_i u)\phi_i.$$

Since $P_1 u$ is a linear polynomial on T,

$$u - \Pi_h u = (u - P_1 u) - \sum_{i=2}^{3} \mu_i(P_i u - P_1 u)\phi_i.$$

The first term on the right hand side is bounded using (4.13):

$$\|u - P_1 u\|_{H^t(T)} \le c h_T^{s-t} \|u\|_{H^s(\tilde{T})}.$$

For the second term we have

$$\|\mu_i(P_i u - P_1 u)\phi_i\|_{H^t(T)} \le \|P_i u - P_1 u\|_{L^\infty(T)} \|\phi_i\|_{H^t(T)},$$

which satisfies the desired bound by (4.12) and (4.15).

For the next section an important special case of (4.11) is

$$\|u - \Pi_h u\|_{L^2(T)} \le c h_T \|u\|_{H^1(\tilde{T})}. \tag{4.16}$$

Another case is H^1 boundedness:

$$\|u - \Pi_h u\|_{H^1(T)} \le c \|u\|_{H^1(\tilde{T})}. \tag{4.17}$$

We draw one more conclusion, which we will need below. Let $\hat{T}$ denote the unit triangle, and $\hat{e}$ one edge of it. The trace theorem then tells us that

$$\|\hat{u}\|_{L^2(\hat{e})}^2 \le c(\|\hat{u}\|_{L^2(\hat{T})}^2 + \|\operatorname{grad} \hat{u}\|_{L^2(\hat{T})}^2), \quad \hat{u} \in H^1(\hat{T}).$$

If we use linear scaling to an arbitrary triangle, we get

$$\|u\|_{L^2(e)}^2 \le c(h_T^{-1}\|u\|_{L^2(T)}^2 + h_T\|\operatorname{grad} u\|_{L^2(T)}^2), \quad u \in H^1(T),$$

where the constant depends only on the shape constant of T. If we now apply this with u replaced by $u - \Pi_h u$ and use (4.16) and (4.17), we get this bound for the Clément interpolant:

$$\|u - \Pi_h u\|_{L^2(e)} \le c h_e^{1/2} \|u\|_{H^1(\tilde{T})}, \tag{4.18}$$

where h_e is the length of the edge e, T is a triangle containing e, and c depends only on the shape constant for the mesh.

9.2. The residual and the error. Consider our usual model problem

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

with a continuous positive coefficient a on $\bar{\Omega}$ and $f \in L^2$. The weak formulation is to find $u \in V$ satisfying

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

where $V = \mathring{H}^1(\Omega)$ and

$$b(w, v) = \int a \operatorname{grad} w \cdot \operatorname{grad} v \, dx, \quad F(v) = \int f v \, dx, \quad w, v \in V.$$

The bilinear form is bounded and coercive on $\mathring{H}^1$:

$$|b(w,v)| \le M\|w\|_1\|v\|_1, \quad b(v,v) \ge \gamma\|v\|_1^2, \quad w,v \in V.$$

Now we suppose that we have computed an approximation U of the solution u and we wish to assess the norm of the error $u-U$ (we are most interested in the case $U = u_h$, the finite element solution). Just as for linear algebra problems, in which we find an approximate solution to a linear system, we shall approach the error through the residual, which is computable. But what do we mean by the *residual* in the solution to such a weakly formulated equation? As discussed at the start of Section 5, the weak formulation may be viewed as an operator equation $Lu = F$, where L is a linear operator for V to V^*. In our case, $V = \mathring{H}^1$ and its dual, V^* is generally denoted H^{-1}, with the dual norm denoted by $\|\cdot\|_{-1}$. Thus the residual $R(U) = F - LU \in H^{-1}$, i.e., it is a linear functional on $\mathring{H}^1$. Specifically,

$$R(U)w = F(w) - b(U,w), \quad w \in V.$$

Clearly $R(U)w = b(u-U,w)$. It follows immediately that $|R(U)w| \le M\|u-U\|_1\|w\|_1$ for all $w \in \mathring{H}^1$, or, equivalently, that $\|R(U)\|_{-1} \le M\|u-U\|_1$. On the other hand, taking $w = u-U$ and using the coercivity, we get $\gamma\|u-U\|_1^2 \le R(U)(u-U) \le \|R(U)\|_{-1}\|u-U\|_1$. Thus

$$M^{-1}\|R(U)\|_{-1} \le \|u-U\|_1 \le \gamma^{-1}\|R(U)\|_{-1}.$$

In short, *the H^{-1} norm of the residual $R(U)$ is equivalent to the H^1 norm of the error $u-U$.*

9.3. Estimating the residual. Now let V_h be the Lagrange finite element subspace of $V = \mathring{H}^1$ corresponding to some mesh $\mathcal{T}_h$ and some polynomial degree r, and let u_h be the corresponding finite element solution. We have just seen that we may estimate the error $\|u-u_h\|_1$ by estimating the H^{-1} error in the residual

$$R(u_h)w = F(w) - b(u_h,w), \quad w \in V.$$

This quantity does not involve the unknown solution u, so we may hope to compute it *a posteriori*, i.e., after we have computed u_h.

We start by integrating by parts on each element T of the mesh to rewrite $R(u_h)w$:

$$\begin{aligned} R(u_h)w &= \sum_{T\in\mathcal{T}_h}\int_T (fw - a\operatorname{grad}u_h\cdot\operatorname{grad}w)\,dx \\ &= \sum_T\int_T (f + \operatorname{div}a\operatorname{grad}u_h)w\,dx - \sum_T\int_{\partial T} a\frac{\partial u_h}{\partial n_T}w\,ds. \end{aligned}$$

Consider the final sum. We can split each integral over ∂T into the sum of the integrals of the three edges of T. Each edge e which is not contained in the boundary comes in twice. For such an edge, let T_- and T_+ be the two triangles which contain e and set

$$R_e(u_h) = -a\left(\frac{\partial u_h|_{T_-}}{\partial n_{T_-}} + \frac{\partial u_h|_{T_+}}{\partial n_{T_+}}\right) \in L^2(e)$$

on e. Since $n_{T_-} = -n_{T_+}$ the term in parenthesis is the *jump* in the normal derivative of u_h across the edge e. Also, for $T \in \mathcal{T}_h$, we set $R_T(u_h) = f + \operatorname{div} a \operatorname{grad} u_h \in L^2(T)$. Then

$$\begin{aligned} R(u_h)w &= \sum_T \int_T R_T(u_h)w\,dx + \sum_{e\in\mathring{\mathcal{E}}} \int_e R_e(u_h)w\,ds. \\ &= \sum_T \Big[\int_T R_T(u_h)w\,dx + \frac{1}{2}\sum_{\substack{e\in\mathring{\mathcal{E}}\\ e\subset T}} \int_e R_e(u_h)w\,ds\Big], \end{aligned} \tag{4.19}$$

where in the last step we used the fact that each $e \in \mathring{\mathcal{E}}$ belongs to 2 triangles.

Next we use *Galerkin orthogonality*: since u_h is the finite element solution, we have

$$b(u - u_h, v) = 0, \quad v \in V_h.$$

In terms of the residual this says that

$$R(u_h)w = R(u_h)(w - v), \quad v \in V_h.$$

In particular, we may choose $v = \Pi_h w$, the Clément interpolant in this equation. Combining with (4.19) (with w replaced by $w - \Pi_h w$) we get

$$\begin{aligned} R(u_h)w &= R(u_h)(w - \Pi_h w) \\ &= \sum_T \Big[\int_T R_T(u_h)(w - \Pi_h w)\,dx + \frac{1}{2}\sum_{\substack{e\in\mathring{\mathcal{E}}\\ e\subset T}} \int_e R_e(u_h)(w - \Pi_h w)\,ds\Big], \end{aligned} \tag{4.20}$$

Next, we bound the terms in the brackets on the right hand side of (4.20) for $w \in \mathring{H}^1$. First we use (4.16) to get

$$\int_T R_T(u_h)(w - \Pi_h w)\,dx \le \|R_T(u_h)\|_{L^2(T)}\|w - \Pi_h w\|_{L^2(T)} \le ch_T\|R_T(u_h)\|_{L^2(T)}\|w\|_{H^1(\tilde T)}.$$

In a similar way, but using (4.18), we obtain for $e \subset T$,

$$\int_e R_e(u_h)(w - \Pi_h w)\,ds \le ch_e^{1/2}\|R_e(u_h)\|_{L^2(e)}\|w\|_{H^1(\tilde T)}.$$

Combining the last three estimates, we get

$$\begin{aligned} |R(u_h)w| &\le c\sum_T \big[h_T\|R_T(u_h)\|_{L^2(T)} + \sum_{e\subset T} h_e^{1/2}\|R_e(u_h)\|_{L^2(e)}\big]\,\|w\|_{H^1(\tilde T)} \\ &\le c\Big\{\sum_T \big[h_T^2\|R_T(u_h)\|^2_{L^2(T)} + \sum_{e\subset T} h_e\|R_e(u_h)\|^2_{L^2(e)}\big]\Big\}^{1/2}\Big(\sum_T \|w\|^2_{H^1(\tilde T)}\Big)^{1/2} \\ &\le c\Big\{\sum_T \big[h_T^2\|R_T(u_h)\|^2_{L^2(T)} + \sum_{e\subset T} h_e\|R_e(u_h)\|^2_{L^2(e)}\big]\Big\}^{1/2}\|w\|_1, \end{aligned}$$

where we invoked the shape regularity in the last step. Since this estimate holds for all $w \in \mathring{H}^1$, we have shown that

$$\|R(u_h)\|_{-1} \le c\Big\{\sum_T \big[h_T^2\|R_T(u_h)\|^2_{L^2(T)} + \sum_{e\subset T} h_e\|R_e(u_h)\|^2_{L^2(e)}\big]\Big\}^{1/2}.$$

In view of the equivalence of the H^{-1} norm of the residual and the H^1 norm of the error established in the preceding subsection, this gives us the a posteriori error estimate

$$\|u - u_h\|_1 \le c\Big\{\sum_T \big[h_T^2\|R_T(u_h)\|_{L^2(T)}^2 + \sum_{e\subset T} h_e\|R_e(u_h)\|_{L^2(e)}^2\big]\Big\}^{1/2}. \tag{4.21}$$

This is a key result. First of all, it gives us a bound on the norm of the error of the finite element solution in terms of quantities that can be explicitly computed (except for the unknown constant c). Second, the error bound is the square root of a sum of terms associated to the individual triangles. Thus, we have a way of assigning portions of the error to the various elements. This will enable us to base our adaptive strategy on refining those triangles for which the corresponding portion of the error for either the triangle itself or for one of its edges is relatively large.

9.4. A posteriori error indicators and adaptivity. Specifically, we associate to each triangle an *error indicator*:

$$\eta_T^2 := h_T^2\|R_T(u_h)\|_{L^2(T)}^2 + \frac{1}{2}\sum_{e\subset\partial T} h_e\|R_e(u_h)\|_{L^2(e)}^2$$

The factor of 1/2 is usually used, to account for the fact that each edge belongs to two triangles. In terms of the error indicators, we can rewrite the a posteriori estimate as

$$\|u - u_h\|_1 \le c\left(\sum_T \eta_T^2\right)^{1/2}.$$

Our basic adaptive strategy then proceeds via the following SOLVE-ESTIMATE-MARK-REFINE loop:

- SOLVE: Given a mesh, compute u_h
- ESTIMATE: For each triangle T compute η_T. If $(\sum_T \eta_T^2)^{1/2} \le tol$, quit.
- MARK: Mark those elements T for which η_T is too large for refinement.
- REFINE: Create a new mesh with the marked elements refined.

We have already seen how to carry out the SOLVE and ESTIMATE steps. There are a number of possible strategies for choosing which elements to mark. One of the simplest is *maximal marking*. We pick a number ρ between 0 and 1, compute $\eta_{\max} = \max_T \eta_T$, and refine those elements T for $\eta_T \ge \rho\eta_{\max}$. Another approach, which is usually preferred, imposes the *Dörfler marking* criterion, which requires that some collection of elements $\mathcal{S}$ is marked so that $\sum_{T\in\mathcal{S}} \eta_T^2 \ge \rho^2 \sum_{T\in\mathcal{T}_h} \eta_T^2$, i.e., we mark enough elements that they account for a given portion ρ of the total error. The program on the next page shows one way to implement this (there are others).

Once we have marked the elements, there is the question of how to carry out the refinement to be sure that all the marked elements are refined and there is not too much additional refinement. In 2-dimensions this is quite easy. Most schemes are based either on dividing each triangle in two, or dividing the marked triangles into 4 congruent triangles. Generally, after refining the marked elements, additional elements have to be refined to avoid hanging nodes in which a vertex of an element fall in the interior of the edge of a neighboring element. In 3-dimensions things are more complicated, but good refinement schemes (which retain shape regularity and avoid hanging nodes) are known.

9.5. Examples of adaptive finite element computations. On the next page we present a bare-bones adaptive Poisson solver written in FEniCS, displayed on the next page. This code uses Lagrange piecewise linear finite elements to solve the Dirichlet problem

$$-\Delta u = 1 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega,$$

with Ω an L-shaped domain and $f \equiv 1$, with the error indicators and marking strategy described above. The solution behaves like $r^{2/3}\sin(2\theta/3)$ in a neighborhood of the reentrant corner, and so is not in H^2. The results can be seen in Figure 4.10. The final mesh has 6,410 elements, all right isoceles triangles, with hypotenuse length h ranging from 0.044 to 0.002. If we used a uniform mesh with the smallest element size, this would require over 3 million elements. Figure 4.11 displays an adaptive mesh in 3D.

```
"""
Adaptive Poisson solver using a residual-based energy-norm error
estimator

  eta_h**2 = sum_T eta_T**2

with

  eta_T**2 = h_T**2 ||R_T||_T**2 + c h_T ||R_dT||_dT**2

where

  R_T =  f + div grad u_h
  R_dT = 2 avg(grad u_h * n)  (2*avg is jump, since n switches sign across edges)

and a Dorfler marking strategy

Adapted by Douglas Arnold from code of Marie Rognes
"""

from dolfin import *
from sys import stdin
from numpy import zeros

# Stop when sum of eta_T**2 < tolerance or max_iterations is reached
tolerance = 0.04
max_iterations = 20

# Create initial mesh
mesh = Mesh("l-shape-mesh.xml")
mesh.order()
figure(0) # reuse plotting window

# Define boundary and boundary value for Dirichlet conditions
u0 = Constant(0.0)
def boundary(x, on_boundary):
    return on_boundary

# SOLVE - ESTIMATE - MARK - REFINE loop
for i in range(max_iterations):

    # *** SOLVE step
    # Define variational problem and boundary condition
    # Solve variational problem on current mesh
    V = FunctionSpace(mesh, "CG", 1)
    u = TrialFunction(V)
    v = TestFunction(V)
    f = Constant(1.0)
    a = inner(grad(u), grad(v))*dx
    L = f*v*dx
    u_h = Function(V)
    solve(a==L, u_h, DirichletBC(V, u0, boundary))
```

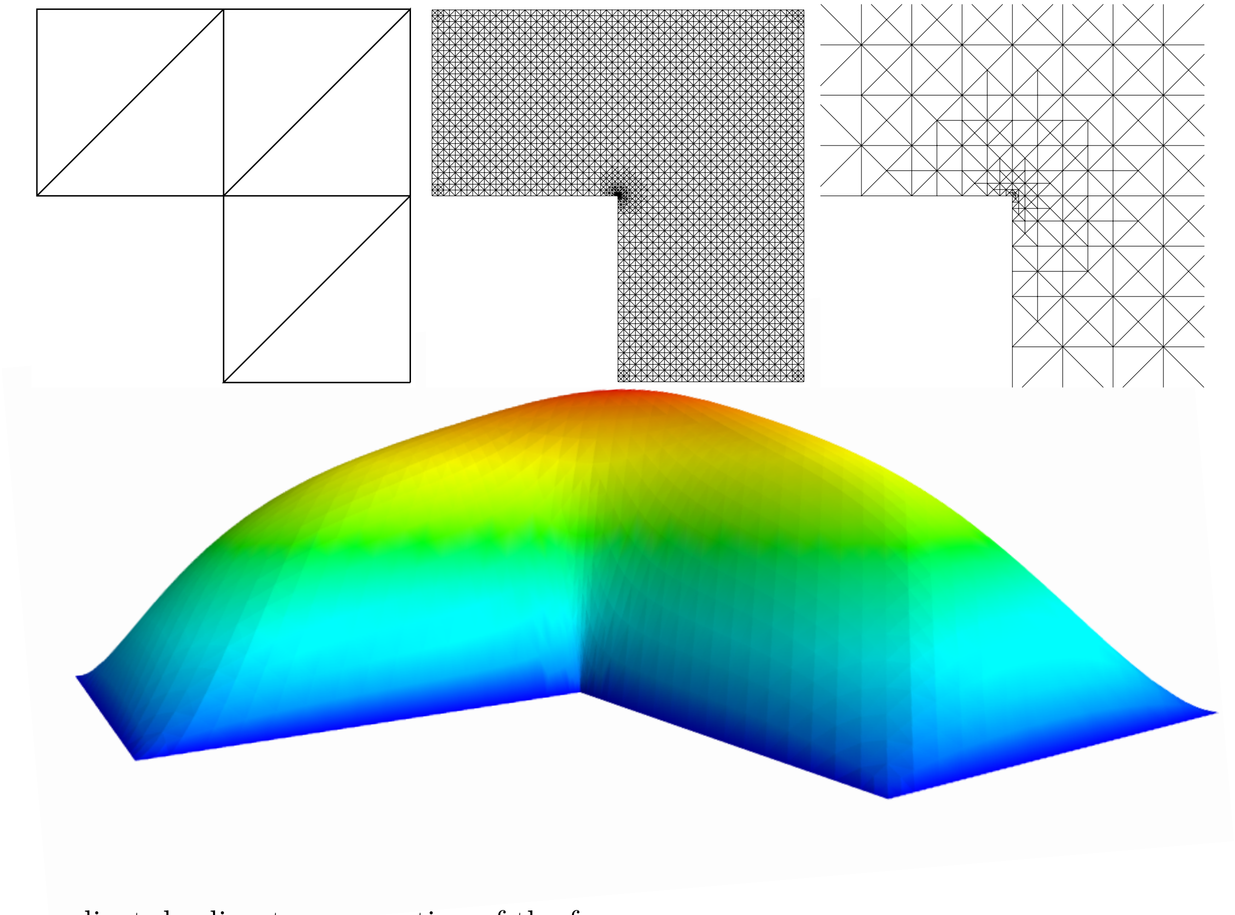
— continued on next page —

```
    # *** ESTIMATE step
    # Define cell and edge residuals
    R_T = f + div(grad(u_h))
    # get the normal to the cells
    n = V.cell().n
    R_dT = 2*avg(dot(grad(u_h), n))
    # Will use space of constants to localize indicator form
    Constants = FunctionSpace(mesh, "DG", 0)
    w = TestFunction(Constants)
    h = CellSize(mesh)
    # Assemble squared error indicators, eta_T^2, and store into a numpy array
    eta2 = assemble(h**2*R_T**2*w*dx + 4.*avg(h)*R_dT**2*avg(w)*dS) # dS is integral over interior edges only
    eta2 = eta2.array()
    # compute maximum and sum (which is the estimate for squared H1 norm of error)
    eta2_max = max(eta2)
    sum_eta2 = sum(eta2)
    # stop error estimate is less than tolerance
    if sum_eta2 < tolerance:
        print "Mesh %g: %d triangles, %d vertices, hmax = %g, hmin = %g, errest = %g" \
           % (i, mesh.num_cells(), mesh.num_vertices(), mesh.hmax(), mesh.hmin(), sqrt(sum_eta2))
        print "\nTolerance achieved.  Exiting."
        break

    # *** MARK step
    # Mark cells for refinement for which eta > frac eta_max for frac = .95, .90, ...;
    # choose frac so that marked elements account for a given part of total error
    frac = .95
    delfrac = .05
    part = .5
    marked = zeros(eta2.size, dtype=’bool’) # marked starts as False for all elements
    sum_marked_eta2 = 0. # sum over marked elements of squared error indicators
    while sum_marked_eta2 < part*sum_eta2:
      new_marked = (~marked) & (eta2 > frac*eta2_max)
      sum_marked_eta2 += sum(eta2[new_marked])
      marked += new_marked
      frac -= delfrac
    # convert marked array to a MeshFunction
    cells_marked = MeshFunction("bool", mesh, mesh.topology().dim())
    cells_marked.array()[:] = marked

    # *** REFINE step
    mesh = refine(mesh, cells_marked)
    plot(mesh, title="Mesh q" + str(i))
    print "Mesh %g: %d triangles, %d vertices, hmax = %g, hmin = %g, errest = %g" \
       % (i, mesh.num_cells(), mesh.num_vertices(), mesh.hmax(), mesh.hmin(), sqrt(sum_eta2))
    stdin.readline()

plot(mesh)
interactive()
```

10. Nonlinear problems

So far we have only discussed linear PDE. In many situations in which PDE models are applied, linear PDE are a simplification, which often is not sufficiently accurate. For example, in our model problem $-\operatorname{div} a \operatorname{grad} u = f$ modeling a steady-state temperature distribution, the thermal conductivity a might depend on the temperature giving a nonlinear equation $-\operatorname{div} a(u(x)) \operatorname{grad} u(x) = 0$. Dependence on the temperature gradient is possible as well, and we might have convection and source terms, which might also depend on the temperature

FIGURE 4.10. Adaptive solution of Poisson's equation by the FEniCS program on the preceding page. Shown are the input mesh, the computed adaptive mesh, and a blow-up of that mesh near the re-entrant corner, as well as the final solution.



or gradient, leading to an equation of the form

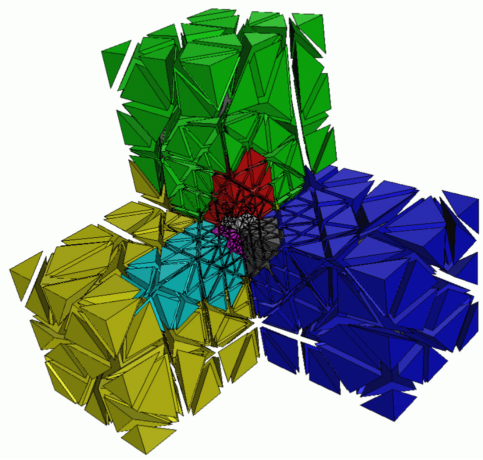
$$-\operatorname{div}[a(u(x), \operatorname{grad} u(x)) \operatorname{grad} u(x)] - f(u(x), \operatorname{grad} u(x)) = 0, \tag{4.22}$$

where a and f are functions of $n+1$ variables in n dimensions (or $2n+1$: they might depend explicitly on x as well). A PDE of the form (4.22) is called *quasilinear*. If a is independent of u and $\operatorname{grad} u$, so the only nonlinearity arises in the lower order terms in f, then it is called semilinear. As long as the coefficient a is everywhere positive (or a symmetric positive definite matrix), (4.22) is elliptic and we can hope to treat it by the sorts of finite element methods discussed heretofore. Some simple examples of such PDE are

$$-\Delta u + \lambda e^u = 0, \quad -\operatorname{div}(1 + e^{-u^2}) \operatorname{grad} u = 0, \quad -\operatorname{div} \frac{1}{\sqrt{1 + |\operatorname{grad} u|^2}} \operatorname{grad} u = 0.$$

The first of these, called Bratu's problem, is semilinear and arises in combustion modeling. The other two are quasilinear. The third is the minimal surface equation, satisfied by functions whose graphs are minimal area surfaces subject to their boundary conditions (like soap bubbles on a frame).

FIGURE 4.11. An adaptive mesh in 3-dimensions produced by Michael Holst using his MC code. (The colors relate to partitioning among processors for parallel computation.)



The theory needed to analyze nonlinear PDE, e.g., to prove existence, uniqueness, continuous dependence, and various qualitative behaviors, is extensive, diverse, often complex, and an area of active research. Many different approaches have been developed in order to address different equations. We will not consider these at all, but briefly consider how one might devise numerical methods to compute the solution to a nonlinear elliptic PDE, assuming that a locally unique solution exists ("locally unique" means that in some neighborhood of the solution no other solution exists). In this case, the general approach to computation is to approximate the solution of the nonlinear problem by solving a sequence of linear problems.

Consider the quasilinear PDE (4.22) subject (for example) to the Dirichlet boundary condition $u = g$ on $\partial\Omega$. We obtain a weak formulation just as in the linear case, by multiplying the equation by a test function v satisfying homogeneous Dirichlet boundary conditions and integrating over Ω by parts. Thus we obtain the nonlinear weak formulation: find u which is equal to g on $\partial\Omega$ and such that

$$F(u,v) := \int a(u, \operatorname{grad} u) \operatorname{grad} u \cdot \operatorname{grad} v \, dx - \int f(u, \operatorname{grad} u) v \, dx = 0, \tag{4.23}$$

for all test functions v vanishing on $\partial\Omega$. We might use the space $H^1(\Omega)$ as the space for the trial and test functions, as in the linear case, although, depending on the nonlinearity, more complicated function spaces may be needed to insure that the integrals all exist. This is an issue for the analysis of the numerical method, but need not concern us here where we will only discuss the formulation of algorithms. Note that the bivariate form $F(u, v)$ is not bilinear. It remains linear in v, but is nonlinear in u.

To solve the nonlinear problem, we use Galerkin's method as in the linear case. Thus we choose a finite dimensional space V_h for the trial and test functions (satisfying the boundary conditions), such as a finite element space based on a mesh and Lagrange finite elements. Then the Galerkin method seeks $u_h \in V_h$ such that

$$F(u_h, v) = 0 \text{ for all } v \in V_h. \tag{4.24}$$

If we choose a basis ϕ_i, $i = 1, \ldots, n$, for V_h and expand $u_h = \sum_{j=1}^n U_j \phi_j$, then the coefficients U_j may be determined from the system of equations

$$F(\sum_j U_j \phi_j, \phi_i) = 0, \quad i = 1, \ldots, n, \tag{4.25}$$

which is a (nonlinear) system of n equations in n unknowns.

10.1. Picard iteration. Let

$$b(w; u, v) = \int a(w, \operatorname{grad} w) \operatorname{grad} u \cdot \operatorname{grad} v \, dx - \int f(w, \operatorname{grad} w) v \, dx$$

so the form in (4.23) is $F(u, v) = b(u; u, v)$. If we fix some $w \in V_h$ the problem of finding $u_h(w) \in V_h$ such that

$$b(w; u_h(w), v) = 0 \text{ for all } v \in V_h,$$

is a standard linear finite element problem. This defines a mapping $w \mapsto u_h$ from V_h to itself. If u_h is a fixed point of this map, then it satisfies $b(u_h; u_h, v) = 0$ for all $v \in V_h$ which is the desired Galerkin equation (4.24). Thus we may try to solve (4.24) by fixed point iteration, with each iteration requiring the solution of a linear finite element system. This approach is called *Picard iteration.* Thus the basic iteration takes the form:

```
choose initial iterate $u_h^0 \in V_h$
for $i = 0, 1, \ldots$
  find $u_h^{i+1} \in V_h$ such that
      $\int a(u_h^i, \operatorname{grad} u_h^i) \operatorname{grad} u_h^{i+1} \cdot \operatorname{grad} v \, dx = \int f(u_h^i, \operatorname{grad} u_h^i) v \, dx dx, \quad v \in V_h.$
end
```

Thus, in each iteration we solve a linear finite element problem, where the coefficients depend on the previously computed iterate. For example, for the minimal surface equation, the nonlinear weak formulation is

$$\int \frac{1}{\sqrt{1 + |\operatorname{grad} u|^2}} \operatorname{grad} u \cdot \operatorname{grad} v \, dx = 0, \quad v \in \mathring{H}^1,$$

so the Picard iteration seeks $u_h^{i+1} \in V_h$ satisfying the Dirichlet boundary conditions and

$$\int \frac{1}{\sqrt{1 + |\operatorname{grad} u_h^i|^2}} \operatorname{grad} u_h^{i+1} \cdot \operatorname{grad} v \, dx = 0, \quad v \in \mathring{V}_h.$$

which is a standard linear finite element system.

The Picard iteration does not always converge, but it often does, especially if the initial guess is reasonably close to the exact solution. When it does converge, it typically does so with a linear rate of convergence to the solution of the nonlinear Galerkin equations. This can be quite slow, requiring many linear solves.

10.2. Newton iteration. We start by recalling what it means to linearize a system of n algebraic equations in n unknowns. We write the system as $G(u) = 0$ where $G : \mathbb{R}^n \to \mathbb{R}^n$ is some function (supposed smooth), and the solution u is sought in $\mathbb{R}^n$. To linearize around some $u^0 \in \mathbb{R}^n$, not too far from the solution u, we replace u in the equations with a perturbation $u^0 + \delta u$ of u^0, where $\delta u \in \mathbb{R}^n$ is expected to be small. Thus we want $G(u^0 + \delta u)$ to vanish, or nearly so. Expanding this quantity via Taylor's theorem gives

$$G(u^0 + \delta u) = G(u^0) + DG(u^0)\delta u + \cdots ,$$

where $DG(u^0)$ is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^n$ (its matrix is $\partial G_i/\partial u_j$ evaluated at u^0), and the dots indicate terms that are quadratic in δu. If we drop the quadratic terms and set the result equal to zero, we get a linear system of n equations in n unknowns to solve for δu:

$$DG(u^0)\delta u = -G(u^0).$$

Newton's method defines u^1 to be $u^0 + \delta u$, which is hopefully an improved approximation of the solution u. It continues by linearizing $G(u) = 0$ about u^1 to find u^2, etc. The typical behavior of Newton's method is that it converges if the initial iterate u^0 is chosen close enough to the solution u, and in that case the convergence is very fast—quadratic. It may, however, be difficult to find a suitably close initial iterate.

Just as for a nonlinear algebraic system, Newton's method may be applied to a nonlinear PDE, or to the finite element discretization of a nonlinear PDE. We can apply it directly to the nonlinear PDE (4.22), and then solve the resulting linear PDE by converting it into weak form, and then discretizing it with Galerkin's method. Alternatively, we can start with the nonlinear weak formulation (4.23), linearize that, obtaining a linear weak formulation, which we can discretize by Galerkin's method. A third possibility is to start with the nonlinear algebraic system (4.24) obtained by discretizing the nonlinear weak formulation, and implement Newton's method for this nonlinear algebraic system. It turns out that all three approaches are equivalent: the final discrete iterates computed are the same for all three. (After reading this section, try to prove this—it is a good test of your understanding.) We prefer to describe the middle alternative: linearization of the nonlinear weak formulation.

For simplicity, consider the case of (4.23) in which f vanishes, so the nonlinear weak formulation seeks $u \in H^1$ with given Dirichlet boundary values such that

$$F(u, v) := \int a(u, \operatorname{grad} u) \operatorname{grad} u \cdot \operatorname{grad} v \, dx = 0, \quad v \in \mathring{H}^1. \tag{4.26}$$

Here we are assuming that the coefficient $a = a(y, z)$ is a smooth real-valued function of a scalar variable y and vector variable z. Now suppose we have an approximation $u^0 \in H^1$ to u, which we suppose satisfies the boundary conditions. Now

$$a\big(u^0 + \delta u, \operatorname{grad}(u^0 + \delta u)\big) = a + \frac{\partial a}{\partial y}\,\delta u + \sum_{j=1}^{n} \frac{\partial a}{\partial z_j}\frac{\partial \delta u}{\partial x_j} + \cdots ,$$

where on the right-hand side a and its partial derivatives are evaluated at $(u^0, \operatorname{grad} u^0)$, and the dots represent terms which are quadratic or higher in δu and its derivatives. Thus the integrand of (4.26) becomes

$$a(u, \operatorname{grad} u) \operatorname{grad} u \cdot \operatorname{grad} v = a\big(u^0 + \delta u, \operatorname{grad}(u^0 + \delta u)\big) \operatorname{grad}(u^0 + \delta u) \cdot \operatorname{grad} v$$

$$= a \operatorname{grad} u^0 \cdot \operatorname{grad} v + a \operatorname{grad} \delta u \cdot \operatorname{grad} v + \frac{\partial a}{\partial y} \delta u \operatorname{grad} u^0 \cdot \operatorname{grad} v + \sum_{j=1}^{n} \frac{\partial a}{\partial z_j} \frac{\partial \delta u}{\partial x_j} \operatorname{grad} u^0 \cdot \operatorname{grad} v + \cdots .$$

$$A \operatorname{grad} \delta u \cdot \operatorname{grad} v + \delta u\, B \cdot \operatorname{grad} v + a \operatorname{grad} u^0 \cdot \operatorname{grad} v + \cdots ,$$

where A is the matrix-valued functions

$$A_{ij} = a + \frac{\partial a}{\partial z_j} \frac{\partial u^0}{\partial x_i},$$

and $B = (\partial a/\partial y) \operatorname{grad} u_0$, both evaluated at $(u^0, \operatorname{grad} u^0)$. Thus $\delta u \in \mathring{H}^1$ is determined by the problem

$$(4.27) \qquad \int (A \operatorname{grad} \delta u \cdot \operatorname{grad} v + \delta u\, B \cdot \operatorname{grad} v)\, dx = -\int a \operatorname{grad} u^0 \cdot \operatorname{grad} v\, dx, \quad v \in \mathring{H}^1,$$

which is the linear weak formulation of a PDE (with homogeneous Dirichlet boundary conditions, even though the nonlinear problem had inhomogeneous boundary conditions).

To implement Newton's method for the finite element method, we start with an approximation u_h^0 in the finite element space, satisfying the Dirichlet boundary conditions, we then define $\delta u_h \in \mathring{V}^h$ by the linear weak formulation (4.27) with the test function v restricted to $\mathring{V}_h$, and set $u_h^1 = u_h^0 + \delta u_h \in V_h$, as the next iterate.

10.3. Convergence of Galerkin's method for the minimal surface equation. We now study the error analysis for Galerkin's method for the minimal surface equation using piecewise linear finite elements. The analysis goes back to a paper of Johnson and Thomée from 1975, and is given as well in the book of Ciarlet on finite element methods. For simplicity we assume that the domain $\Omega \subset \mathbb{R}^2$ is polygonal and that there is a solution $u : \Omega \to \mathbb{R}$ belonging to $H^2(\Omega) \cap W^{1,\infty}(\Omega)$ satisfying the Dirichlet problem for the minimal surface equation:

$$-\operatorname{div} a(\operatorname{grad} u) \operatorname{grad} u = 0 \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega,$$

where

$$a : \mathbb{R}^2 \to \mathbb{R}, \quad a(z) = (1 + |z|^2)^{-1/2}.$$

The weak formulation characterizes $u \in H^1$ such that $u = g$ on $\partial\Omega$ by

$$\int a(\operatorname{grad} u) \operatorname{grad} u \cdot \operatorname{grad} v\, dx = 0, \quad v \in \mathring{H}^1(\Omega).$$

For the Galerkin method, we consider a shape-regular quasi-uniform family of triangulations $\mathcal{T}_h$ of Ω and let V_h be the corresponding space of Lagrange finite elements of degree 1. The Galerkin solution is determined as $u_h \in V_h$ such that u_h equals g at the boundary vertices and

$$(4.28) \qquad \int a(\operatorname{grad} u_h) \operatorname{grad} u_h \cdot \operatorname{grad} v\, dx = 0, \quad v \in \mathring{H}^1(\Omega).$$

Note that set u_h equal to g at boundary vertices is the same as requiring that $u_h = I_h$ on $\partial\Omega$ where $I_h : C(\bar{\Omega}) \to V_h$ is the interpolation operator.

The first question we should ask is whether the Galerkin equations, which can be viewed as a system of finitely many nonlinear algebraic equations, have a unique solution. This can be proven by taking a step backwards to the optimization problem that led to the minimal surface equations, namely minimizing the surface area

$$J(u) = \int_\Omega \sqrt{1 + |\operatorname{grad} u|^2}\,dx.$$

If we minimize $J(u_h)$ over all $u_h \in V_h$ satisfying the discrete boundary condition, at the minimum we obtain the Galerkin equations (4.28). A minimizer must exist, since $J(u_h) \to \infty$ as $u_h \to \infty$ (in any norm, as all norms are equivalent on V_h). In fact, by computing the Hessian of $J(u)$ one can check that it is convex, and so there exists a unique minimum.

The main question we wish to consider is an error estimate for $u - u_h$. We shall prove that there exists a constant C, which may depend on u and the shape constant and quasiuniformity constant of the mesh, but not otherwise, such that

$$\|u - u_h\|_1 \le Ch. \tag{4.29}$$

Note that this question of convergence of the Galerkin solution u_h to the exact solution u as the mesh is refined, has nothing to do with the question of convergence of the Picard iteration or Newton iteration to u_h.

We begin with a simple calculus lemma.

Lemma 4.8.

$$|a(\operatorname{grad} u) - a(\operatorname{grad} u_h)| \le a(\operatorname{grad} u)a(\operatorname{grad} u_h)|\operatorname{grad} u - \operatorname{grad} u_h| \ \ \textit{on}\ \Omega.$$

Proof. First we note that the real function $t \mapsto \sqrt{1+t^2}$ has derivative everywhere less than 1 in absolute value, so

$$|\sqrt{1+s^2} - \sqrt{1+t^2}| \le |s - t|, \quad s, t \in \mathbb{R}.$$

Now

$$\left|\frac{1}{\sqrt{1+s^2}} - \frac{1}{\sqrt{1+t^2}}\right| = \left|\frac{|\sqrt{1+s^2} - \sqrt{1+t^2}|}{\sqrt{1+s^2}\sqrt{1+t^2}}\right| \le \frac{|s-t|}{\sqrt{1+s^2}\sqrt{1+t^2}}.$$

Setting $s = |\operatorname{grad} u|$ and $t = |\operatorname{grad} u_h|$ gives the lemma. □

To prove (4.29), as usual we consider, instead of the error $u - u_h$ the difference between u_h and a representative of u in the subspace, which we naturally take to be the interpolant. Thus we want to bound $|I_h u - u_h|_1$ (the H^1 seminorm is sufficient, since $I_h u - u_h$ vanishes on the boundary). As a first step, we bound $I_h u - u_h$ in a slightly weaker norm, more closely related to the Galerkin method. Specifically, for each h, we define:

$$(v, w)_{1h} := \int_\Omega a(u_h)\operatorname{grad} v \cdot \operatorname{grad} w\,dx, \quad |v|_{1h} := \sqrt{(v,v)_{1h}}, \quad v, w \in H^1(\Omega).$$

Note that $|v|_h \le |v|_1$. In this notation we may write the Galerkin equations as

$$(u_h, v)_{1h} = 0, \quad v \in V_h,$$

while the weak formulation becomes

$$(u, v)_{1h} = \int [a(\operatorname{grad} u_h) - a(\operatorname{grad} u)]\operatorname{grad} u \cdot \operatorname{grad} v\,dx,$$

or

$$(I_h u, v)_{1h} = (I_h u - u, v)_{1h} + \int [a(\operatorname{grad} u_h) - a(\operatorname{grad} u)] \operatorname{grad} u \cdot \operatorname{grad} v \, dx,$$

We then subtract these two and take $v = I_h u - u_h$ to get

$$|I_h u - u_h|_{1h}^2 = (I_h u - u, I_h u - u_h)_{1h} + \int [a(\operatorname{grad} u_h) - a(\operatorname{grad} u)] \operatorname{grad} u \cdot \operatorname{grad}(I_h u - u_h) \, dx.$$

Calling the two terms on the right hand side T_1 and T_2, we bound the first using the Cauchy–Schwartz inequality: $T_1 \le |I_h u - u|_{1h} |I_h u - u_h|_{1h}$. For the second we use the lemma and find

$$T_2 \le \gamma(u) |u - u_h|_{1h} |I_h u - u_h|_{1h} \le \gamma(u) (|I_h u - u|_{1h} |I_h u - u_h|_{1h} + |I_h u - u_h|_{1h}^2)$$

where

$$\gamma(u) := \|a(\operatorname{grad} u) \operatorname{grad} u\|_{L^\infty} < 1.$$

Combining these equations gives

$$|I_h u - u_h|_{1h}^2 \le [1 + \gamma(u)] |I_h u - u|_{1h} |I_h u - u_h|_{1h} + \gamma(u) |I_h u - u_h|_{1h}^2,$$

whence

$$|I_h u - u_h|_{1h}^2 \le \frac{1 + \gamma(u)}{1 - \gamma(u)} |I_h u - u|_{1h} |I_h u - u_h|_{1h},$$

or

$$|I_h u - u_h|_{1h} \le \frac{1 + \gamma(u)}{1 - \gamma(u)} |I_h u - u|_{1h}.$$

With the triangle inequality this becomes

$$|u - u_h|_{1h} \le \left[1 + \frac{1 + \gamma(u)}{1 - \gamma(u)}\right] |I_h u - u|_{1h}.$$

Of course, for the interpolation error we have

$$|I_h u - u|_{1h} \le |I_h u - u|_1 \le C h \|u\|_2,$$

so altogether

$$|u - u_h|_{1h} \le C(u) h, \tag{4.30}$$

which is the desired estimate (4.29) except in a slightly weaker norm than the H^1 seminorm.

To show that this norm is equivalent to the full H^1 seminorm, it is sufficient to show that the coefficient $a(\operatorname{grad} u_h)$ is bounded below, or, equivalently, that the piecewise constant function $\operatorname{grad} u_h$ is bounded above (uniformly over all meshes). Now let K be any triangle in any of the meshes $\mathcal{T}_h$. The area of K satisfies $c^{-1} h^2 \le |K| \le c h^2$ where c depends on the quasiuniformity and shape regularity of the mesh. Then

$$\int_K \frac{|\operatorname{grad} u_h|^2}{\sqrt{1 + |\operatorname{grad} u_h|^2}} dx \le 2 \int_K \frac{|\operatorname{grad} u - \operatorname{grad} u_h|^2}{\sqrt{1 + |\operatorname{grad} u_h|^2}} dx + 2 \int_K \frac{|\operatorname{grad} u|^2}{\sqrt{1 + |\operatorname{grad} u_h|^2}} dx.$$

The first term on the right hand side is part of $|u - u_h|_{1h}^2$ and so is bounded by a u-dependent constant time h^2. The second is bounded by $|u|_{1,\infty}^2 |K|$, which is the same form. We have thus shown that there is a constant M, depending only on u and mesh regularity, such that

$$\int_K \frac{|\operatorname{grad} u_h|^2}{\sqrt{1 + |\operatorname{grad} u_h|^2}} dx \le M |K|$$

on every triangle K. Since u_h is piecewise linear, the integrand is constant on K, and this constant value cannot exceed M. But $t^2/\sqrt{1+t^2} \to \infty$ as $t \to \infty$, so it follows that $|\operatorname{grad} u_h|$ must remain bounded.

This concludes the proof of (4.29).

It is possible to use duality techniques to prove an $O(h^2)$ estimate for $\|u - u_h\|_{L^2}$. This was partially done in Johnson–Thomée 1975 and completed by Rannacher in 1977.

CHAPTER 5

Time-dependent problems

So far we have considered the numerical solution of elliptic PDEs. In this chapter we will consider some parabolic and hyperbolic PDEs.

1. Finite difference methods for the heat equation

In this section we consider the Dirichlet problem for the heat equation: find $u : \bar{\Omega} \times [0,T] \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t}(x,t) - \Delta u(x,t) = f(x,t), \quad x \in \Omega, \quad 0 \le t \le T, \tag{5.1}$$

$$u(x,t) = 0, \quad x \in \partial\Omega, \quad 0 \le t \le T. \tag{5.2}$$

Since this is a time-dependent problem, we also need an initial condition:

$$u(x,0) = u_0(x), \quad x \in \Omega.$$

For simplicity, we will assume $\Omega = (0,1) \times (0,1)$ is the unit square in $\mathbb{R}^2$ (or the unit interval in $\mathbb{R}$). Let us consider first discretization in space only, which we have already studied. Following the notations we used in Chapter 2, we use a mesh with spacing $h = 1/N$, and let Ω_h be the set of interior mesh points, Γ_h the set of boundary mesh points, and $\bar{\Omega}_h = \Omega_h \cup \Gamma_h$. The semidiscrete finite difference method is: find $u_h : \bar{\Omega}_h \times [0,T] \to \mathbb{R}$ such at

$$\begin{gathered}\frac{\partial u_h}{\partial t}(x,t) - \Delta_h u_h(x,t) = f(x,t), \quad x \in \Omega_h, \quad 0 \le t \le T, \\ u_h(x,t) = 0, \quad x \in \partial\Omega_h, \quad 0 \le t \le T. \\ u_h(x,0) = u_0(x), \quad x \in \Omega_h.\end{gathered}$$

If we let $U_{mn}(t) = u_h((mh,nh),t)$, then we may write the first equation as

$$U'_{mn}(t) - \frac{U_{m+1,n}(t) + U_{m-1,n}(t) + U_{m,n+1}(t) + U_{m,n-1}(t) - 4U_{mn}(t)}{h^2} = f_{mn}(t),$$
$$0 < m,n < N, \; 0 \le t \le T.$$

Thus we have a system of $(n-1)^2$ ordinary differential equations, with given initial conditions.

One could feed this system of ODEs to an ODE solver. But we shall consider simple ODE solution schemes, which are, after all, themselves finite difference schemes, and analyze them directly. We shall focus on three simple schemes, although there are much more sophisticated possibilities.

For a system of ODEs, find $u : [0,T] \to \mathbb{R}^m$ such that

$$u'(t) = f(t,u(t)), \quad 0 \le t \le T, \quad u(0) = u_0,$$

(where $f : [0, T] \times \mathbb{R}^m \to \mathbb{R}^m$, $u_0 \in \mathbb{R}^m$) the simplest discretization is Euler's method. For a given timestep $k > 0$, let $t_j = jk$, $j = 0, 1, \ldots$, and define $u_j = u_h(t_j) \in \mathbb{R}_m$ for $j = 0, 1, \ldots$ by $u_h(0) = u(0)$, and

$$\frac{u_{j+1} - u_j}{k} = f(t_j, u_j), \quad j = 0, 1, \ldots.$$

Explicitly,

$$u_{j+1} = u_j + kf(t_j, u_j), \quad j = 0, 1, \ldots.$$

An alternative method is the backward Euler method or implicit Euler method

$$\frac{u_{j+1} - u_j}{k} = f(t_{j+1}, u_{j+1}), \quad j = 0, 1, \ldots.$$

This method involves solving the algebraic system

$$u_{j+1} - kf(t_{j+1}, u_{j+1}) = u_j, \quad j = 0, 1, \ldots.$$

This is a linear or nonlinear algebraic system according to whether f is linear or nonlinear in u, i.e., according to whether the original ODE system is linear or nonlinear.

1.1. Forward differences in time. Now consider the application of Euler's method to the semidiscretized heat equation. We take the timestep $k = T/M$ for some integer M, so that the discrete time values are $0, k, 2k, \ldots, T = Mk$. Writing U^j_{mn} for $u_h((mh, nh), jk)$ we get the explicit method

$$\frac{U^{j+1}_{mn} - U^j_{mn}}{k} - (\Delta_h U)^j_{mn} = f^j_{mn}, \tag{5.3}$$

i.e.,

$$U^{j+1}_{mn} = U^j_{mn} + k[(\Delta_h U)^j_{mn} + f^j_{mn}], \quad 0 < m, n < N, \ j = 0, 1 \ldots.$$

This is called the forward-centered difference method for the heat equation, because it uses forward differences in time and centered differences in space.

We shall analyze the forward-centered scheme (5.3) as usual, by establishing consistency and stability. Let $u^j_{mn} = u((mh, nh), t_j)$ denote the restriction of the exact solution, the consistency error is just

$$E^j_{mn} := \frac{u^{j+1}_{mn} - u^j_{mn}}{k} - (\Delta_h u)^j_{mn} - f^j_{mn} \tag{5.4}$$

$$= \left[\frac{u^{j+1}_{mn} - u^j_{mn}}{k} - (\Delta_h u)^j_{mn}\right] - \left[\left(\frac{\partial u}{\partial t} - \Delta u\right)((mh, nh), jk)\right]. \tag{5.5}$$

In Chapter 2 we used Taylor's expansion to get

$$\left|(\Delta_h u)^j_{mn} - \Delta u((mh, nh), jk)\right| \le c_1 h^2,$$

where

$$c_1 = (\|\partial^4 u/\partial x^4\|_{L^\infty(\bar\Omega \times [0,T])} + \|\partial^4 u/\partial y^4\|_{L^\infty(\bar\Omega \times [0,T])})/12.$$

Even easier is

$$\left|\frac{u^{j+1}_{mn} - u^j_{mn}}{k} - \frac{\partial u}{\partial t}((mh, nh), jk)\right| \le c_2 k,$$

where $c_2 = \|\partial^2 u/\partial t^2 u\|_{L^\infty}/2$. Thus

$$|E^j_{mn}| \le c(k + h^2),$$

with $c = \max(c_1, c_2)$.

Next we establish a stability result. Suppose that a mesh function U^j_{mn} satisfies (5.3). We want to bound an appropriate norm of the mesh function in terms of an appropriate norm of the function f^j_{mn} on the right hand side. For the norm, we use the max norm:

$$\|U\|_{L^\infty} = \max_{0\le j\le M} \max_{0\le m,n\le N} |U^j_{mn}|.$$

Write $K^j = \max_{0\le m,n\le N} |U^j_{mn}|$, and $F^j = \max_{0\le m,n\le N} |f^j_{mn}|$. From (5.3), we have

$$U^{j+1}_{mn} = (1 - \frac{4k}{h^2})U^j_{mn} + \frac{k}{h^2}(U^j_{m-1,n} + U^j_{m+1,n} + U^j_{m,n-1} + U^j_{m,n+1}) + kf^j_{mn}.$$

Now we make the assumption that $4k/h^2 \le 1$. Then the 5 coefficients of U on the right hand side are all nonnegative numbers which add to 1, so it easily follows that

$$K^{j+1} \le K^j + kF^j.$$

Therefore $K^1 \le K^0 + kF^0$, $K^2 \le K^0 + k(F^0 + F^1)$, etc. Thus $\max_{0\le j\le M} K^j \le K^0 + T \max_{0\le j<M} F^j$, where we have used that $kM = T$. Thus, if U satisfies (5.3), then

$$\|U\|_{L^\infty} \le \|U^0\|_{L^\infty(\Omega_h)} + T\|f\|_{L^\infty}, \tag{5.6}$$

which is a stability result. We have thus shown stability under the condition that $k \le h^2/4$. We say that the forward-centered difference method for the heat equation is *conditionally stable.*

Now we apply this stability result to the error $e^j_{mn} = u^j_{mn} - U^j_{mn}$, which satisfies

$$\frac{e^{j+1}_{mn} - e^j_{mn}}{k} - (\Delta_h e)^j_{mn} = E^j_{mn},$$

with E the consistency error (so $\|E\|_{L^\infty} \le c(k + h^2)$). Note that $E^0_{mn} = 0$, so the stability result gives

$$\|e\|_{L^\infty} \le T\|E\|_{L^\infty}.$$

Using our estimate for the consistency error, we have proven the following theorem.

THEOREM 5.1. *Let u solve the heat equation* (5.1)*, and let U^j_{mn} be determined by the forward-centered finite difference method with mesh size $h = 1/N$ and timestep $k = T/M$. Suppose that $k \le h^2/4$. Then*

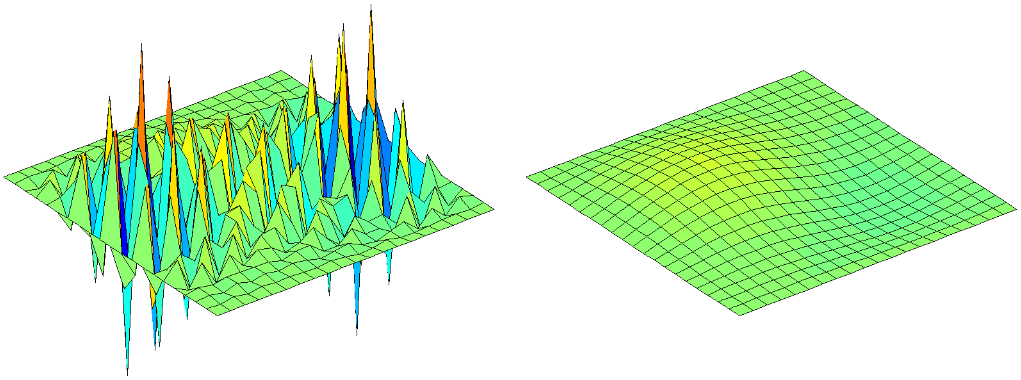
$$\max_{0\le j\le M} \max_{0\le m,n\le N} |u((mh, nh), jk) - U^j_{mn}| \le C(k + h^2),$$

where $C = cT$ with c as above.

In short, $\|u - u_h\|_{L^\infty} = O(k + h^2)$.

The requirement that $4k/h^2 \le 1$ is not needed for consistency. But it is required to prove stability, and a simple example shows that it is necessary for convergence. See Figure 5.1. An even simpler example would be the 1D case, for which we obtain stability under the condition $k \le h^2/2$.

FIGURE 5.1. Centered differences in space and forward differences in time for the Dirichlet problem for the heat equation on the unit square. The mesh size is $h = 1/20$, and the timestep is $k = 1/800$ on the left and $k = 1/1600$ on the right. On the left we show the result after 18 time steps, i.e., $t = 18/800 = .0225$. On the right we show the result after 36 time steps (the same time). The computation on the right remains stable for long times.



1.2. Backward differences in time. Next we consider of backward differences in time, i.e., the backward Euler method to solve the semidiscrete system. Then (5.3) becomes

$$\frac{U^{j+1}_{mn} - U^{j}_{mn}}{k} - (\Delta_h U)^{j+1}_{mn} = f^{j+1}_{mn}. \tag{5.7}$$

This is now an *implicit* method. Instead of writing down u^{j+1}_h explicitly in terms of u^j_h, we need to solve for the vector of its values, U^{j+1}_{mn}, from a system of linear equations:

$$U^{j+1}_{mn} - k\Delta_h U^{j+1}_{mn} = U^{j}_{mn} + k f^{j+1}_{mn}, \quad 0 < m, n < N,$$

or, written out,

$$(1+4\mu)U^{j+1}_{mn} - \mu(U^{j+1}_{m+1,n} + U^{j+1}_{m-1,n} + U^{j+1}_{m,n+1} + U^{j+1}_{m,n-1}) = U^{j}_{mn} + k f^{j+1}_{mn}, \tag{5.8}$$

with $\mu = k/h^2$. This is a sparse system of $(N-1)^2$ equations in $(N-1)^2$ unknowns with the same sparsity pattern as Δ_h. Since $-\Delta_h$ is symmetric and positive definite, this system, whose matrix is $I - k\Delta_h$, is as well.

REMARK. The computational difference between explicit and implicit methods was very significant before the advent of fast solvers, like multigrid. Since such solvers reduce the computational work to the order of the number of unknowns, they are not so much slower than explicit methods.

Now suppose that (5.8) holds. Let K^j again denote the maximum of $|U^j_{mn}|$. Then there exists m, n such that $K^{j+1} = sU^{j+1}_{mn}$ where $s = \pm 1$. Therefore $sU^{j+1}_{mn} \geq sU^{j+1}_{m+1,n}$ and similarly for each of the other three neighbors. For this particular m, n, we multiply (5.8) by s and obtain

$$K^{j+1} = sU^{j+1}_{mn} \leq sU^{j}_{mn} + skf^{j+1}_{mn} \leq K^j + k\|f^{j+1}\|_{L^\infty}.$$

From this we get the stability result (5.6) as before, but now the stability is unconditional: it holds for any $h, k > 0$. We immediately obtain the analogue of the convergence theorem Theorem 5.1 for the backward-centered method.

THEOREM 5.2. *Let u solve the heat equation (5.1), and let U^j_{mn} be determined by the backward-centered finite difference method with mesh size $h = 1/N$ and timestep $k = T/M$. Then*

$$\max_{0\le j\le M}\max_{0\le m,n\le N}|u((mh,nh),jk) - U^j_{mn}| \le C(k+h^2),$$

where $C = cT$ with c as above.

1.3. Fourier analysis. As we did for the Poisson problem, we can use Fourier analysis to analyze difference schemes for the heat equation. Recall that for a mesh function v on Ω_h we defined the norm

$$\|u\|_h^2 = h^2\sum_{m=1}^{N-1}\sum_{n=1}^{N-1}|u(mh,nh)|^2,$$

and the corresponding inner product. We then showed that $-\Delta_h$ had an orthogonal basis of eigenfunctions ϕ_{mn} with corresponding eigenvalues satisfying

$$2\pi^2 \approx \lambda_{1,1} \le \lambda_{mn} \le \lambda_{N-1,N-1} < 8/h^2.$$

Consider now the forward-centered difference equations for the heat equation $\partial u/\partial t = \Delta u$, with homogeneous Dirichlet boundary data, and given initial data u_0 (for simplicity we assume $f \equiv 0$). We have

$$U^{j+1}_{mn} = GU^j_{mn},$$

where the $G = I + k\Delta_h$. By iteration, we have

$$\|U^j\| \le \|G\|^j\|U^0\|.$$

Thus we have L^2 stability if and only if the spectral radius of G is bounded by 1. Now the eigenvalues of G are $\mu_{mn} = 1 - k\lambda_{mn}$, so they satisfy $1 - 8k/h^2 < \mu_{mn} < 1$, so the spectral radius condition is satisfies if $1 - 8k/h^2 \ge -1$, i.e., $k \le h^2/4$. In this way, Fourier analysis leads us to the same conditional stability condition we obtained above.

If we consider instead the backward-centered difference scheme, then the corresponding operator G is $G = (I + k\Delta_h)^{-1}$ with eigenvalues $(1 + k\lambda_{mn})^{-1}$, which has spectral radius bounded by 1 for any $k > 0$. Thus we obtain unconditional stability.

1.4. Crank–Nicolson. Although we are free to choose the timestep and space type as we like for the backward-centered method, accuracy considerations still indicate that we should take $k = O(h^2)$, so very small timesteps. It is natural to seek a method which is second order in time as well as space. If we use the trapezoidal method for time discretization, the resulting fully discrete scheme is called the Crank–Nicolson method. It is an implicit method given by

$$\frac{U^{j+1}_{mn} - U^j_{mn}}{k} - \frac{1}{2}[(\Delta_h U)^{j+1}_{mn} + (\Delta_h U)^j_{mn}] = \frac{1}{2}[f^j_{mn} + f^{j+1}_{mn}].$$

We leave it as an exercise to show that the Crank–Nicolson scheme is unconditionally stable scheme with respect to the L^2 norm, with error $O(h^2 + k^2)$.

2. Finite difference methods for the advection equation

In this section we consider finite difference methods for the 1D advection equation, which seeks $u(x,t)$ satisfying

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0.$$

This is the simplest hyperbolic equation. The solutions profile given by the initial data $u_0(x)$ simply travels to the right with speed 1 without changing shape: $u(x,t) = u_0(x-t)$. Thus, if the problem is posed on the whole real line then the initial value $u_0(x)$ determines a unique solution. If x ranges over a finite interval than we need to impose a Dirichlet boundary condition on the left end point of the interval, but not on the right. Note that we have made a tiny simplification in assuming the wave speed is one. With little extra effort we could handle any constant wave velocity c, positive or negative, i.e., the equation $\partial u/\partial t + c\,\partial u/\partial x = 0$.

Although the 1D advection equation is particularly simple, its numerical analysis has a lot in common with the numerical analysis of the 1D wave equation and hyperbolic equations in higher dimensions.

2.1. The CFL condition. A simple explicit finite difference method uses forward differences in time, and either forward, backward, or centered differences in space. In this context, the use of backward differences in time (which will turn out to be the more stable) is called *upwind differences* (if the velocity were negative, so the wave travels to the left, the upwind difference would be the forward difference). Thus the forward difference/upwind difference method is

$$\frac{U_n^{j+1} - U_n^j}{k} + \frac{U_n^j - U_{n-1}^j}{h} = 0,$$

giving the explicit update formula

$$U_n^{j+1} = (1 - \frac{k}{h})U_n^j + \frac{k}{h}U_{n-1}^j. \tag{5.9}$$

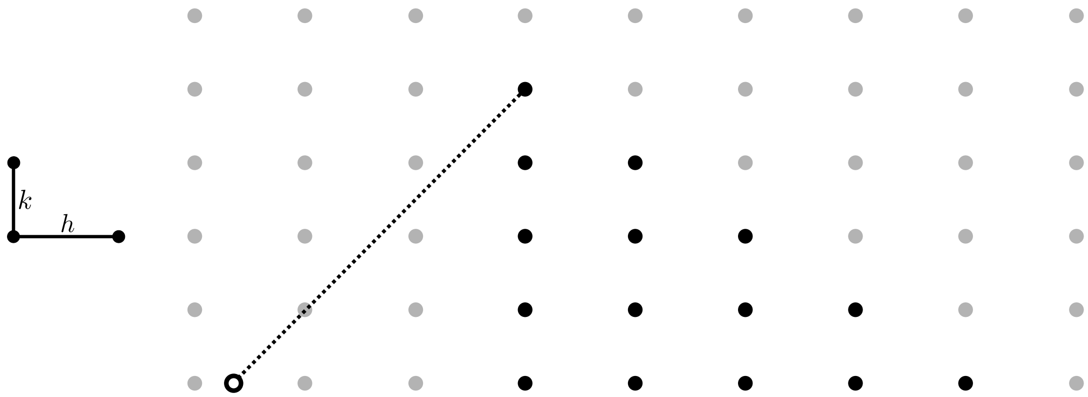
Here, of course, $U_n^j = u_h(nh, jk)$, where h is the mesh size and k is the time step.

By contrast the forward difference/downwind difference method would be

$$U_n^{j+1} = (1 + \frac{k}{h})U_n^j - \frac{k}{h}U_{n+1}^j. \tag{5.10}$$

The latter method is never used for the very good reason that it is not convergent. To see this, consider the discrete solution u_h at a grid point x_0 and time step t_0. From (5.10), $u_h(x_0,t_0)$ is determined by $u_h(x_0,t_0-k)$ and $u_h(x_0+h,t_0-k)$. These are, in turn, determined by $u_h(x_0,t_0-2k)$, $u_h(x_0-h,t_0-2k)$, and $u_h(x_0-2h,t_0-2k)$. Continuing in this way, we see that $u_h(x_0,t_0)$ is determined by the initial data u_0 only at the grid point x_0 and points to the right of x_0. But for the exact solution, we have $u(x_0,t_0) = u_0(x_0-t_0)$, i.e., the solution depends on the initial data at a point to the left of x_0. Imagine now that we pick initial data which is equal to 1 at $x_0 - t_0$ but equal to zero for $x \geq x_0$. Then $u_h(x_0,t_0)$ will vanish for any choice of h and k, and so will not converge to $u(x_0,t_0) = 1$. This argument is quite general. Given h, k, and a particular grid point, we define the *numerical domain of dependence* of the numerical solution as the set grid points for which the value of the initial data affects the value of the numerical solution at the given point. Then a necessary condition for the convergence of a numerical method is that, as h and k are brought to zero, the numerical domain of dependence must come arbitrarily close to every point in the domain of dependence of the true solution. In short, the numerical domain of dependence must encompass the true domain of dependence in the limit. This is a *necessary condition* for convergence, pointed out in 1928 in a paper of Courant, Friedrichs, and Lewy, and known as the *CFL condition.*

FIGURE 5.2. The downwind difference scheme, whose stencil is shown at left, does not fulfil the CFL condition, so cannot be convergent.



We can apply the CFL condition to the forward/upwind difference method as well. In this case we see that its satisfaction depends on a relation between h and k. Suppose that $k = \mu h$ as h, k tend to zero for some positive constant μ (so now $\mu = k/h$, while, when we analyzed the heat equation, we set $\mu = k/h^2$). If $\mu \le 1$, then the CFL condition holds, while if $\mu > 1$, the CFL condition is violated and the method definitely does not converge.

A third method is the forward/central difference method:

$$\frac{U_n^{j+1} - U_n^j}{k} + \frac{U_{n+1}^j - U_{n-1}^j}{2h} = 0.$$

This satisfies the CFL condition under the same condition as the forward/upwind method: $\mu \le 1$.

2.2. Fourier analysis. We now apply Fourier analysis to three methods above. For simplicity we will consider periodic solutions on an interval, to avoid dealing with boundary conditions. We also take the period to be 2π, which makes the eigenfunction simpler. Thus we are studying solutions of the advection equation which are defined on the whole real line by are 2π-periodic, and this determined by their restriction to the interval $(0, 2\pi)$ (or any other interval of length 2π.

To discretize we define the mesh size $h = 2\pi/N$ for some positive integer N, and let $I_h = \{\, nh \,|\, n \in \mathbb{Z} \,\}$ denote the set of all mesh points. Let $C(I_h)$ denote the set of mesh functions $u : I_h \to \mathbb{C}$ which are 2π-periodic. Such a function has an obvious set of degrees of freedom: we may assign any complex numbers for its values at $0, h, 2h, \dots, (N-1)h$, and this determines it uniquely. Thus the dimension of $C(I_h)$ over the complex numbers is

N. We take as norm of $v \in C(I_h)$, $\|v\|^2 = h\sum_{n=0}^{N-1} h|v(nh)|^2$. The corresponding norm is $\langle v, w\rangle = h\sum v(nh)\bar{w}(nh)$.

Now we introduce a Fourier basis for $C(I_h)$. The formula for basis functions is a bit simpler than we used before, because of the choice of period 2π. This would allow us to use the restrictions to the mesh of the functions $\sin mx$ and $\cos mx$ as the basis. But it is slightly more convenient to allow complex-valued functions and use as the basis functions the complex exponentials $\phi_m \in C(I_h)$ defined by

$$\phi_m(x) = e^{imx}, \quad x \in I_h.$$

Since all the values of ϕ_m have modulus 1, we have $\|\phi_m\| = 1$. We will see shortly that $\langle \phi_m, \phi_n\rangle = 0$ if $0 \le m < n < N$, so the ϕ_m. These facts togethern show that the ϕ_m for $0 \le m < N$ form an orthonormal basis for $C(I_h)$.

Define the *shift operators* $K_+, K_- : C(I_h) \to C(I_h)$ by

$$(K_+v)(x) = v(x+h), \quad K_-v(x) = v(x-h), \qquad x \in I_h.$$

We may write the difference methods above in terms of the shift operators. For example, writing u_h^j for $u_h(\,\cdot\,, jk) \in C(I_h)$, the forward/upwind method (5.9) becomes

$$u_h^{j+1} = (1 - \frac{k}{h})u_h^j + \frac{k}{h}K_-u_h^j = [(1-\frac{k}{h})I + \frac{k}{h}K_-]u_h^j.$$

The operator (matrix) $G = [(1-\frac{k}{h})I + \frac{k}{h}K_-]$ is the *amplification matrix* for the method, and we have, by iteration, that $u_h^j = G^j u_h^0$.

The utility of the Fourier basis is that it consists of eigenfunctions of the shift operator. Indeed we immediately see that

$$K_\pm\phi_m(x) = \phi_m(x \pm h) = e^{im(x\pm h)} = e^{\pm imh}e^{imx} = e^{\pm imh}\phi_m(x),$$

so ϕ_m is an eigenfunction of $K_\pm$ with eigenvalue $e^{\pm imh}$. (The fact that the ϕ_m are eigenfunctions of the symmetric operator $K_+ + K_-$ can be used to show that they are orthonormal.)

Since the amplification matrix G is a linear combination of the identity and K_-, the ϕ_m are eigenfunctions, and its eigenvalues are the numbers

$$\mu_m = (1-\mu) + \mu e^{-imh},$$

where $\mu = k/h > 0$. Note that the μ_m all belong to a circle in the complex plane with center $1-\mu$ and radius μ. It is then easy to see that all the $|\mu_m| \le 1$ if $\mu \le 1$ and not otherwise. Note that the forward/upwind method is thus stable if and only if $k \le h$, i.e., exactly when the CFL condition is satisfied.

Consider next the forward/centered difference method

$$u_h^{j+1} = u_h^j - \frac{k}{2h}(K_+u_h^j - K_-u_h^j) = [I - \frac{k}{2h}(K_+ - K_-)]u_h^j.$$

The eigenvalues of the amplification matrix are thus

$$1 - \frac{\mu}{2}(e^{imh} - e^{-imh}) = 1 - i\mu\sin mh,$$

so the eigenvalues are complex numbers with real part 1 and imaginary part between $-\mu$ and μ. The largest modulus exceeds 1 for any $\mu > 0$, so the forward/centered difference method is *unconditionally unstable* for the advection equation, despite the fact that it satisfies the

CFL condition whenever $k \le h$. Thus we cannot use the CFL condition as a way to ensure stability: it is necessary, but not sufficient.

There is a way to restore stability to the forward/centered difference scheme: we leave the centered difference in space, but change the forward difference in time to

$$\frac{\partial u}{\partial t}(nh, jk) \approx \frac{U_n^{j+1} - (U_{n-1}^j + U_{n+1}^j)/2}{k}.$$

This gives the *Lax–Friedrichs method*:

$$\frac{U_n^{j+1} - (U_{n-1}^j + U_{n+1}^j)/2}{k} + \frac{U_{n+1}^j - U_{n-1}^j}{2h} = 0.$$

The amplification matrix is then

$$G = \frac{1+\mu}{2} e^{-imh} + \frac{1-\mu}{2} e^{imh},$$

so the eigenvalues are $\cos mh - i\mu \sin mh$, which lie on an ellipse centered at the origin with principal radii 1 and μ. They thus remain within the unit circle if and only if $\mu \le 1$, that is, $k \le h$. So, like the forward/upwind method, the Lax-Friedrichs method is conditionally stable, with the condition coinciding with the CFL condition for the method, namely $k \le h$.

3. Finite element methods for the heat equation

In this section we consider the initial boundary value problem for the heat equation

$$\frac{\partial u}{\partial t}(x,t) - \operatorname{div} a(x) \operatorname{grad} u(x,t) = f(x,t), \quad x \in \Omega, \quad 0 \le t \le T,$$
$$u(x,t) = 0, \quad x \in \partial\Omega, \quad 0 \le t \le T,$$
$$u(x,0) = u_0(x), \quad x \in \Omega.$$

We have allowed the thermal conductivity a to be variable, assuming only that it is bounded above and below by positive constants, since this will cause no additional complications. We could easily generalize further, allowing a variable specific heat (coefficient of $\partial u/\partial t$), lower order terms, and different boundary conditions.

Now we consider finite elements for spatial discretization. To derive a weak formulation, we multiply the heat equation (5.1) by a test function $v(x) \in \mathring{H}^1(\Omega)$ and integrate over Ω. This gives

$$\int_\Omega \frac{\partial u}{\partial t}(x,t) v(x)\, dx + \int_\Omega a(x) \operatorname{grad} u(x,t) \cdot \operatorname{grad} v(x)\, dx = \int_\Omega f(x,t) v(x)\, dx, \quad 0 \le t \le T.$$

Writing $\langle \cdot\,, \cdot \rangle$ for the $L^2(\Omega)$ inner product and $b(w,v) = \int_\Omega a \operatorname{grad} w \cdot \operatorname{grad} v\, dx$, we may write the weak formulation as

$$\langle \frac{\partial u}{\partial t}, v \rangle + b(u,v) = \langle f, v \rangle, \quad v \in \mathring{H}^1(\Omega), \quad 0 \le t \le T.$$

For the finite element method it is often useful to think of $u(x,t)$ as a function of t taking values in functions of x. Specifically, we may think of u mapping $t \in [0,T]$ to $u(\cdot,t) \in \mathring{H}^1(\Omega)$. Specifically, we may seek the solution $u \in C^1([0,T], \mathring{H}^1(\Omega))$, which means that $u(\cdot,t) \in \mathring{H}^1(\Omega)$ for each t, that u is differentiable with respect to t, and that $\partial u(\cdot,t)/\partial t \in \mathring{H}^1(\Omega)$ for all t. Thus when we write $u(t)$, we mean the function $u(\cdot,t) \in \mathring{H}^1(\Omega)$.

Now let $V_h \subset \mathring{H}^1(\Omega)$ denote the usual space of Lagrange finite elements of degree r with respect to a triangulation of Ω. For a *semidiscrete finite element approximation* we seek u_h mapping $[0,T]$ into V_h, i.e., $u_h \in C^1([0,T],V_h)$, satisfying

$$\langle \frac{\partial u_h}{\partial t}, v\rangle + b(u_h,v) = \langle f, v\rangle, \quad v \in V_h, \quad 0 \le t \le T. \tag{5.11}$$

We also need to specify an initial condition for u_h. For this we choose some $u_h^0 \in V_h$ which approximates u_0 (common choices are the L^2 projection or the interpolant).

We now show that this problem may be viewed as a system of ODEs. For this let ϕ_i, $1 \le i \le D$ be a basis for V_h (for efficiency we will choose a local basis). We may then write

$$u_h(x,t) = \sum_{j=1}^{D} \alpha_j(t)\phi_j(x).$$

Plugging this into (5.11) and taking the test function $v = \phi_i$, we get

$$\sum_j \langle \phi_j, \phi_i\rangle\, \alpha_j'(t) + \sum_j b(\phi_j, \phi_i)\, \alpha_j(t) = \langle f, \phi_i\rangle.$$

In terms of the *mass matrix*, stiffness matrix, and load vector:

$$M_{ij} = \langle \phi_j, \phi_i\rangle, \quad A_{ij} = \langle \phi_j, \phi_i\rangle, \quad F_i(t) = \langle f, \phi_i\rangle,$$

this can be written

$$M\alpha'(t) + A\alpha(t) = F(t), \quad 0 \le t \le T.$$

This is a system of linear ODEs for the unknown coefficients $\alpha = (\alpha_j(t))$. The initial condition $u_h(0) = u_h^0$ can be written $\alpha(0) = \alpha^0$ where $u_h^0 = \sum_j \alpha_j^0 \phi_j$.

We now turn to *fully discrete approximation* using the finite element method for discretization in space, and finite differences for discretization in time. Consider first using Euler's method for time discretization. This leads to the system

$$M\frac{\alpha^{j+1} - \alpha^j}{k} + A\alpha^j = F^j,$$

or

$$M\alpha^{j+1} = M\alpha^j + k(-A\alpha^j + F^j).$$

Notice that for finite elements this method is not truly explicit, since we have to solve an equation involving the mass matrix at each time step.

The backward Euler's method

$$M\frac{\alpha^{j+1} - \alpha^j}{k} + A\alpha^{j+1} = F^{j+1},$$

leads to a different linear system at each time step:

$$(M + kA)\alpha^{j+1} = M\alpha^j + kF^{j+1},$$

while Crank–Nicolson would give

$$(M + \frac{k}{2}A)\alpha^{j+1} = (M - \frac{k}{2}A)\alpha^j + \frac{k}{2}(F^j + F^{j+1}).$$

3.1. Analysis of the semidiscrete finite element method. Before analyzing a fully discrete finite element scheme, we analyze the convergence of the semidiscrete scheme, since it is less involved. The key to the analysis of the semidiscrete finite element method is to compare u_h not directly to u, but rather to an appropriate representative $w_h \in C^1([0,T],V_h)$. For w_h we choose the *elliptic projection* of u, defined by

$$b(w_h,v)=b(u,v),\quad v\in V_h,\quad 0\le t\le T. \tag{5.12}$$

From our study of the finite element method for elliptic problems, we have the L^2 estimate

$$\|u(t)-w_h(t)\|\le ch^{r+1}\|u(t)\|_{r+1},\quad 0\le t\le T. \tag{5.13}$$

If we differentiate (5.12), we see that $\partial w_h/\partial t$ is the elliptic projection of $\partial u/\partial t$, so

$$\|\frac{\partial u}{\partial t}(t)-\frac{\partial w_h}{\partial t}(t)\|\le ch^{r+1}\|\frac{\partial u}{\partial t}(t)\|_{r+1},\quad 0\le t\le T.$$

Now

$$\begin{aligned}\langle\frac{\partial w_h}{\partial t},v\rangle+b(w_h,v)&=\langle\frac{\partial w_h}{\partial t},v\rangle+b(u,v)\\&=\langle\frac{\partial(w_h-u)}{\partial t},v\rangle+\langle f,v\rangle,\quad v\in V_h,\quad 0\le t\le T.\end{aligned} \tag{5.14}$$

Let $y_h=w_h-u_h$. Subtracting (5.11) from (5.14), we get

$$\langle\frac{\partial y_h}{\partial t},v\rangle+b(y_h,v)=\langle\frac{\partial(w_h-u)}{\partial t},v\rangle,\quad v\in V_h,\quad 0\le t\le T.$$

Now, for each t we choose $v=y_h(t)\in V_h$. Note that for any function $y\in C^1([0,T];L^2(\Omega))$,

$$\|y\|\frac{d}{dt}\|y\|=\frac{1}{2}\frac{d}{dt}\|y\|^2=\langle\frac{\partial y}{\partial t},y\rangle.$$

Thus we get

$$\|y_h\|\frac{d}{dt}\|y_h\|+b(y_h,y_h)=\langle\frac{\partial(w_h-u)}{\partial t},y_h\rangle\le\|\frac{\partial(w_h-u)}{\partial t}\|\|y_h\|, \tag{5.15}$$

so

$$\frac{d}{dt}\|y_h\|\le\|\frac{\partial(w_h-u)}{\partial t}\|\le ch^{r+1}\|\frac{\partial u}{\partial t}(t)\|_{r+1}.$$

This holds for each t. Integrating over $[0,t]$, we get

$$\|y_h(t)\|\le\|y_h(0)\|+ch^{r+1}\|\frac{\partial u}{\partial t}\|_{L^1([0,T];H^{r+1}(\Omega))}.$$

For $y_h(0)$ we have

$$\|y_h(0)\|=\|w_h(0)-u_h(0)\|\le\|w_h(0)-u(0)\|+\|u_0-u_h(0)\|\le ch^{r+1}\|u_0\|_{r+1}+\|u_0-u_h(0)\|.$$

Thus, assuming that the exact solution is sufficiently smooth and the initial data $u_h(0)$ is chosen so that $\|u_0-u_h(0)\|=O(h^{r+1})$, we have

$$\|y_h\|_{L^\infty([0,T];L^2(\Omega))}=O(h^{r+1}).$$

Combining this estimate with the elliptic estimate (5.13) we get an estimate on the error

$$\|u-u_h\|_{L^\infty([0,T];L^2(\Omega))}=O(h^{r+1}).$$

Remark. We can put this analysis into the framework of consistency and stability introduced in Section 3. We take our discrete solution space X_h as $C^1([0,T];V_h)$, and the discrete operator $L_h : X_h \to Y_h := C([0,T];V_h^*)$ is

$$(L_h u_h)(v) = \langle \frac{\partial u_h}{\partial t}, v\rangle + b(u_h, v), \quad u_h \in X_h, \quad v \in V_h, \quad 0 \le t \le T.$$

Thus our numerical method is to find $u_h \in X_h$ such that $L_h u_h = F_h$, where

$$F_h(v) = \int f v\, dx, \quad v \in V_h, \quad 0 \le t \le T.$$

As a representative $U_h \in X_h$ of the exact solution u we use the elliptic projection w_h. Then the consistency error is given by

$$E(v) := \langle \frac{\partial w_h}{\partial t}, v\rangle + b(w_h, v) - \langle f, v\rangle, \quad v \in V_h, \quad 0 \le t \le T.$$

In the first part of our analysis we showed that

$$E(v) = \langle \frac{\partial (w_h - u)}{\partial t}, v\rangle,$$

so $|||E||| = O(h^{r+1})$, where the norm we use on Y_h is

$$|||E||| = \int_0^T \sup_{0\ne v\in V_h} \frac{|E(v)|}{\|v\|}\, dt.$$

The second part of the analysis was a stability result. Essentially we showed that if $u_h \in X_h$ and $F_h \in Y_h$ satisfy $L_h u_h = F_h$, then

$$\max_{0\le t\le T} \|u_h\| \le \|u_h(0)\| + |||F_h|||.$$

Remark. In the case of finite elements for elliptic problems, we first got an estimate in H^1, then an estimate in L^2, and I mentioned that there are others possible. In the case of the parabolic problem, there are many other estimates we could derive in different norms in space or time or both. For example, by integrating (5.15) in time we get that $\|y_h\|_{L^2([0,T];H^1(\Omega))} = O(h^{r+1})$. For the elliptic projection we have $\|u - w_h\|_{H^1(\Omega)} = O(h^r)$ for each t, so the triangle inequality gives $\|u - u_h\|_{L^2([0,T];H^1(\Omega))} = O(h^r)$.

3.2. Analysis of a fully discrete finite element method. Now we turn to the analysis of a fully discrete scheme: finite elements in space and backward Euler in time. Writing u_h^j for $u_h(\cdot, jk)$ (with k the time step), the scheme is

$$\langle \frac{u_h^{j+1} - u_h^j}{k}, v\rangle + b(u^{j+1}, v) = \langle f^{j+1}, v\rangle, \quad v \in V_h, \quad j = 0, 1, \dots. \tag{5.16}$$

We initialize the iteration by choosing $u_h^0 \in V_h$ to be, e.g., the interpolant, L^2 projection, or elliptic projection. Notice that, at each time step, we have to solve the linear system

$$(M + kA)\alpha^{j+1} = M\alpha^j + kF^{j+1},$$

where α^j is the vector of coefficients of u_h^j with respect to a basis, and M, A, and F, are the mass matrix, stiffness matrix, and load vector respectively.

To analyze this scheme, we proceed as we did for the semidiscrete scheme, with some extra complications coming from the time discretization. In particular, we continue to use the elliptic projection w_h as a representative of u. Thus the consistency error is given by

$$\begin{aligned}\langle\frac{w_h^{j+1}-w_h^j}{k},v\rangle + b(w_h^{j+1},v) - \langle f^{j+1},v\rangle\\ &= \langle\frac{u^{j+1}-u^j}{k},v\rangle + b(u^{j+1},v) - \langle f^{j+1},v\rangle + \langle\frac{(w_h^{j+1}-u^{j+1})-(w_h^j-u^j)}{k},v\rangle\\ &= \langle\frac{u^{j+1}-u^j}{k}-\frac{\partial u^{j+1}}{\partial t},v\rangle + \langle\frac{(w_h^{j+1}-u^{j+1})-(w_h^j-u^j)}{k},v\rangle = \langle z^j,v\rangle,\end{aligned}$$

where the last line gives the definition of z^j. Next we estimate the two terms that comprise z^j, in L^2. First we have

$$\|\frac{u^{j+1}-u^j}{k}-\frac{\partial u^{j+1}}{\partial t}\| \le \frac{k}{2}\|\frac{\partial^2 u}{\partial t^2}\|_{L^\infty(L^2)},$$

by Taylors theorem. Next,

$$\frac{(w_h^{j+1}-u^{j+1})-(w_h^j-u^j)}{k} = \frac{1}{k}\int_{jk}^{(j+1)k}\frac{\partial}{\partial t}[w_h(s)-u(s)]\,ds,$$

so

$$\|\frac{(w_h^{j+1}-u^{j+1})-(w_h^j-u^j)}{k}\| \le ch^{r+1}\|\frac{\partial u}{\partial t}\|_{L^\infty([jk,(j+1)k];H^{r+1}(\Omega)}.$$

Thus we have obtained a bound on the consistency error:

$$\langle\frac{w_h^{j+1}-w_h^j}{k},v\rangle + b(w_h^{j+1},v) - \langle f^{j+1},v\rangle = \langle z^j,v\rangle, \quad v\in V_h, \quad j=0,1,\ldots.$$

with

$$\|z^j\| \le c(k\|\frac{\partial^2 u}{\partial t^2}\|_{L^\infty([0,T];L^2(\Omega))} + h^{r+1}\|\frac{\partial u}{\partial t}\|_{L^\infty([0,T];H^{r+1}(\Omega)}) =: E, \quad j=0,1,\ldots.$$

Combining with the scheme (5.16), we get (for $y_h = w_h - u_h$)

$$\langle\frac{y_h^{j+1}-y_h^j}{k},v\rangle + b(y_h^{j+1},v) = \langle z^j,v\rangle, \quad v\in V_h.$$

We conclude the argument with a stability argument. Choose $v = y_h^{j+1}\in V_h$. This becomes:

$$\|y_h^{j+1}\|^2 + kb(y_h^{j+1},y_h^{j+1}) = \langle y_h^j + kz^j, y_h^{j+1}\rangle,$$

so

$$\|y_h^{j+1}\| \le \|y_h^j\| + kE,$$

and, by iteration,

$$\max_{0\le j\le M}\|y_h^j\| \le \|y_h^0\| + TE.$$

In this way we prove that

$$\max_{0\le j\le M}\|u^j - u_h^j\| = O(k+h^{r+1}).$$

Exercise for the reader: analyze the convergence of the fully discrete finite element method using Crank–Nicolson for time discretization. In the stability argument, you will want to use the test function $v = (y_h^{j+1} + y_h^j)/2$.

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

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

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 $r \ge 1$.

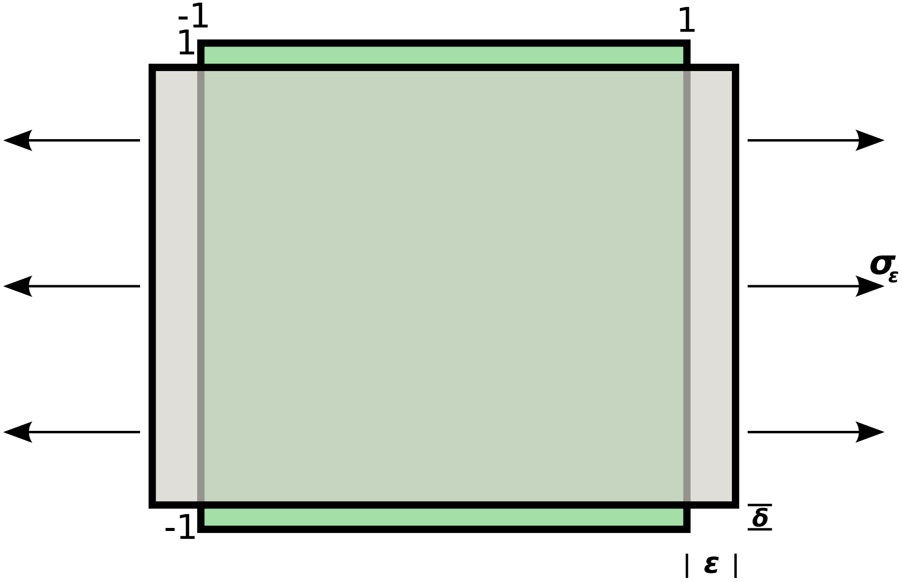
We bounded the approximation error for the Lagrange finite element spaces V_h using the Bramble–Hilbert lemma and scaling. Putting together the above considerations, for the model scalar second order elliptic PDE, $-\operatorname{div} a \operatorname{grad} u + cu = f$, we obtained H^1 error estimates. 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 ν. 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 1.0×10^7 psi, or 6.9×10^{10} pascals.

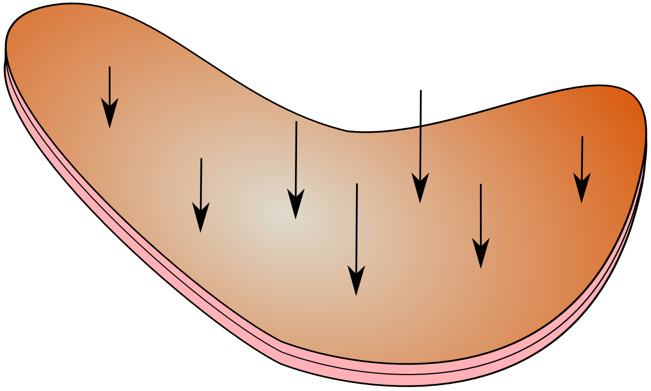
FIGURE 6.1. Elastic cube under tension σ_ϵ. Strain is ϵ in the direction of tension, $-\delta$ in the normal directions. Young's modulus is $E = \sigma_\epsilon/\epsilon$. Poisson ratio is $\nu = \delta/\epsilon$.



Under the same tension test, Poisson's ratio is the ratio of the 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 2×2 Hessian matrix of w. (Warning: sometimes the notation ∇^2 is used for the Laplacian, but we do not follow this usage.) For a matrix τ 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 ϕ, let $\operatorname{grad}\phi$ denote the Jacobian matrix field $(\partial\phi_i/\partial x_j)$, and for a matrix field τ, 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 Ω 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 Ω_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 $\bar{w} = (r^4 - 4r^2 + 3)/64$.

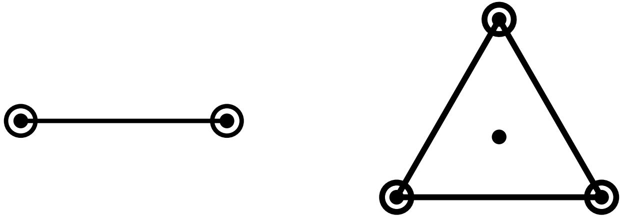
3. Conforming finite elements for the plate problem

Since the weak formulation of the plate problem (with either clamped 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.

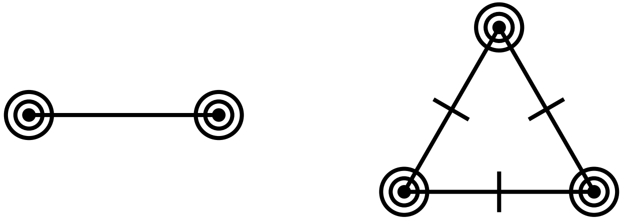


First we show the unisolvence of the proposed DOFs. Suppose $u \in \mathcal{P}_3(T)$ for some triangle T, and all the DOFs for u vanish. For an edge e of T, let $v = u|_e$. Using the distance along e as a coordinate, we may view e as an interval, and v belongs to $\mathcal{P}_3(e)$, and both v and its 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 λ_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 $I_T u \in \mathcal{P}_5(T)$ 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 θ. 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

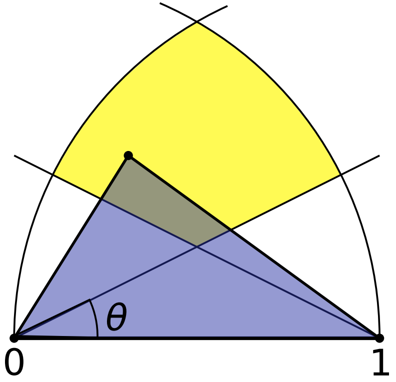
$$|u - I_{\hat{T}} u|_r \le c|u|_s, \tag{6.4}$$

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

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

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 θ. Consequently the third vertex must lie in the compact region shown in yellow.



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 Ω 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 ϕ 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.

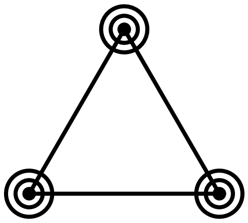
3.2. Reduced Hermite quintic. The difficulties with Hermite quintic elements (many DOFs, need for second derivatives, complicated) motivate the search for simpler elements. It turns out that one slight simplification can be made fairly easily. Define

$$\mathcal{P}_5'(T) = \{\, u \in \mathcal{P}_5(T) \,|\, \partial u/\partial n \in \mathcal{P}_3(e) \text{ on each edge } e \,\}.$$

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 $\partial u/\partial n \in \mathcal{P}_4(e)$ is a homogeneous linear equation which must be satisfied by 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.

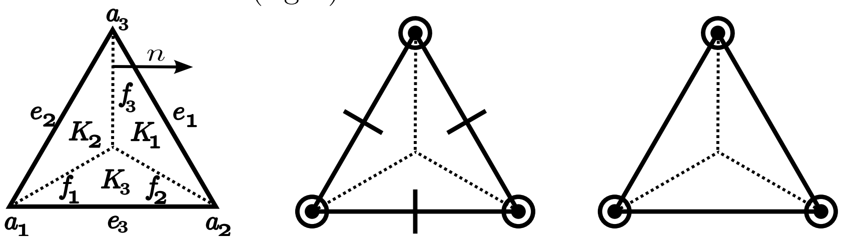
FIGURE 6.6. Reduced Hermite quintic element.



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

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

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 which, if satisfied by the 30 coefficients, insure that the piecewise cubic u is a C^1 function. Thus the dimension of the space of shape functions is ≥ 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 adapted 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 μ_i^2, where μ_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 $u_i = p_i\mu_i^2$, where $p_i \in \mathcal{P}_1$. Now μ_1 and μ_2 agree along the edge f_3, since both are linear functions which are zero at a_3 and one at b. 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 function on T which is equal to p_i on K_i is continuous.

Now let n be the unit normal to f_3 pointing from K_2 into K_1 and consider the continuity of $\partial_n u := \partial u/\partial n$ across f_3. This gives

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

where we have used the fact that $\mu_1 = \mu_2$ and $p_1 = p_2$ on f_3. Dividing by the polynomial μ_1 then gives

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

and, passing to the vertex a_3 of f_3, where $\mu_1 = \mu_2 = 0$, we obtain

$$p_1(a_3)\partial_n\mu_1 = p_1(a_3)\partial_n\mu_2.$$

Now the constant $\partial_n\mu_1$ and $\partial_n\mu_2$ are not equal. In fact, the former is negative and the latter is positive. Therefore, the preceding 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 $p_1 = c_1\mu_1$ for some constant c_1. Then $u_1 = p_1\mu_1^2 = c_1\mu_1^3$. Evaluating both sides at the barycenter b, we see that $c_1 = u(b)$, so $u_1 = u(b)\mu_1^3$. Similarly $u_2 = u(b)\mu_2^3$ and $u_3 = u(b)\mu_3^3$, and to complete the proof, it remains to show that $u(b) = 0$. For this, we equate $\partial_n u_1(b)$ and $\partial_n u_2(b)$ to find

$$3u(b)\mu_1^2(b)\partial_n\mu_1 = 3u(b)\mu_2^2(b)\partial_n\mu_2.$$

Since $\mu_1(b) = \mu_2(b) = 1$ and $\partial_n\mu_1 \neq \partial_n\mu_2$, this implies that $u(b) = 0$.

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 $\inf_{v\in V_h}\|u - v\|_2 \le Ch^2\|u\|_4$ 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 Δv 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

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

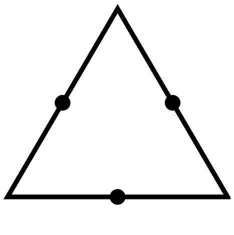
Not surprisingly, this method does not work in general. However, as we shall see, if we take elements which are in some sense "nearly C^1", we obtain a convergent method.

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 our motivation now is 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

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

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

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

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:

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

Next we add over T:

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

In other words

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

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

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

Taking $v = U_h - u_h$, we get

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

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

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

and so

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

For the approximation error $U_h - u$ we could take U_h to be the interpolant into V_h and use a Bramble–Hilbert argument. But even easier, we take U_h 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 $\|U_h - u\|_h \le ch\|u\|_2$. 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 ∂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 Ω, 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 Ω) 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

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

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 ψ). 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 ϕ 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 ϕ 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

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

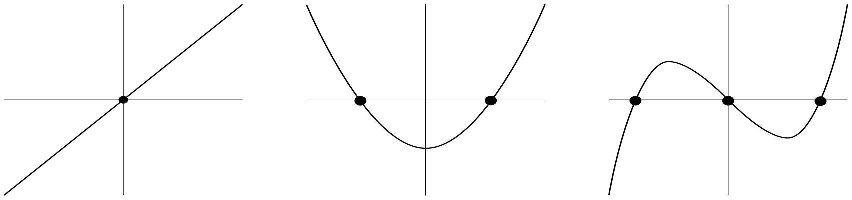
which gives $\|u-u_h\| \le ch\|u-u_h\|_h$, and so the theorem. □

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

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

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 $\pm1/\sqrt{3}$. 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 rth degree Legendre polynomial, if and only if it vanishes at the r Gauss points (zeros of the rth degree Legendre polynomial). See Figure 7.2.

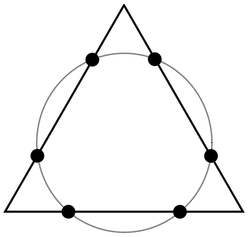
FIGURE 7.2. Legendre polynomials of degree 1, 2, and 3, and their roots, the Gauss points.



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

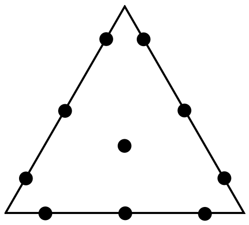
There is however one issue. The space V_h defined by (7.4) is a finite element space, definable through shape functions and DOFs, for r odd, but *not* for r even. To see what goes wrong in the even case, take $r = 2$. The shape function space is, of course, $\mathcal{P}_2(T)$, and the natural choice 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.

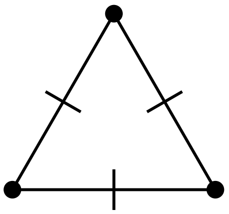


of T. Thus v is a multiple of the bubble function on T, and so the DOF at the barycenter implies $v \equiv 0$.

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 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.

Figure 7.5. The Morley nonconforming plate element.



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). We remark that the unisolvency argument implies that if $v \in V_h$, then $\partial v/\partial x$ is a piecewise linear which is continuous at the midpoints of edges, i.e., belongs to the nonconforming $\mathcal{P}_1$ space we studied in the previous section.

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

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

where is this section we denote by broken H^2-like norm:

$$\|v\|_h = \|\nabla_h^2 v\|.$$

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 ϕ 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:

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

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 ν, 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, θ the rotation of vertical fibers, M the bending moment tensor, and ζ 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 also for second order equations.

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 τ and the second by a scalar test function v, and integrate over Ω. 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 σ and u such that

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

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 (σ, u) and the test function (τ, v) which is symmetric: find (σ, 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 only the product or σ and τ, 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 σ). 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

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

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 over the Hilbert space $H(\operatorname{div}) \times L^2$, as in (8.2).

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 (σ, 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 (σ, 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 σ 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 σ 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 τ. 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.

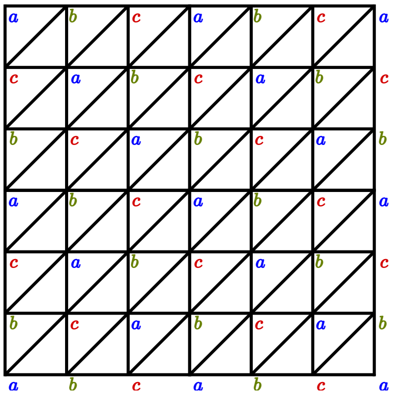


FIGURE 8.1. A piecewise linear which is orthogonal to all piecewise constants $(a + b + c = 0)$.

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 σ 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 of a relatively coarse mesh).

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

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

Notice that in one dimension $H^1 = H(\operatorname{div})$. 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 σ 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

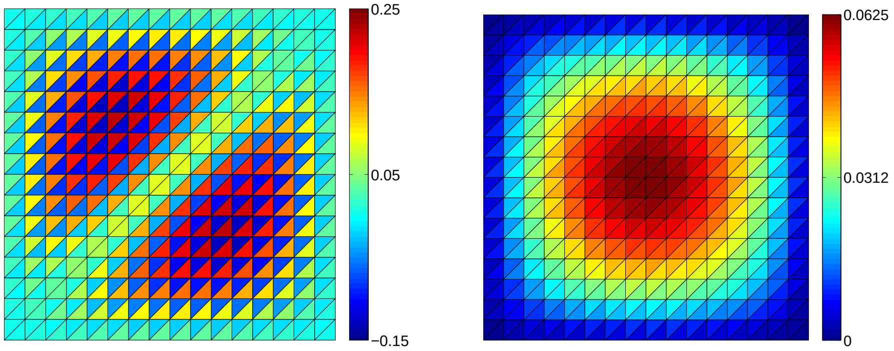


FIGURE 8.2. Approximation of the mixed formulation for Poisson's equation using piecewise constants for u and for σ using either continuous piecewise linears (left), or Raviart–Thomas elements (right). The plotted quantity is u in each case.

quadratics for σ and piecewise constants for u. That method is clearly unstable. (Our test problem has $u(x) = \cos(\pi x/2)$.)

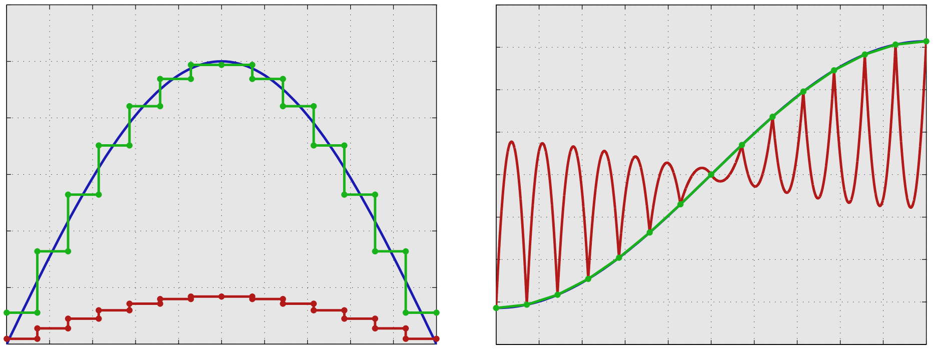


FIGURE 8.3. Approximation of the mixed formulation for $-u'' = f$ in one dimension with two choices of elements, piecewise constants for u and piecewise linears for σ (a stable method, shown in green), or piecewise constants for u and piecewise quadratics for σ (unstable, shown in red). The left plot shows u and the right plot shows σ, 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 α 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 σ and τ, and we define the space

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

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

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

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 $\tau \in H(\operatorname{div})$. 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

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

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

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

We can extend this result to all of $H(\operatorname{div})$ 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 $H(\operatorname{div})$). *The map $\gamma\tau = \tau \cdot n$ extends to a bounded linear map from $H(\operatorname{div})$ 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 Δ 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 σ 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 $\|T\|_{\mathcal{L}(V,W)} = \|T^{**}\|_{\mathcal{L}(V^{**},W^{**})} \le \|T^*\|_{\mathcal{L}(W^*,V^*)}$, 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:

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

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

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

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 $v \in V$,

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

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

$$\begin{tikzcd}
Z \arrow[r, "i_Z"] \arrow[d, "\cong"] & V \arrow[d, "\cong"] & & V \arrow[r, "\pi_Z"] \arrow[d, "\cong"] & Z \arrow[d, "\cong"] \\
Z^* \arrow[r, "\pi_Z^*"] & V^* & & V^* \arrow[r, "i_Z^*"] & Z^*
\end{tikzcd}$$

where the vertical maps are the Riesz isomorphisms. This says, that Z^* may be viewed simply as a subspace of V^* with π_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:

$$\begin{tikzcd}
V \arrow[r, "T"] \arrow[d, "\pi_{Z^\perp}"] & W \\
Z^\perp \arrow[r, "\tilde{T}"] & Y \arrow[u, "i_Y"']
\end{tikzcd}$$

Taking duals we get the commuting diagram

$$\begin{tikzcd}
V^* & W^* \arrow[l, "T^*"'] \arrow[d, "\pi_{Y^*}"] \\
(Z^\perp)^* \arrow[u, "i_{(Z^\perp)^*}"] & Y^* \arrow[l, "\tilde{T}^*"']
\end{tikzcd}$$

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 described 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 σ, u, and there is a constant such that

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

Theorem 8.7 (Brezzi's theorem in operator form). *Let $Z = \mathcal{N}(B)$ and define $A_{ZZ} : Z \to Z^*$ by $A_{ZZ} = \pi_{Z^*} \circ A \circ i_Z$. 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 $\|A_{ZZ}^{-1}\|$, $\|(B \circ i_{Z^\perp})^{-1}\|$, and $\|A\|$.*

Proof. In addition to A_{ZZ}, define maps $A_{Z\perp} = \pi_{Z^{\perp *}} \circ A \circ i_Z : Z \to Z^{\perp *}$ and, similarly, $A_{\perp Z}$ and $A_{\perp\perp}$. We also define $B_\perp = B \circ i_{Z^\perp} : Z^\perp \to W^*$. (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

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

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

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

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

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

Note that γ_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 $\|B^* w\| \ge \gamma_2 \|w\|$ for all $w \in W$, which is equivalent to, that for all $w \in W$ there exists $0 \ne v \in V$ such that $b(v, w) = B^* w(v) \ge \gamma_2 \|w\| \|v\|$, or, finally:

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

In this case γ_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 $\|A\|$, γ_1^{-1}, and γ_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:

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

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

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

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

For the second Brezzi condition we show that for any $v \in L^2$ we can find $\tau \in H(\operatorname{div})$ with $\operatorname{div} \tau = v$ and $\|\tau\|_{H(\operatorname{div})} \leq c\|v\|$. 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

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

and

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

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 theorem gives the fundamental estimate for mixed methods. It provides a quasioptimality result analogous to that we obtained for a coercive bilinear form, e.g., for the error in the H^1 norm for a standard Galerkin method for the Poisson equation. A major message of the theorem is that, unlike for coercive formulations, for saddle point problems it is n ot true we do not obtain quasioptimality for every choice of the Galerkin subspaces V_h and W_h. Rather, the subspaces must be chosen so that (8.13) and (8.14) hold. That is, the Galerkin subspaces must be chosen with a view not only to approximation, but also to stability.

For the standard finite element method for the Poisson equation, after we obtained the basic H^1 estimate, we used additional arguments (Aubin-Nitsche duality) to obtain an additional estimate (in L^2), which was an estimate of one order higher. For the mixed method case as well, it is often possible to use duality arguments to obtain improved error estimates. These separate the errors in σ and u, and/or consider the errors in weaker norms. Below we will see specific examples.

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 out $\gamma_{2,h} \to 0$ 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

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

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 τ. 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 τ 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 τ and $\phi \in \mathring{C}^\infty(\Omega)$. If τ 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 τ. This shows that the weak divergence of τ 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 τ to belong to $H(\mathrm{div})$ 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 τ 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 $V_h \subset H(\mathrm{div})$ 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 π_h as $H^1(\Omega; \mathbb{R}^2)$ rather than $H(\mathrm{div})$. The reason for this is that $\int_e \tau \cdot n_e \, ds$ need not be defined for $\tau \in H(\mathrm{div})$, 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:

$$\begin{tikzcd} H^1 \arrow[r, "\operatorname{div}"] \arrow[d, "\pi_h"] & L^2 \arrow[d, "P_h"] \\ V_h \arrow[r, "\operatorname{div}"] & W_h. \end{tikzcd}$$

We shall also prove below that π_h is bounded on H^1:

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

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

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*

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

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 $\|\tau\|_{H(\operatorname{div})} \le c\|v\|$. First we find $\sigma \in H^1(\Omega; \mathbb{R}^2)$ with $\operatorname{div} \sigma = v$, $\|\sigma\|_1 \le c\|v\|$. 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,

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

□

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*

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

For the second infimum, we of course have

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

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 π_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: π_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

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

10.2.1. *Improved estimates for σ.* This theorem gives first order convergence for σ 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 σ or u depend on the H^1 norm of $\operatorname{div}\sigma$. Here we show how to obtain improved estimates for σ 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 σ.

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

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

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

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

and

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

□

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

Theorem 8.18 (Polynomial de Rham sequence). *For any $r\ge1$, the complex of maps*

$$\mathcal{P}_r(\mathbb{R}^2)\xrightarrow{\operatorname{curl}}\mathcal{P}_{r-1}(\mathbb{R}^2;\mathbb{R}^2)\xrightarrow{\operatorname{div}}\mathcal{P}_{r-2}(\mathbb{R}^2)\to0$$

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

$$0\to\mathbb{R}\xrightarrow{\subset}\mathcal{P}_r(\mathbb{R}^2)\xrightarrow{\operatorname{curl}}\mathcal{P}_{r-1}(\mathbb{R}^2;\mathbb{R}^2)\xrightarrow{\operatorname{div}}\mathcal{P}_{r-2}(\mathbb{R}^2)\to0.$$

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 curl kills the constant functions (which is obvious), and that $\operatorname{div}\circ\operatorname{curl}=0$, 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 curl 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

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

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

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

□

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*

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

is a resolution of the constants.

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

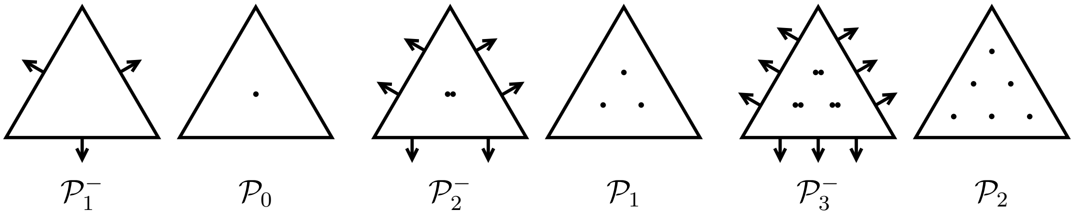
$$\tau \mapsto \int_e \tau \cdot n_e p(s)\, ds, \quad p \in \mathcal{P}_{r-1}(e), \tag{8.18}$$

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 $3r + r(r-1) = r(r+2) = \dim \mathcal{P}_{r-1}^-(T;\mathbb{R}^2)$ altogether. So, to show unisolvence, all we need to do is show that if all the DOFs vanish, then $\tau \in \mathcal{P}_{r-1}^-(T;\mathbb{R}^2)$ 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 ∂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 $\tau = \operatorname{curl}\phi$, 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 ϕ 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

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

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

$$\begin{tikzcd} H^1 \arrow[r, "\operatorname{div}"] \arrow[d, "\pi_h"] & L^2 \arrow[d, "P_h"] \\ V_h \arrow[r, "\operatorname{div}"] & W_h. \end{tikzcd}$$

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 $r \ge 1$. Assuming a smooth solution, we thus get

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

The improved estimates for σ 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

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

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 Ω be convex.

First we recall the error equations

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

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

Note that we have replaced u with P_hu 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 $\|w\|_2 + \|\rho\|_1 \le c\|P_h u - u_h\|$.

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

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

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

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

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

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

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

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

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 $u \in \mathring{H}^1(\Omega;\mathbb{R}^2)$, $p \in \hat{L}^2(\Omega)$, 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 $\Pi_h : \overset{\circ}{H}{}^1 \to V_h$ 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 $\tilde{\pi}_h : \overset{\circ}{H}{}^1 \to V_h$ by

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

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 π_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

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

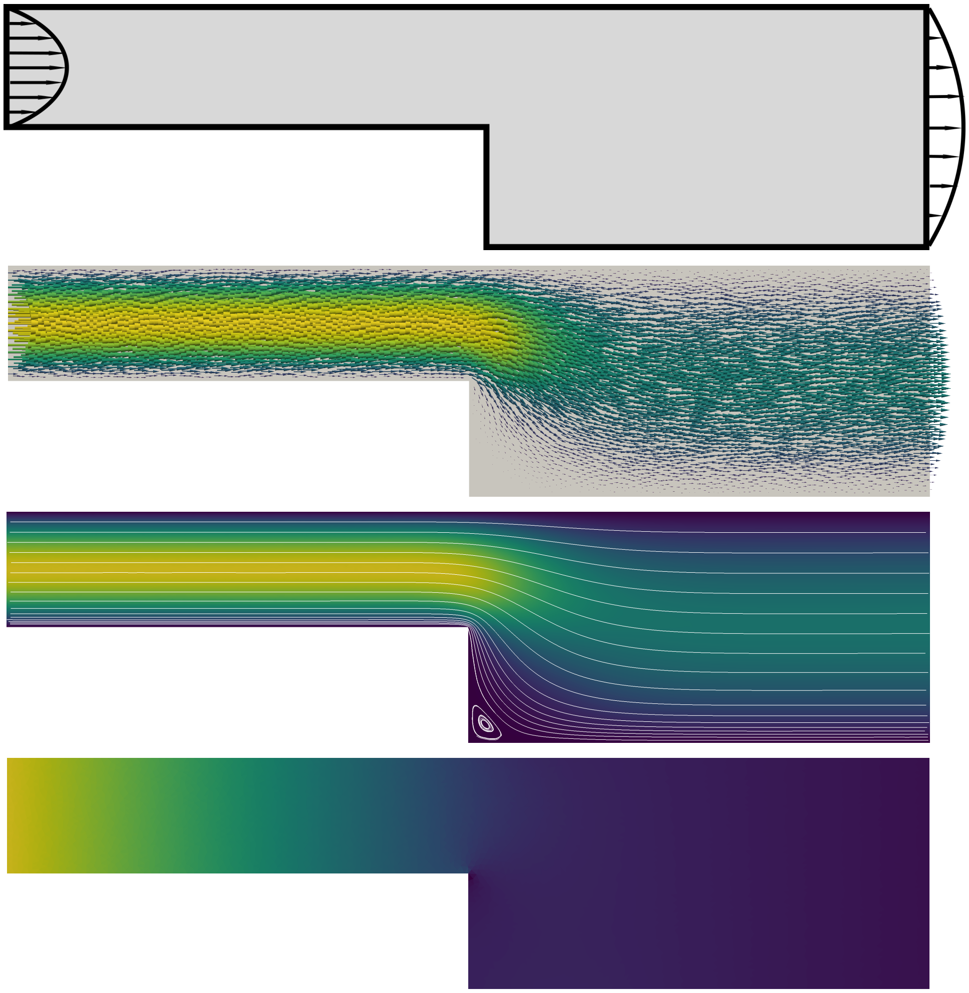
Notice that the rate of converge is only $O(h)$, the same as we would get for the best approximation using $\mathcal{P}_1$ Lagrange elements. The method in fact does not achieve $\|u - u_h\|_1 = O(h^2)$, because of the low order of pressure approximation.

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 ($x_1 = -4$) 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 $x_1 = 4$, 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 is unstructured but not adapted to the flow, has 10,783 elements. The velocity field and the pressure field were visualized using Paraview, with the color showing the magnitude of the velocity or pressure. In the third plot, the streamlines, as computed in Paraview from the discrete velocity field, are shown.

The $\mathcal{P}_2$-$\mathcal{P}_0$ element may be generalized to $\mathcal{P}_r$-$\mathcal{P}_{r-2}$ and analyzed with similar techniques. It is a stable element for any $r \ge 2$. For smooth solutions, it converges with rate $O(h^{r-1})$ for the velocity in H^1 and for the pressure in L^2.

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

FIGURE 8.5. Flow over a step computed using P_2-P_0 elements on an non-adaptive unstructured mesh of 10,783 triangles. The first figures shows the domain and the given velocity boundary conditions. The second figure shows the velocity field and the third figures adds the streamlines computed from the discrete velocity field. The final figure shows the pressure field.



of the T and is positive in the interior. It may be written as $\lambda_1\lambda_2\lambda_3$ where the λ_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 usual 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 to rewrite this as

$$\int_\Omega \pi_h v \cdot \operatorname{grad} q\, dx = \int_\Omega v \cdot \operatorname{grad} q\, dx, \quad q \in W_h.$$

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

$$\int_T \pi_h v\, dx = \int_T v\, dx.$$

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. The usual scaling argument then gives

$$\|\pi_T v\|_{L^2(T)} \le c\|v\|_{L^2(T)}, \quad |\pi_T v|_{H^1(T)} \le ch^{-1}\|v\|_{L^2(T)}$$

so $\|\pi_T v\|_{H^1(T)} \le ch^{-1}\|v\|_{L^2(T)}$. 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 Π_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

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

We can also use an Aubin–Nitsche duality argument to get

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

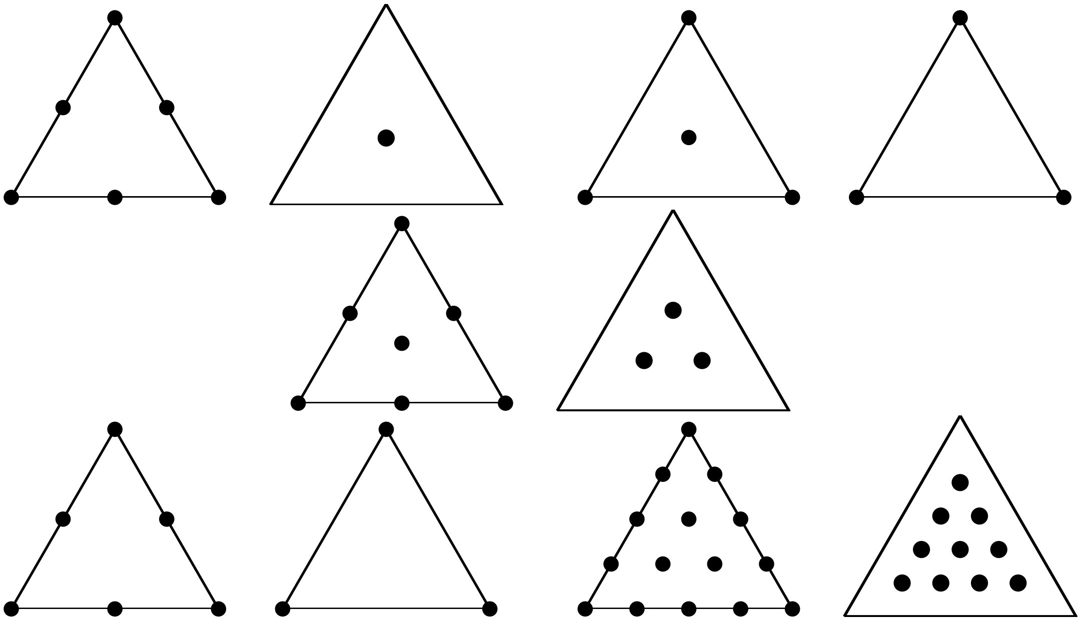
We do *not* get second order convergence for p in L^2.

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 $r \ge 2$ 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 σn for a tensor (matrix) σ is known as Cauchy's Theorem. The same theorem shows that, as a consequence of the conservation of angular momentum, the matrix σ 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 deformation of the body. For an *elastic* material, the stress tensor σ 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 ν:

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

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 ν 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 Γ_D and Γ_N be disjoint open subsets of $\partial\Omega$ whose closures cover $\partial\Omega$. We assume that Γ_D is not empty (it may be all of $\partial\Omega$). On Γ_D we impose the displacement

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

with $g : \Gamma_D \to \mathbb{R}^n$ given. On Γ_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 τ and vector field v. If τ 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, τ is symmetric, and so orthogonal to all skew-symmetric matrices). Thus for symmetric τ,

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

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

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

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 Γ_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 $b(v,v) \ge 0$. In fact, since we assumed that C is positive definite on $\mathbb{R}^{n\times n}_{\text{symm}}$, we have

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

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 Ω be a domain with Lipschitz boundary. Then there exists a constant c such that*

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

PROOF.

$$\begin{aligned}\|\,\epsilon(v)\|^2 &= \frac14\int[\operatorname{grad} v + (\operatorname{grad} v)^T] : [\operatorname{grad} v + (\operatorname{grad} v)^T]\,dx\\ &= \frac14\|\operatorname{grad} v\|^2 + \frac14\|(\operatorname{grad} v)^T\|^2 + \frac12\int \operatorname{grad} v : (\operatorname{grad} v)^T\,dx \qquad (9.8)\\ &= \frac12\|\operatorname{grad} v\|^2 + \frac12\int \operatorname{grad} v : (\operatorname{grad} v)^T\,dx.\end{aligned}$$

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

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

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 Ω be a domain with a Lipschitz boundary and Γ_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 Ω, Dirichlet boundary Γ_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

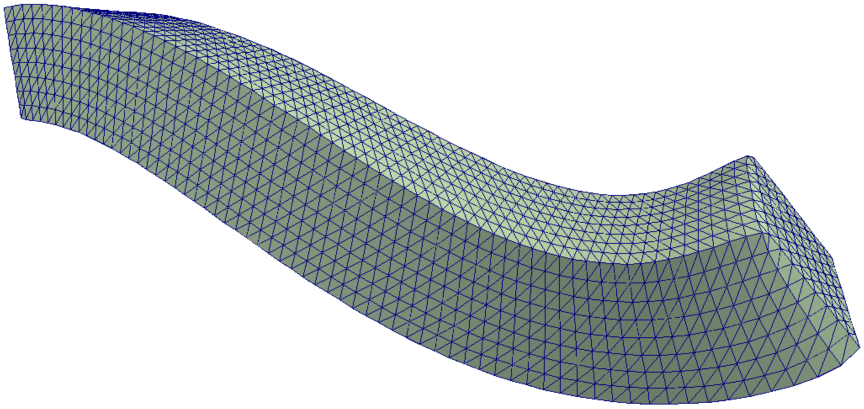
$$\|u - u_h\|_1 \leq c h^r \|u\|_{r+1}.$$

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

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

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 a multiple of 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 100 or 1000.

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

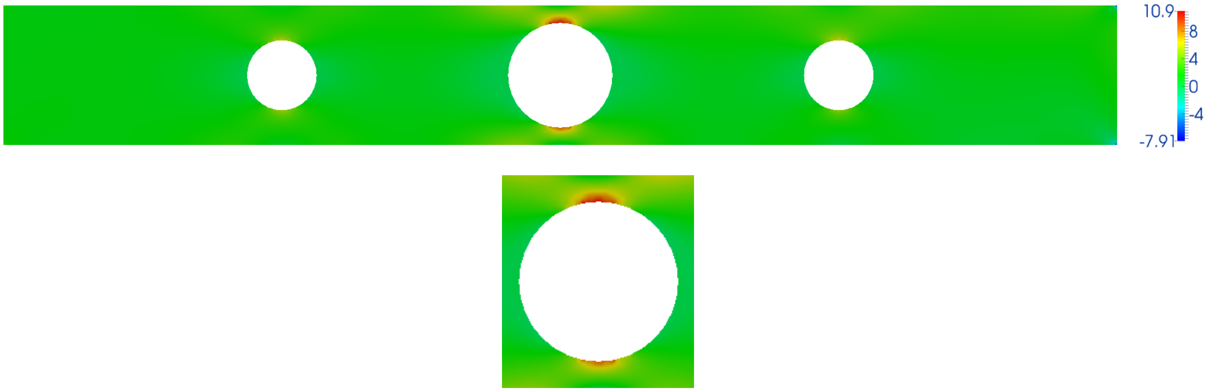


The second example is the analogous problem in two dimensions, except that the domain is the rectangle $(0, 8) \times (0, 1)$ with three circular cut-outs removed. Figure 9.2 shows the longitudinal stress, i.e., the stress component σ_{11}, which gives the tension in the x_1 direction (or the compression, 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 $\lambda \ge 0$, 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 λ increases toward $+\infty$, or, equivalently, as the Poisson ratio ν 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

FIGURE 9.2. Longitudinal stress of a 2D elastic bar with cut-outs with left face clamped and a rigid displacement applied to the right face. Detail shows stress concentration around at the top and bottom of middle cut-out.



the example just computed, with the stress shown in Figure 9.2, but now take the Poisson 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, using a slightly coarser mesh so that the grid scale oscillations can be clearly seen.

We know that the displacement method gives the error estimate

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

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 λ grows, even though the exact solution does not degenerate.

Let us investigate the dependence onλ 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

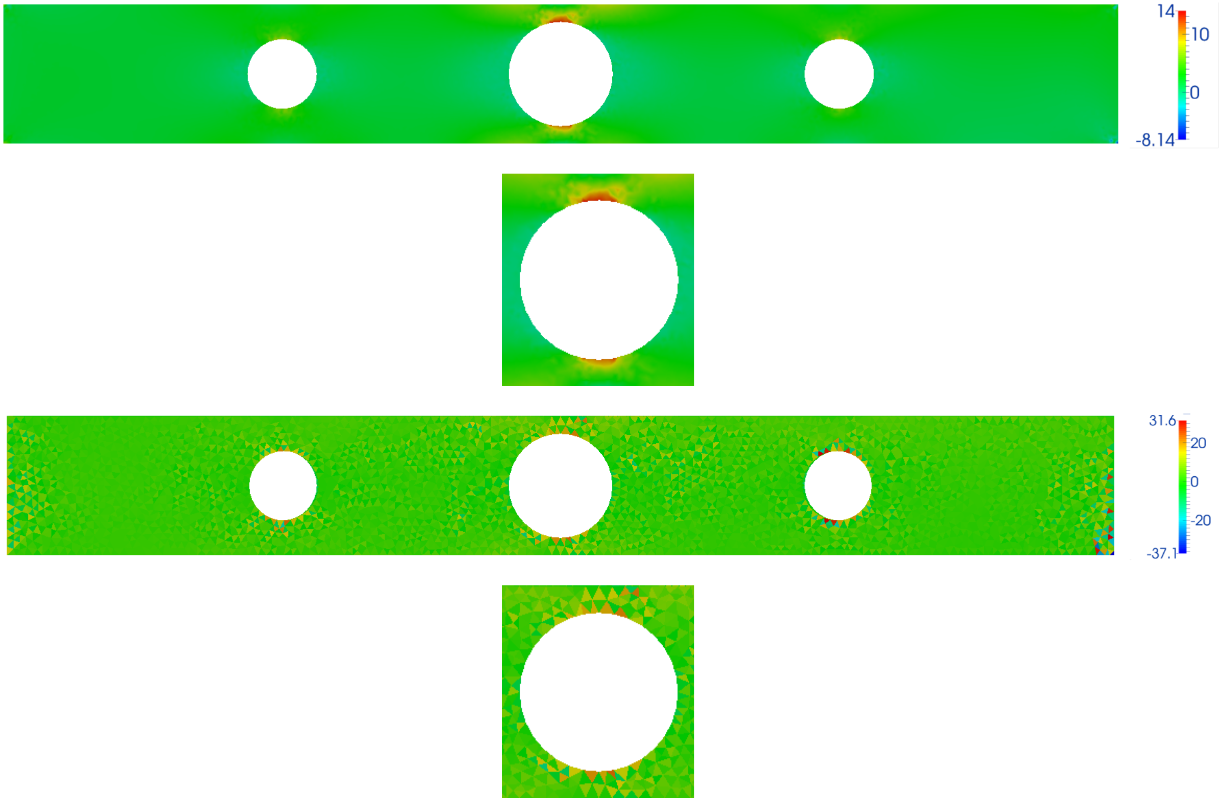
$$b(u, v) = 2\mu \int \epsilon(u) : \epsilon(v)\, dx + \lambda \int (\operatorname{div} u)(\operatorname{div} v)\, dx,$$

so

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

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

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).



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 λ. But finally we get to the culprit. Since the coefficient λ enters the bilinear form b, $\|b\|$ tends to ∞ with λ.

5. The Airy stress function and compatibility of strain

Before turning to mixed finite elements for elasticity, we establish some analytic results relevant to elasticity which we shall need. Recall that we proved above, in Theorem 8.18, the exactness of the polynomial de Rham complex, i.e., that the sequence

$$0 \to \mathbb{R} \xrightarrow{\subset} \mathcal{P}_r(\mathbb{R}^2) \xrightarrow{\text{curl}} \mathcal{P}_{r-1}(\mathbb{R}^2; \mathbb{R}^2) \xrightarrow{\text{div}} \mathcal{P}_{r-2}(\mathbb{R}^2) \to 0$$

is an exact complex. This gathered in one-statement several useful relations among the polynomial spaces. The continuous version of this statement is the exactness of the de Rham complex, i.e., the complex

$$0 \to \mathbb{R} \xrightarrow{\subset} H^1(\Omega) \xrightarrow{\text{curl}} H(\text{div}, \Omega) \xrightarrow{\text{div}} L^2(\Omega) \to 0,$$

for any *simply-connected domain*. Let us verify this. We know that div maps $H^1(\Omega; \mathbb{R}^2)$ onto $L^2(\Omega)$, so, a fortiori, it maps $H(\text{div})$ onto L^2. The verification of the identity $\text{div}\,\text{curl} = 0$ is immediate. The remaining point is that if $u \in H(\text{div})$ and $\text{div}\,u = 0$, then $u = \text{curl}\,\phi$ for some $\phi \in H^1$. This fact is typically proved in vector calculus using a line integral to define ϕ. The divergence theorem and the simple-connectivity insure that the line integral is path independent and so that ϕ is well-defined.

In our development of mixed methods for elasticity we will rely on an analogous sequence in which the scalar functions are replaced by vector functions and the vector function by a symmetric matrix field. The divergence operator in the de Rham complex, which maps vector fields to scalar functions, will be replaced by divergence from symmetric matrix fields to vector field, defined by applying the divergence to each row of the matrix. However the curl operator, which in the de Rham complex maps scalar functions to vector fields, will be replaced by a *second order operator* called the *Airy stress function*, defined

$$J\phi = \text{curl}\,\text{curl}\,\phi = \begin{pmatrix} \partial^2\phi/\partial x_2^2 & -\partial^2\phi/\partial x_1\partial x_2 \\ -\partial^2\phi/\partial x_1\partial x_2 & \partial^2\phi/\partial x_1^2 \end{pmatrix}.$$

Here $\text{curl}\,\phi$ is a vector field so curl applies to each component of it, to give the two rows of the matrix field $\text{curl}\,\text{curl}\,\phi$.

THEOREM 9.4 (The elasticity complex in two dimensions). *If Ω is a simply-connected domain, then the complex*

$$0 \to \mathcal{P}_1(\Omega) \xrightarrow{\subset} H^2(\Omega) \xrightarrow{J} H(\text{div}, \Omega; \mathbb{R}^{2\times 2}_{symm} \xrightarrow{\text{div}} L^2(\Omega; \mathbb{R}^2) \to 0,$$

is an exact sequence.

PROOF. Note that the components of $J\phi$ are the same as the components of the Hessian of ϕ, except for order and sign, so that the kernel of J is indeed $\mathcal{P}_1(\Omega)$. It it simple to check that $\text{div}\,J\phi = 0$ for all ϕ, so there are two points to check. 1) If $\tau \in H(\text{div}, \Omega; \mathbb{R}^{2\times 2}_{\text{symm}}$ and $\text{div}\,\tau = 0$, then $\tau = J\phi$ for some $\phi \in H^2$. 2) Every $v \in L^2(\Omega; \mathbb{R}^2)$ equals $\text{div}\,\tau$ for some $\tau \in H(\text{div}, \Omega; \mathbb{R}^{2\times 2}_{\text{symm}}$.

Proof of 1): Since $\text{div}\,\tau = 0$, each row of τ is a divergence free vector field, so $(\tau_{11}, \tau_{12}) = \text{curl}\,\psi_1$ and $(\tau_{21}, \tau_{22}) = \text{curl}\,\psi_2$, for some ψ_1, $\psi_2 \in H^1$. Now $\partial\psi_1/\partial x_1 = \tau_{12} = \tau_{21} = -\partial\psi_2/\partial x_2$, so $\text{div}\,\psi = 0$, so $\psi = \text{curl}\,\phi$ for some $\phi \in H^2$.

Proof of 2): Each component of the vector v is the divergence of an H^1 vector field, so $v = \text{div}\,\rho$ for some H^1 matrix field ρ. However ρ need not be symmetric. Let $\tau = \rho + \text{curl}\,u$ where $u \in H^1(\Omega; \mathbb{R}^2)$, so $\text{curl}\,u$ is a matrix. Note that $\text{div}\,\tau = \text{div}\,\rho = v$. Direct calculation shows

$$\tau_{12} - \tau_{21} = \rho_{12} - \rho_{21} - \text{div}\,u,$$

so if we choose u such that $\text{div}\,u = \rho_{12} - \rho_{21}$, then τ is symmetric, and $\text{div}\,\tau = v \in L^2(\Omega; \mathbb{R}^2)$, i.e., $\tau \in H(\text{div}, \Omega; \mathbb{R}^{2\times 2}_{\text{symm}}$ as desired. □

Note that verification of exactness of the elasticity complex is based on repeated use of the exactness of the de Rham complex.

There is also an adjoint result. The formal adjoint of $J = \operatorname{curl}\operatorname{curl}$ maps a symmetric matrix field to a scalar function by the formula

$$\operatorname{rot}\operatorname{rot}\rho = \frac{\partial^2 \rho_{11}}{\partial x_2^2} - 2\frac{\partial^2 \rho_{11}}{\partial x_1 \partial x_2} + \frac{\partial^2 \rho_{22}}{\partial x_1^2}.$$

Here the first rot takes a vector field v to $-\operatorname{div} v^\perp = \partial v_2/\partial x_1 - \partial v_1/\partial x_2$, and the second rot takes a matrix field to a vector field by applying the same formula to each row. It is easy to check that if $u \in H^1(\Omega;\mathbb{R}^2)$, then $\operatorname{rot}\operatorname{rot}\epsilon(u) = 0$, i.e., every strain tensor $\epsilon(u)$ is in the kernel of rot rot. In the following theorem, we verify that the kernel of rot rot consists of precisely the strain tensors $\epsilon(u)$, $u \in H^1(\Omega,\mathbb{R}^2)$. In this sense, rot rot measures the compatibility of strain.

To state the theorem, we need some definitions. A linear vector field of the form $v(x) = a + bx^\perp = (a_1 - bx_2, a_2 + bx_1)$ for some $a \in \mathbb{R}^2$, $b \in \mathbb{R}$ is called an *infinitesmal rigid motion.* The space $RM(\Omega)$ is defined to be the space of restrictions of such infinitesmal rigid motions to Ω. Clearly $\dim RM(\Omega) = 3$. We also define $H(\operatorname{rot}\operatorname{rot},\Omega;\mathbb{R}^{2\times 2}_{\text{symm}})$ to be the space of L^2 symmetric matrix fields ρ for which $\operatorname{rot}\operatorname{rot}\rho \in L^2$.

THEOREM 9.5. *If Ω is a simply-connected domain, then the complex*

$$0 \to RM(\Omega) \xrightarrow{\subset} H^1(\Omega,\mathbb{R}^2) \xrightarrow{\epsilon} H(\operatorname{rot}\operatorname{rot},\Omega;\mathbb{R}^{2\times 2}_{symm}) \xrightarrow{\rho\operatorname{rot}} L^2(\Omega) \to 0,$$

is an exact sequence.

PROOF. It is easy to check that $\epsilon(v) = 0$ if v is a rigid motion, and that $\operatorname{rot}\rho\,\epsilon(v) = 0$ for all $v \in \mathcal{H}^1(\Omega;\mathbb{R}^2)$. So we need to verify three things: 1) If $\epsilon(v) = 0$, then v is a rigid motion. 2) If $\operatorname{rot}\operatorname{rot}\tau = 0$, then $\tau = \epsilon(v)$ for some vector field v. 3) Every L^2 function is $\operatorname{rot}\operatorname{rot}\tau$ for some symmetric matrix field τ.

Proof of 1): Let v be any vector field, and let $\epsilon = \epsilon(v)$ be the associated strain tensor. Then $v_{1,11} = \epsilon_{11,1}$, $v_{1,12} = \epsilon_{11,2}$ and $v_{1,22} = 2\,\epsilon_{12,2} - \epsilon_{22,1}$. Thus if $\epsilon(v) = 0$, all three partial derivatives of v vanish, which means that $v \in \mathcal{P}_1(\Omega;\mathbb{R}^2)$. Thus $v(x) = a + Bx$ for some $a \in \mathbb{R}^2$, $B \in \mathbb{R}^{2\times 2}$. But then $\epsilon(v) = (B + B^T)/2$, so B is skew symmetric, i.e.,

$$B = \begin{pmatrix} 0 & -b \\ b & 0 \end{pmatrix}$$

for some $b \in \mathbb{R}$. Thus, indeed, $v(x) = a + bx^\perp$.

Proof of 2): Suppose τ is a symmetric matrix field and $\operatorname{rot}\operatorname{rot}\tau = 0$. Since $\operatorname{rot}\tau$ is a rotation-free, there exists ψ such that $\operatorname{grad}\psi = \operatorname{rot}\tau$. Define

$$\eta = \begin{pmatrix} 0 & -\psi \\ \psi & 0 \end{pmatrix}, \quad \rho = \tau - \eta.$$

Then $\operatorname{rot}\eta = \operatorname{grad}\psi$, so

$$\operatorname{rot}\rho = \rho\tau - \operatorname{grad}\psi = 0.$$

Therefore, $\rho = \operatorname{grad} u$ for some H^1 vector field u. Since τ is symmetric and η is skew-symmetric, the symmetric part of ρ is τ, while the symmetric part of $\operatorname{grad} u$ is, by definition, $\epsilon(u)$. Thus $\tau = \epsilon(u)$.

Proof of 3): Any $\phi \in L^2$ is the rotation of an H^1 vector field, and each component of this vector field is rotation of an H^1 vector field, so we obtain an H^1 matrix field (not necessarily symmetric), so that $\operatorname{rot}\operatorname{rot}\tau = \phi$. Now choose an H^1 vector field u such that $\operatorname{rot} u = \tau_{12} - \tau_{21}$. Then $\tau - \operatorname{grad} u$ is a symmetric L^2 vector field, and, since $\operatorname{rot}\operatorname{grad} u = 0$, $\operatorname{rot}\operatorname{rot}(\tau - \operatorname{grad} u) = \phi$. Thus $\tau - \operatorname{grad} u \in H(\operatorname{rot}\operatorname{rot}, \Omega; \mathbb{R}^{2\times 2}_{\text{symm}})$ is the desired symmetric matrix field. □

6. Mixed finite elements for elasticity

The mixed formulation for elasticity directly treats the two fundamental first order equations, the equilibrium equation (9.1) and the constitutive equation (9.2), treating both the stress σ and the displacement u as unknowns. For simplicity we consider first the case of homogeneous Dirichlet boundary conditions. We start by inverting the constitutive relation to get

$$A\sigma = \epsilon(u),$$

where $A = C^{-1} : \mathbb{R}^{n\times n}_{\text{symm}} \to \mathbb{R}^{n\times n}_{\text{symm}}$, which is symmetric positive definite, is called the *compliance tensor*. Testing this function against a tensor field τ, and testing the equilibrium equation (9.1) against a vector field v, we get the weak formulation: find $\sigma \in H(\operatorname{div}, \mathbb{R}^{n\times n}_{\text{symm}})$, $u \in L^2(\Omega, \mathbb{R}^n)$ such that

$$\int A\sigma : \tau \, dx + \int \operatorname{div}\tau \cdot u \, dx = 0, \quad \tau \in H(\operatorname{div}, \mathbb{R}^{n\times n}_{\text{symm}}), \tag{9.11}$$

$$\int \operatorname{div}\sigma \cdot v \, dx = -\int f \cdot v \, dx, \quad v \in L^2(\Omega, \mathbb{R}^n). \tag{9.12}$$

Notice that this problem fits into the abstract framework for saddle point (8.6) we studied earlier, with $V = H(\operatorname{div}, \mathbb{R}^{n\times n}_{\text{symm}})$, $W = L^2(\Omega, \mathbb{R}^n)$, $a(\sigma,\tau) = \int A\sigma : \tau \, dx$, $b(\tau, v) = \int \operatorname{div}\tau \cdot v \, dx$.

It remains to complete the notes beyond this point…