Patrick E. Farrell

C6.4 Hillary Term 2021

Finite Element Methods for PDEs

Contents

1 Variational formulations and Galerkin approximation

1.1 Prelude

Many physical phenomena are accurately and concisely described by partial differential equations (PDEs). While we can often write down the laws of physics in this form, most of the time we cannot exactly solve them in cases of practical interest. We therefore turn to numerical methods for the approximation of their solutions.

The finite element method is one of the most popular, general, powerful and elegant approaches for approximating the solutions of PDEs. Unlike finite difference methods, it naturally handles complicated domains (useful for engines and aeroplanes) and minimally regular data (such as discontinuous forcing terms). It permits an insightful error analysis, allowing practitioners to understand the cost of the approximations they make and in some cases to automatically control them.

An excellent general reference for the material covered in these lectures is Brenner and Scott[1].

[1] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag New York, third edition edition, 2008

There are four basic ingredients in the finite element method:

1. Variational formulation in an infinite-dimensional space V;
2. Variational formulation in a finite-dimensional space $V_h \subset V$;
3. The construction of a basis for V_h;
4. The assembly and solution of the resulting linear system of equations.

We discuss these in turn.

1.2 Variational formulation

1.2.1 Motivation

Let Ω be an open, bounded, connected subset of Euclidean space $\mathbb{R}^n$, $n = 1 \ldots 3$, with Lipschitz boundary[2] $\Gamma = \partial\Omega$. Consider the model problem: given $f \in C(\Omega)$ and $g \in C(\Gamma)$, find $u \in C^2(\Omega) \cap C(\overline{\Omega})$ such that

[2] We will define this later.

$$\begin{aligned} -\nabla^2 u &= f \text{ in } \Omega, \\ u &= g \text{ on } \Gamma. \end{aligned} \tag{P}$$

Definition 1.2.1 (classical solution)**.** *A solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfying* (P) *is said to be a* classical solution *of this equation.*

This is Poisson's equation with Dirichlet boundary conditions. If $g = 0$, it describes the deformation of a stretched elastic membrane clamped to the wireframe boundary of shape Γ subject to a load $f(x)$. It also relates the gravitational potential and mass density in Newtonian mechanics, and the electric potential and charge distribution in electrostatics. It describes incompressible inviscid irrotational flow, and many other things besides.

Let us be explicit about what equality means in (P). In (P), we mean pointwise equality: in the first equation there are two functions $-\nabla^2 u$ and f, and we insist that

$$(-\nabla^2 u)(x) = f(x) \quad \text{for all } x \in \Omega. \tag{1.2.1}$$

Thus, we have to be able to evaluate f and $-\nabla^2 u$ at points, and this is why we demand that $u \in C^2(\Omega)$, $f \in C(\Omega)$ and $g \in C(\Gamma)$.

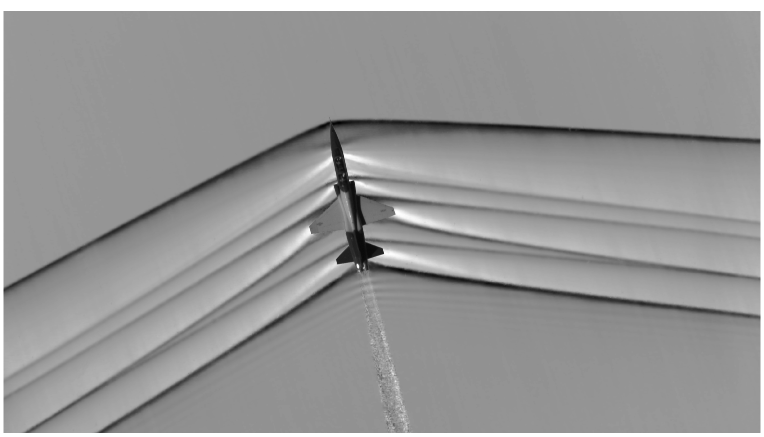

Figure 1.1: A supersonic aircraft induces a shock wave; physical reality is best described by a discontinuous function. The blurring of the interface is an artefact of an averaging process in the photography; the true interface is sharper. Credit: `https://www.nasa.gov/centers/armstrong/features/shock_and_awesome.html`

Unfortunately, the map $-\nabla^2 : C^2(\mathbb{R}^2) \to C^0(\mathbb{R}^2)$ is not invertible[3]. These continuity requirements are too strict, and do not capture certain reasonably physical scenarios: after all, discontinuities happen in real life. In Part A Differential Equations, you saw examples of first-order hyperbolic equations where a kink[4] in the data propagates

[3] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, third edition, 2001

[4] A discontinuity in the derivative.

along characteristics and induces a kink in the solution. But if the solution is not differentiable there, in what sense is it the solution of a PDE? This is not an abstract matter of no physical importance. For example, a supersonic aircraft induces a shock wave, which is a discontinuity in the pressure, temperature and density, figure 1.1. How can we interpret this as the solution of a PDE (in this case, the compressible Navier–Stokes)?

We will turn to an alternative interpretation of what it means for two functions to be equal. Rather than basing our formulation on pointwise evaluation, we will instead base it on *integration* against *test functions*. Integration is more forgiving than pointwise evaluation; we can ignore bad behaviour on a set of measure zero[5]. A classical solution will satisfy our new variational formulation, but the variational formulation will be more general: it will permit solutions like that shown in figure 1.1, and more besides.

It will turn out (by the end of the course you will agree) that the variational formulation to be described is the natural one; it is the sense in which PDEs should be understood. The success of the finite element method largely arises from its foundations on this bedrock.

[5] Loosely speaking, measure means length in one dimension, area in two dimensions, and volume in three dimensions. For a formal development, see A4, Integration.

1.2.2 Prelude: linear algebra

Imagine you have two vectors $a, b \in \mathbb{R}^n$, $n < \infty$, but you are not allowed to examine their entries. You can, however, compute their inner products $a \cdot v$, $b \cdot v$ against any test vector $v \in \mathbb{R}^n$ that you like. It is trivial to prove[6] that

$$a = b \iff a \cdot v = b \cdot v \text{ for all } v \in \mathbb{R}^n.$$

The first statement is like pointwise equality; the coefficients of a and b must match in any basis. The second statement is a *variational* statement: we demand that *when tested with any v varying in some set, the projections of a and b onto v must match.*

We will now do exactly the same thing with functions.

[6] This exercise appears on the first problem sheet of the Geometry course, in the first term of first year.

1.2.3 Casting into variational form

Let v be any sufficiently regular function such that $v|_\Gamma = 0$[7]. For now, assume that $g = 0$[8] and that Ω is polytopic (i.e. a polygon or polyhedron)[9]. Let us multiply both sides of the equation by v and integrate:

$$-\int_\Omega v\nabla^2 u \, \mathrm{d}x = \int_\Omega fv \, \mathrm{d}x. \tag{1.2.2}$$

We wish to reduce the regularity requirements on u by shifting one of the derivatives from u onto v. This is achieved by *integration by parts*.

[7] The precise definition of *sufficiently regular* will be specified later.

[8] The case of inhomogeneous boundary data is a straightforward extension and will be discussed in later lectures.

[9] This is so that we do not have to wrestle now with the issue of approximating the domain as well as the solution; it greatly complicates the error estimates.

First, recall the divergence theorem: if w is a sufficiently regular vector field defined on a Lipschitz domain Ω, then

$$\int_\Omega \nabla \cdot w \, \mathrm{d}x = \int_\Gamma w \cdot n \, \mathrm{d}s, \tag{1.2.3}$$

where n is the unit outward facing normal to Ω on Γ. Now applying the product rule to the following quantity

$$\nabla \cdot (v \nabla u) = v \nabla^2 u + \nabla u \cdot \nabla v, \tag{1.2.4}$$

integrating, and applying the divergence theorem, we find

$$\int_\Gamma v \nabla u \cdot n \, \mathrm{d}s = \int_\Omega v \nabla^2 u \, \mathrm{d}x + \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x, \tag{1.2.5}$$

or rearranged,

Theorem 1.2.2 (Integration by parts)**.** *For a Lipschitsz domain Ω and functions u, v:*

$$\boxed{-\int_\Omega v \nabla^2 u \, \mathrm{d}x = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x - \int_\Gamma v \nabla u \cdot n \, \mathrm{d}s.} \tag{IBP}$$

Applying this to our original problem, we can rephrase (1.2.2) as

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x - \int_\Gamma v \nabla u \cdot n \, \mathrm{d}s = \int_\Omega f v \, \mathrm{d}x. \tag{1.2.6}$$

As we cleverly chose v to vanish on Γ, this reduces to

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x, \tag{1.2.7}$$

or in abstract notation

$$a(u, v) = F(v) \tag{1.2.8}$$

where in this case

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x \tag{1.2.9}$$

and

$$F(v) = \int_\Omega f v \, \mathrm{d}x. \tag{1.2.10}$$

At this point, we inspect our variational problem and decide on the *function space* V such that the problem makes sense for $u, v \in V$. For now, we will define

$$V = \{ v : \Omega \to \mathbb{R} | \, a(v, v) < \infty \text{ and } F(v) < \infty \text{ and } v|_\Gamma = 0 \} \tag{1.2.11}$$

and postpone the discussion of exactly what this function space is to the next lectures.

Definition 1.2.3 (abstract variational formulation)**.** *A linear variational equality is the problem:*

$$\text{find } u \in V \text{ such that } a(u,v) = F(v) \text{ for all } v \in V. \tag{Q}$$

In particular, the variational formulation of (P) is of the form (Q) when $g = 0$.

Certainly (P) implies (Q), but not necessarily the other way around. We have weakened the regularity requirements on the solution u; instead of requiring the existence and continuity of two derivatives, we merely require the square-integrability of one derivative. For this reason, the variational formulation is also referred to as the *weak formulation*. However, it will transpire in later lectures that if u and f happen to be sufficiently regular (i.e. in $C^2(\Omega) \cap C(\overline{\Omega})$ and $C(\Omega)$ respectively) then (Q) does indeed imply (P).

There are many fundamental questions to address before we begin to discretise. Does (Q) have a solution? A unique solution? Is that solution stable with respect to the data (f and g)? We will address these points in subsequent lectures; but for now, let us accept that (Q) is well posed, and discuss how the finite element method proceeds from here.

1.3 *Galerkin approximation*

Let $V_h \subset V$ be a closed finite dimensional subspace of V. Instead of seeking a solution in the infinite-dimensional space V, we will instead look for an approximation inside a more manageable finite-dimensional subspace; as we make V_h larger and larger, we expect that our approximation will get better and better.

Definition 1.3.1 (Galerkin approximation)**.** *Given $V_h \subset V$, the* Galerkin approximation[10] *of* (Q) *is*

$$\text{find } u_h \in V_h \text{ such that } a(u_h,v_h) = F(v_h) \text{ for all } v_h \in V_h. \tag{G}$$

[10] Also referred to as the Ritz–Galerkin approximation.

This discrete, finite-dimensional problem possesses the same structure as the infinite-dimensional one: the *operators in the PDE are the same, merely restricted to subspaces*. As we will see later, this means that useful properties of the PDE like symmetry and positive-definiteness are automatically inherited by the discretisation.

There are many questions to be investigated at this point. Does this discrete problem have a unique, stable solution? What can be learned about the approximation error? We will see in subsequent lectures (Céa's Lemma and subsequent variants) that for many problems this approximation is *quasi-optimal*: u_h is the best it could be, up to some problem-dependent constants. But for now, we will continue on our tour of the big picture.

1.4 Construction of function spaces

This idea of Galerkin approximation is extremely general: V_h could be constructed in any number of ways (a spectral expansion, wavelets, a problem-specific choice, ...). The *finite element method* is a particular choice of Galerkin approximation, where the discrete function space V_h is constructed by *equipping a mesh of the domain* Ω *with piecewise polynomial basis functions*.

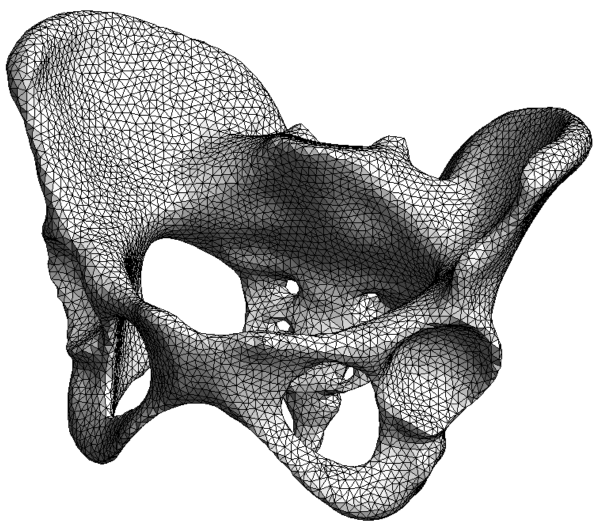

Figure 1.2: A mesh of a human pelvis, produced using the `gmsh` software.

Definition 1.4.1 (mesh, informal)**.** *A* mesh *is a geometric decomposition of a domain* Ω *into a finite collection of* cells $\{K_i\}$ *such that*

1. $\mathrm{int}(K_i) \cap \mathrm{int}(K_j) = \emptyset$ *if* $i \neq j$*, and*

2. $\cup_i K_i = \bar{\Omega}$.

With enough cells, one can approximate very complicated domains: the volume of the ocean, or the combustion chamber of an engine, or a steam turbine. For an example, see figure 1.2.

The cells are chosen to be simple geometric shapes over which we know how to integrate, e.g. triangles, tetrahedra, prisms, quadrilaterals, hexagons. For an illustration of some cells, see figure 1.3, taken from Logg et al.[11]

[11] A. Logg, K. A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2011

Once we have this geometric decomposition, we can define a discrete function space by equipping it with piecewise polynomial basis

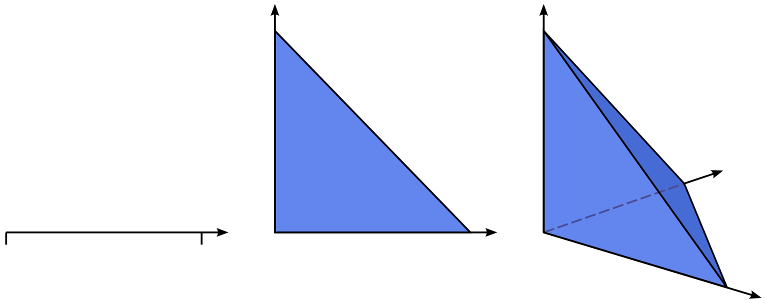

Figure 1.3: Some simplicial cells in one, two and three dimensions. These are purely geometric objects.

functions. For our current example, we will use a triangular mesh and the space

$$V_h = \{\text{all continuous functions that are piecewise linear when restricted to a cell}\}.$$

Such a function is uniquely determined by its values at the vertices of the mesh. That is, the function has one degree of freedom that we will solve for at each the vertex of the mesh, figure 1.4. With other choices of V_h, the degrees of freedom will be associated with different geometric entities; for example, if we chose V_h to be the space of discontinuous functions that are piecewise constant over each cell, each cell would have exactly one degree of freedom (conventionally taken at its barycentre).

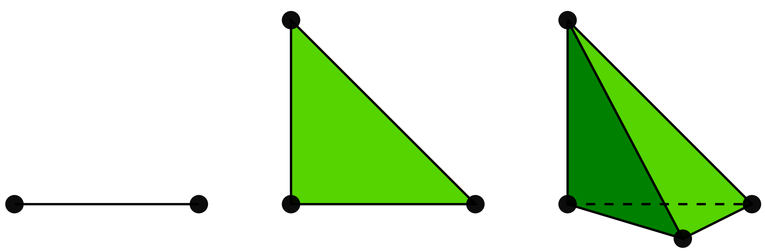

Figure 1.4: The linear Lagrange element in one, two and three dimensions. The black circles denote pointwise evaluation. These pictures describe what values need to be stored to define a function on the cell.

As a basis for the space, we will use the *nodal* basis, which is defined by the following property.

Definition 1.4.2 (nodal basis)**.** *Given the locations of N degrees of freedom x_i, $i = 0, \ldots, N-1$, the associated nodal basis ϕ_i, $i = 0, \ldots, N-1$ satisfies*

$$\phi_i(x_j) = \delta_{ij}, \tag{1.4.1}$$

where δ_{ij} is the Kronecker delta.

This choice of discrete function space has many advantages. Most basis functions decouple, i.e. their supports do not intersect, which will yield *sparsity* of the resulting linear system (most entries in our matrix will be zero). By allowing for arbitrary geometric decompositions, complicated geometries can be well approximated. Lastly, the function space can be made larger in one of two ways: either the mesh can be refined (suitable for problems with nonsmooth solutions), or the order of the interpolating polynomials can be increased (suitable for problems with smooth solutions). The flexibility offered by these two strategies is very powerful for difficult problems.

1.5 Linear algebraic formulation

We now have decided on our V_h and chosen a basis for the space:

$$V_h = \text{span}\{\phi_1, \dots, \phi_N\}. \tag{1.5.1}$$

Let us expand u_h and v_h in terms of this basis, i.e.

$$u_h = \sum_{j=1}^{N} U_j \phi_j \tag{1.5.2}$$

and

$$v_h = \sum_{i=1}^{N} V_i \phi_i. \tag{1.5.3}$$

Our aim is to calculate U_i, the coefficients of our approximate solution. We will now show that the Galerkin approximation (G) is equivalent to a linear system of equations.

First expand v_h in our Galerkin approximation:

$$a(u_h, v_h) = F(v_h) \tag{1.5.4}$$

$$\implies a(u_h, \sum_i V_i \phi_i) = F(\sum_i V_i \phi_i) \tag{1.5.5}$$

$$\implies \sum_i V_i a(u_h, \phi_i) = \sum_i V_i F(\phi_i). \tag{1.5.6}$$

As this has to hold for all possible values of V_i, this is equivalent to

$$a(u_h, \phi_i) = F(\phi_i) \text{ for } i = 1, \dots, N. \tag{1.5.7}$$

Each test function ϕ_i will yield *one row* of the resulting matrix.

Now expand u_h:

$$a(\sum_j U_j \phi_j, \phi_i) = F(\phi_i) \tag{1.5.8}$$

$$\implies \sum_j a(\phi_j, \phi_i) U_j = F(\phi_i) \tag{1.5.9}$$

$$\tag{1.5.10}$$

or in matrix notation

$$AU = b, \tag{1.5.11}$$

where

$$A_{ij} = a(\phi_j, \phi_i) \tag{1.5.12}$$

and

$$b_i = F(\phi_i). \tag{1.5.13}$$

This linear system can then be solved with techniques from numerical linear algebra (e.g. Gaussian elimination or a Krylov method).

1.6 Outlook

We have now sketched the finite element method, but many questions remain. When is the abstract variational formulation (Q) well-posed? When is the Galerkin approximation (G) well-posed? What can be said about the approximation error $u - u_h$, as measured in different norms? How can we solve nonlinear problems, or coupled PDEs?

2 Elements of functional analysis: Lebesgue spaces

We now embark on a campaign to prove that (Q) is indeed well-posed. Our goal for the next few lectures is to prove the *Lax-Milgram* Theorem, a fundamental result in PDEs and variational analysis. Lax-Milgram will give us well-posedness of both (Q) and its Galerkin approximation (G). In order to state and prove the theorem, we must first introduce some basic concepts of functional analysis[1].

[1] This material is treated more thoroughly in C4.3, Functional Analytic Methods for PDEs. However, C4.3 is not a prerequisite for this course, and we will introduce all of the functional analysis we need as we need it.

2.1 Banach spaces

Definition 2.1.1 (normed vector space)**.** *A* normed vector space X *is a vector space equipped with a norm* $\|\cdot\| : X \to \mathbb{R}$ *that satisfies the following properties:*

1. $\|x\| \geq 0$, *and* $\|x\| = 0 \iff x = 0$;
2. $\|\alpha x\| = |\alpha|\|x\|$ *for any scalar* $\alpha \in \mathbb{R}$;
3. $\|x + y\| \leq \|x\| + \|y\|$.

Recall that *completeness* of a normed vector space X means that all Cauchy sequences[2] in X converge in X.

[2] Recall that a *Cauchy sequence* is one where its elements get arbitrarily close to each other.

Definition 2.1.2 (Banach space)**.** *A Banach space is a complete normed vector space.*

Example 2.1.3. *Euclidean space* $\mathbb{R}^n$ *equipped with the 1-norm* $\|\cdot\|_1$*, the 2-norm* $\|\cdot\|_2$*, or the supremum norm* $\|\cdot\|_\infty$ *are all Banach spaces.*

Example 2.1.4. *The space of continuous functions from a domain* Ω *to* $\mathbb{R}$ *equipped with the supremum norm*

$$\|f\|_\infty = \sup\{|f(x)| : x \in \Omega\} \tag{2.1.1}$$

is a Banach space.

2.2 *Hilbert spaces*

Definition 2.2.1 (inner product space)**.** *An* inner product space X *is a vector space equipped with an inner product* $(\cdot,\cdot) : X \times X \to \mathbb{R}$ *that satisfies the following properties:*

1. $(u, v) = (v, u)$;
2. $(\alpha u + \beta v, w) = \alpha(u, w) + \beta(v, w)$ *for* $\alpha, \beta \in \mathbb{R}$;
3. $(u, u) \geq 0$ *with* $(u, u) = 0 \iff u = 0$.

An inner product induces a norm, $\|u\| = \sqrt{(u, u)}$, and hence a metric.

Definition 2.2.2 (Hilbert space)**.** *A* Hilbert space X *is a complete inner product space.*

Thus a Hilbert space is also a Banach space. In this course we will restrict ourselves to real Hilbert spaces.

Example 2.2.3. *The canonical example of a Hilbert space is* $\mathbb{R}^n$ *with inner product*

$$(u, v)_{\mathbb{R}^n} = u^T v. \tag{2.2.1}$$

This induces the $\|\cdot\|_2$ *norm.*

Example 2.2.4. *The space of square-integrable functions on a domain,* $L^2(\Omega)$*, is a Hilbert space with inner product*

$$(u, v)_{L^2(\Omega)} = \int_\Omega uv \,\mathrm{d}x. \tag{2.2.2}$$

Example 2.2.5. *The space* $H_0^1(\Omega)$ *of square-integrable functions that are zero on the boundary and that have square-integrable derivatives is a Hilbert space with inner product*

$$(u, v)_{H_0^1(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x. \tag{2.2.3}$$

This satisfies the last condition of the definition of an inner product because if $\nabla u = 0$ *then* u *must be constant; but the only constant function in* $H_0^1(\Omega)$ *is the zero function, because of the boundary conditions.*

Example 2.2.6. *For* $\Omega \subset \mathbb{R}^n$*, the space* $H(\mathrm{div}, \Omega)$ *of square-integrable vector-valued functions with square-integrable divergence*

$$H(\mathrm{div}, \Omega) = \{v \in L^2(\Omega; \mathbb{R}^n) : \nabla \cdot v \in L^2(\Omega)\} \tag{2.2.4}$$

is a Hilbert space with inner product

$$(u, v)_{H(\mathrm{div},\Omega)} = \int_\Omega u \cdot v + \nabla \cdot u \nabla \cdot v \,\mathrm{d}x. \tag{2.2.5}$$

Example 2.2.7. *For $\Omega \subset \mathbb{R}^3$, the space $H(\mathrm{curl}, \Omega)$ of square-integrable vector-valued functions with square-integrable curl*

$$H(\mathrm{curl}, \Omega) = \{v \in L^2(\Omega; \mathbb{R}^3) : \nabla \times v \in L^2(\Omega; \mathbb{R}^3)\} \tag{2.2.6}$$

is a Hilbert space with inner product

$$(u, v)_{H(\mathrm{curl},\Omega)} = \int_\Omega u \cdot v + \nabla \times u \cdot \nabla \times v \, \mathrm{d}x. \tag{2.2.7}$$

An important fact about Hilbert spaces is the Cauchy–Schwarz inequality.

Theorem 2.2.8 (Cauchy–Schwarz inequality)**.** *For a Hilbert space X and any $u, v \in X$,*

$$|(u, v)_X| \leq \|u\|_X \|v\|_X. \tag{2.2.8}$$

Proof. Let $\lambda \in \mathbb{R}$. Then

$$\begin{aligned} 0 \leq \|u + \lambda v\|_X^2 &= (u + \lambda v, u + \lambda v)_X \\ &= (u, u) + (u, \lambda v) + (\lambda v, u) + (\lambda v, \lambda v) \\ &= \|u\|_X^2 + 2\lambda(u, v) + \lambda^2 \|v\|_X^2. \end{aligned} \tag{2.2.9}$$

The right-hand side is a quadratic polynomial in λ with real coefficients, and it is non-negative for all $\lambda \in \mathbb{R}$. Therefore its discriminant is non-positive; it can only be zero or negative. Thus,

$$|2(u, v)_X|^2 - 4\|u\|_X^2 \|v\|_X^2 \leq 0, \tag{2.2.10}$$

yielding the desired inequality. □

The only property of a norm that is not immediately obvious from the properties of an inner product is the triangle inequality. In fact, the triangle inequality follows from Cauchy–Schwarz:

Corollary 2.2.9 (Triangle inequality)**.** *Let $u, v \in X$. Then*

$$\|u + v\| \leq \|u\| + \|v\|. \tag{2.2.11}$$

Proof.

$$\|u + v\|^2 = (u + v, u + v) = \|u\|^2 + 2(u, v) + \|v\|^2 \tag{2.2.12}$$

$$\leq \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 \tag{2.2.13}$$

$$= (\|u\| + \|v\|)^2. \tag{2.2.14}$$

□

The Cauchy–Schwarz inequality also ensures that the definition of angle makes sense in Hilbert spaces. We define the angle θ between two vectors u and v by

$$\cos\theta = \frac{(u, v)_X}{\|u\|_X \|v\|_X}, \tag{2.2.15}$$

which is always in $[-1, 1]$ by Cauchy–Schwarz.

2.3 Dual of a Hilbert space

Definition 2.3.1 (Linear functional on a Hilbert space)**.** *Given a Hilbert space X, a* linear functional *j on X is a function $j : X \to \mathbb{R}$ that satisfies*

$$j(\alpha u + \beta v) = \alpha j(u) + \beta j(v). \tag{2.3.1}$$

Example 2.3.2. *Integration over a fixed domain, evaluation at a fixed point x, and evaluation of the derivative at a point x in the direction v are all examples of linear functionals (when they are defined!).*

Definition 2.3.3 (Bounded linear functional)**.** *A* bounded *linear functional $j : X \to \mathbb{R}$ is one for which there exists $L \in [0, \infty)$ such that*

$$|j(u)| \leq L\|u\|_X \quad \forall\, u \in X. \tag{2.3.2}$$

Lemma 2.3.4 (Boundedness and continuity)**.** *Boundedness is equivalent to continuity.*

Proof. First, suppose a linear functional j is bounded and let $u, v \in X$. Then

$$0 \leq |j(u+v) - j(u)| = |j(v)| \leq L\|v\|_X \tag{2.3.3}$$

which goes to zero as $v \to 0$. Thus j is continuous at u.

Now suppose j is continuous. In particular, it is continuous at 0. Thus, there exists a $\delta > 0$ such that if $\|h - 0\| \leq \delta$, then $|j(h) - j(0)| = |j(h)| \leq 1$. Then, for any $u \in X$,

$$|j(u)| = \left| \frac{\|u\|}{\delta} j\left(\delta \frac{u}{\|u\|}\right) \right| = \frac{\|u\|}{\delta} \left| j\left(\delta \frac{u}{\|u\|}\right) \right| \leq \frac{1}{\delta}\|u\|, \tag{2.3.4}$$

which shows that j is bounded. □

Thus, the words boundedness and continuity are used interchangeably.

Definition 2.3.5 (Dual of a Hilbert space)**.** *The* dual X^* of a Hilbert space *X is the space of all bounded linear functionals on X. This has a natural norm induced by the norm on the underlying space:*

$$\|j\|_{X^*} = \sup_{\|u\|_X = 1} |j(u)|. \tag{2.3.5}$$

In fact, this is itself a Hilbert space[3]*.*

Given a $j \in X^*$, denote the action of j on u (or equivalently u on j) by

$$\langle j, u \rangle = j(u). \tag{2.3.6}$$

This is called the *duality pairing*.

[3] The inner product on X^* can be constructed using the parallelogram law and the polarisation identity; we won't need these.

Example 2.3.6. *The functional* $j : L^2(\Omega) \to \mathbb{R}$

$$j(v) = \int_\Omega v \, \mathrm{d}x \tag{2.3.7}$$

is an element of the dual space of $L^2(\Omega)$.

Example 2.3.7. *The functional* $F(v)$ *defined in* (1.2.10) *is an element of the dual space of* $V = H_0^1(\Omega)$.

2.4 The Riesz Representation theorem

There is a fundamental connection between a Hilbert space and its dual. This connection is captured in the Riesz Representation Theorem.

Theorem 2.4.1 (Riesz Representation Theorem)**.** *Any bounded linear functional* $j \in X^*$ *can be uniquely represented by a* $g \in X$*, via*

$$\langle j, u \rangle = (g, u). \tag{2.4.1}$$

Moreover, the norms agree: $\|j\|_{X^*} = \|g\|_X$.

This defines a canonical linear map, the *Riesz map* $\mathcal{R} : X^* \to X$, that maps $j \mapsto g$. This Riesz map is an isometric isomorphism[4].

[4] In the complex case, you have to take some conjugates of scalars, and so it is called an anti-isomorphism.

Example 2.4.2. *Let* $X = L^2(\Omega)$ *and let*

$$j(v) = \langle j, v \rangle = \int_\Omega v \, \mathrm{d}x. \tag{2.4.2}$$

Then its $L^2(\Omega)$ *Riesz representation is the constant function* $g(x) = 1$.

2.5 Lebesgue spaces

The Lebesgue spaces, denoted $L^p(\Omega)$, are fundamental function spaces that appear throughout analysis. Their purpose (for us) is to finely capture the allowed rate of blow-up of a function at singularities: the larger the p, the slower the blow-up allowed.

Definition 2.5.1 (Lebesgue p-norm, finite p)**.** *Let* $p \in [1, \infty)$*. The* $L^p(\Omega)$ *norm is defined by*

$$\|u\|_{L^p(\Omega)} = \left(\int_\Omega |u|^p \, \mathrm{d}x \right)^{1/p}. \tag{2.5.1}$$

Definition 2.5.2 (Lebesgue p-norm, infinite p)**.** *The* $L^\infty(\Omega)$ *norm is defined by*

$$\|u\|_{L^\infty(\Omega)} = \mathit{ess\ sup}\{|u(x)| : x \in \Omega\}, \tag{2.5.2}$$

where the essential supremum is the smallest real number c *such that* $\{x \in \Omega : |u(x)| > c\}$ *has measure zero, i.e. has zero volume (in 3D) or area (in 2D)*[5].

[5] If you haven't done a course in measure theory, just think of it as the supremum and forget about it.

Definition 2.5.3 (Lebesgue space)**.** *For $p \in [1, \infty]$, the $L^p(\Omega)$ space is defined by*

$$L^p(\Omega) = \{u : \Omega \to \mathbb{R} \ : \ \|u\|_{L^p(\Omega)} < \infty\}. \tag{2.5.3}$$

Remark 2.5.4 (for measure theorists)**.** *The above definition is not* quite *true. Since we are concerned with integrability, and the Lebesgue integral ignores anything happening on a set of measure zero, we actually have to be a little more subtle. Define two measurable functions to be* equivalent *if they differ only on a set of measure zero. This is easily checked to be an equivalence relation. The actual definition of $L^p(\Omega)$ consists of the definition above, quotiented out by this equivalence relation. Thus, an element $[f] \in L^p(\Omega)$ is not a single function; it is an* entire equivalence class of functions, *all of which differ up to sets of measure zero.*

The important consequence of this is that *functions in $L^p(\Omega)$ cannot be evaluated pointwise* in the traditional sense, as a point has measure zero and for any equivalence class $[f]$ you can get any answer you like for $f(x)$. In order to evaluate a function pointwise (for that is quite a useful thing to do sometimes), we have to prove that any equivalence class in the function space has a continuous representative, which we can then evaluate[6].

[6] We'll now go back to saying a function lives in $L^p(\Omega)$, rather than the equivalence class of the function.

Theorem 2.5.5. *All Lebesgue spaces for $p \in [1, \infty]$ are Banach spaces. The choice $p = 2$ is also a Hilbert space, as observed above.*

Let's build some intuition about these spaces. Let us restrict ourselves to bounded Ω, i.e. domains of finite measure; we will always pose our PDEs on bounded domains, and so this is the case of interest to us[7].

[7] The relationships between $L^p(\Omega)$ spaces are different in the case of unbounded Ω: you have to worry that the function decays to zero sufficiently fast at infinity, as well as its blow-up properties.

Example 2.5.6. *First, note that the function $f(x) = 1$ is in $L^p(\Omega)$ for all p:*

$$\|1\|_{L^p(\Omega)} = \left(\int_\Omega 1^p \, \mathrm{d}x\right)^{1/p} \tag{2.5.4}$$

$$= V^{1/p} < \infty, \tag{2.5.5}$$

where V is the measure (e.g. volume or area) of the domain.

Let's consider some examples of functions that distinguish between different $L^p(\Omega)$ spaces.

Example 2.5.7. *Let $\Omega = (0, 1)$ and let*

$$f_q(x) = x^{-q}. \tag{2.5.6}$$

Then $f_q \in L^p(\Omega) \iff q < 1/p$. That is,

- $\frac{1}{x} \notin L^1(\Omega)$*, but* $\frac{1}{x^{0.999}} \in L^1(\Omega)$*;*

- $\frac{1}{\sqrt{x}} \notin L^2(\Omega)$, *but* $\frac{1}{x^{0.4999}} \in L^2(\Omega)$, *etc.*

In other words, the larger the p, the slower the allowed rate of blow-up at singularities. This process continues until $L^\infty(\Omega)$, for which no blow-up whatsoever is allowed.

Next, we state a fundamental result in the analysis of $L^p(\Omega)$ spaces.

Theorem 2.5.8 (Hölder's inequality)**.** *Let $p, q \in [1, \infty]$ such that*

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{2.5.7}$$

The elements of such a pair are called Hölder conjugates[8]*. If $f \in L^p(\Omega)$ and $g \in L^q(\Omega)$, then $fg \in L^1(\Omega)$ and*

$$\|fg\|_{L^1(\Omega)} \leq \|f\|_{L^p(\Omega)} \|g\|_{L^q(\Omega)}. \tag{2.5.8}$$

[8] By convention here, $1/\infty = 0$, so 1 and ∞ are conjugate.

Corollary 2.5.9. *For $1 < p < \infty$, the dual of $L^p(\Omega)$ is isomorphic to $L^q(\Omega)$, where q is conjugate to p. This means that these $L^p(\Omega)$ spaces are* reflexive*, i.e.*

$$(L^p(\Omega))^{**} = L^p(\Omega) \quad \text{for } 1 < p < \infty. \tag{2.5.9}$$

The dual of $L^1(\Omega)$ is $L^\infty(\Omega)$, but the dual of $L^\infty(\Omega)$ is not $L^1(\Omega)$ (it is a much larger space); these spaces are not reflexive.

With this, we can prove the fundamental inclusion for $L^p(\Omega)$ spaces on bounded domains.

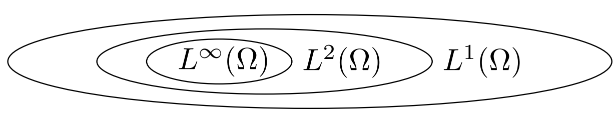

Figure 2.1: The nesting of $L^p(\Omega)$ spaces for bounded domains.

Theorem 2.5.10 (Inclusion of Lebesgue spaces)**.** *Let Ω be bounded. Let $1 \leq p < q \leq \infty$. If $f \in L^q(\Omega)$, then $f \in L^p(\Omega)$.*

Proof.

$$\begin{aligned} \|f\|^p_{L^p(\Omega)} &= \int_\Omega |f|^p \,\mathrm{d}x && (2.5.10)\\ &= \int_\Omega |f|^p 1 \,\mathrm{d}x && (2.5.11)\\ &\leq \|f^p\|_{L^\alpha(\Omega)} \|1\|_{L^\beta(\Omega)} && (2.5.12) \end{aligned}$$

for any choice α and β that are Hölder conjugates, so long as the quantities are defined. Choose $\alpha = q/p$, $\beta = q/(q-p)$. Then

$$\begin{aligned} \|f^p\|_{L^\alpha(\Omega)} &= \left(\int_\Omega (|f|^p)^{q/p} \,\mathrm{d}x\right)^{p/q} && (2.5.13)\\ &= \left(\int_\Omega |f|^q \,\mathrm{d}x\right)^{p/q} && (2.5.14)\\ &= \|f\|^p_{L^q(\Omega)} < \infty, && (2.5.15) \end{aligned}$$

and since the volume of the domain is finite, we have

$$\|f\|_{L^p(\Omega)} < \infty. \tag{2.5.16}$$

$\square$

Example 2.5.11 (A function that is in $L^p(\Omega)$ for all finite p but not in $L^\infty(\Omega)$)**.** *A natural question to ask is the following: if $f \in L^p(\Omega)$ for all $p \in [1, \infty)$, does it follow that $f \in L^\infty(\Omega)$? The answer is no. Let $\Omega = (0, 1)$, and consider*

$$f(x) = \log(x). \tag{2.5.17}$$

To see that $f \in L^p(\Omega)$ for all finite p, observe that

$$\lim_{x \to 0^+} \frac{\log(x)}{x^{-1/2p}} = 0 \tag{2.5.18}$$

by l'Hôpital's rule. Thus, $|\log(x)|^p < x^{-1/2}$ for sufficiently small x, and the $L^p(\Omega)$ norm is finite. However, f is unbounded below on $(0, 1)$, so $f \notin L^\infty(\Omega)$.

3 Elements of functional analysis: Sobolev spaces

We saw earlier that the classical notion of differentiability, e.g. $u \in C^2(\mathbb{R}^n)$, was too restrictive: the Poisson equation didn't always have a solution for $f \in C^0(\mathbb{R}^n)$ data. We need to introduce a new concept of differentiation with respect to space to ensure that this equation is indeed well-posed.

3.1 Weak derivatives

To motivate the definition, first suppose $f \in C^1(a,b)$, i.e. f' exists in $C^0(a,b)$. Let ϕ be a differentiable function that is zero on the boundary $\{a,b\}$. Then integration by parts tells us

$$\int_a^b f'\phi \, \mathrm{d}x = -\int_a^b f\phi' \, \mathrm{d}x, \tag{3.1.1}$$

i.e. we can swap the differentiation operator onto the test function ϕ. This is how we will *define* the weak derivative f' in Lebesgue spaces.

Definition 3.1.1 (Compact support in Ω). *A function $\phi \in C(\Omega)$ has* compact support *iff*

$$\textit{supp}(\phi) = \textit{closure}\{x \in \Omega : \phi(x) \neq 0\} \tag{3.1.2}$$

is compact (i.e. is bounded, as it is closed by construction) and is a subset of the interior *of Ω. In particular, this means that ϕ vanishes on Γ.*

Definition 3.1.2 (Bump functions). *The set of* bump functions*, denoted $C_0^\infty(\Omega)$, is the set of $C^\infty(\Omega)$ functions that have compact support in Ω.*

Example 3.1.3. *Let $\Omega = (-2,2)$. The function*

$$\Psi(x) = \begin{cases} \exp\left(-\frac{1}{1-x^2}\right) & \textit{if } |x| < 1 \\ 0 & \textit{otherwise} \end{cases} \tag{3.1.3}$$

is in $C_0^\infty(\Omega)$[1].

[1] See Brenner & Scott page 27 for the proof.

We are going to use these (very nice, infinitely differentiable, zero at the boundaries) functions to define the weak derivatives of functions that aren't so smooth. Before we do this, though, we will define the set of functions for which we will define weak derivatives.

Definition 3.1.4 (Locally integrable functions)**.** *Given a domain* Ω*, the set of* locally integrable functions *is defined by*

$$L^1_{loc}(\Omega) = \{f : f \in L^1(K) \text{ for all compact } K \subset \text{ interior } \Omega\}. \tag{3.1.4}$$

This set includes $L^1(\Omega)$ and $C^0(\Omega)$ as subsets.

Let's first define the first derivative, then generalise that definition to higher derivatives.

Definition 3.1.5 (Weak first derivative)**.** *Let* $\Omega \subset \mathbb{R}^n$*. We say that a given function* $f \in L^1_{loc}(\Omega)$ *has a weak* i^{th} *partial derivative* $\partial f/\partial x_i$ *if there exists a function* $g \in L^1_{loc}(\Omega)$ *such that*

$$\int_\Omega g\phi \, \mathrm{d}x = -\int_\Omega f \frac{\partial \phi}{\partial x_i} \, \mathrm{d}x \quad \text{for all } \phi \in C_0^\infty(\Omega). \tag{3.1.5}$$

Compare this definition to (3.1.1).

Theorem 3.1.6 (Uniqueness of weak derivatives)**.** *Weak derivatives are unique, up to a set of measure zero.*

Example 3.1.7. *Any strongly differentiable function has a weak derivative.*

Example 3.1.8. *Let* $\Omega = (-1, 1)$ *and take* $f(x) = |x|$*. Then it has a weak derivative* f' *given by*

$$f' = \begin{cases} -1 & x < 0 \\ 1 & x > 0. \end{cases} \tag{3.1.6}$$

To verify this, break the interval into the two parts in which f is smooth, and integrate by parts. Let $\phi \in C_0^\infty(\Omega)$. Then

$$\int_{-1}^{1} f(x)\phi'(x) \, \mathrm{d}x = \int_{-1}^{0} f(x)\phi'(x) \, \mathrm{d}x + \int_{0}^{1} f(x)\phi'(x) \, \mathrm{d}x \tag{3.1.7}$$

$$= -\int_{-1}^{0} (-1)\phi(x) \, \mathrm{d}x + [f\phi]_{-1}^{0} - \int_{0}^{1} (+1)\phi(x) \, \mathrm{d}x + [f\phi]_{0}^{1} \tag{3.1.8}$$

$$= -\int_{-1}^{1} f'(x)\phi(x) \, \mathrm{d}x + \big((f\phi)(0^-) - (f\phi)(0^+)\big) \tag{3.1.9}$$

$$= -\int_{-1}^{1} f'(x)\phi(x) \, \mathrm{d}x. \tag{3.1.10}$$

Example 3.1.9. *Extending the previous example, any continuous piecewise-differentiable function is weakly differentiable. This is important to us because we will approximate the solutions of our PDEs with continuous, piecewise-differentiable functions (differentiable on each element)*[2]*.*

[2] It is sometimes useful to approximate functions with discontinuous polynomials, in the so-called discontinuous Galerkin discretisation. This is beyond the scope of this course.

Example 3.1.10. *A counterexample: take* $\Omega = (-1, 1)$ *and take* $f(x) = sign(x)$*, i.e.*

$$f(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0. \end{cases} \tag{3.1.11}$$

This function has no weak derivative.

An informal proof: the only candidate f' would be $f' \equiv 0$, but the discontinuity at $x = 0$ means that the extra terms arising from integration by parts do not vanish.

3.2 *Higher weak derivatives*

To compactly define higher derivatives, we first need to introduce multi-index notation.

Definition 3.2.1 (multi-index)**.** *Let* $\Omega \subset \mathbb{R}^n$*. A multi-index* α *is a tuple of* n *non-negative integers*

$$\alpha = (\alpha_1, \ldots, \alpha_n). \tag{3.2.1}$$

Given a multi-index α *and* $\phi \in C^\infty(\Omega)$*, define*

$$\partial_x^\alpha \phi = \phi^{(\alpha)} = D^\alpha \phi = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} \phi. \tag{3.2.2}$$

The length *of* α *is the order of the derivative,*

$$|\alpha| = \sum_{i=1}^n \alpha_i. \tag{3.2.3}$$

Example 3.2.2. *The multi-index* $(1, 0)$ *corresponds to* $\partial/\partial x_1$*. The multi-index* $(0, 1)$ *corresponds to* $\partial/\partial x_2$*. A sum over* $|\alpha| = 1$ *means to sum over all first order derivatives.*

With this, we can now neatly define weak derivatives of any order.

Definition 3.2.3 (Weak derivative)**.** *Let* $\Omega \subset \mathbb{R}^n$*. We say that a given function* $f \in L^1_{loc}(\Omega)$ *has a weak derivative* $D^\alpha f$ *provided there exists a function* $g \in L^1_{loc}(\Omega)$ *such that*

$$\int_\Omega g\phi \, \mathrm{d}x = (-1)^{|\alpha|} \int_\Omega f\phi^{(\alpha)} \, \mathrm{d}x \quad \textit{for all } \phi \in C_0^\infty(\Omega). \tag{3.2.4}$$

3.3 *Sobolev spaces*

We saw before that Lebesgue spaces control the allowed rate of blowup of their functions, described by a parameter p. We will extend this idea to *Sobolev spaces*. Sobolev spaces are indexed by two parameters: k will describe the number of (weak) derivatives demanded, while p will describe the allowed rate of blowup of the function and all its derivatives. Sobolev spaces are important because they turn out to be the natural space in which to look for solutions of PDEs, and provide the natural norms in which to measure approximation errors.

Definition 3.3.1 (Sobolev norm)**.** *Let k be a non-negative integer. Let $f \in L^1_{loc}(\Omega)$. Suppose that the weak derivatives $D^\alpha f$ exist for all $|\alpha| \leq k$. For $p \in [1, \infty)$, define the* Sobolev norm

$$\|f\|_{W^k_p(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha f\|^p_{L^p(\Omega)} \right)^{1/p} \tag{3.3.1}$$

and in the case $p = \infty$

$$\|f\|_{W^k_p(\Omega)} = \max_{|\alpha| \leq k} \|D^\alpha f\|_{L^\infty(\Omega)}. \tag{3.3.2}$$

Definition 3.3.2 (Sobolev space)**.** *Define the* Sobolev space $W^k_p(\Omega)$ *as*

$$W^k_p(\Omega) = \{f \in L^1_{loc}(\Omega) : \|f\|_{W^k_p(\Omega)} < \infty\}. \tag{3.3.3}$$

Theorem 3.3.3. *The Sobolev space $W^k_p(\Omega)$ is a Banach space.*

Proof. See theorem 1.3.2 of Brenner & Scott. □

Theorem 3.3.4. *The Sobolev spaces with $p = 2$ are Hilbert spaces. These are denoted by*

$$H^k(\Omega) = W^k_2(\Omega). \tag{3.3.4}$$

Let's now start to become acquainted with these fundamental objects. We'll begin by stating some relationships between different Sobolev spaces that follow easily from the definition.

Example 3.3.5. *The space $W^0_p(\Omega) = L^p(\Omega)$. That is, if we ask for no weak derivatives, we just get the $L^p(\Omega)$ space back.*

Example 3.3.6. *The space $W^1_\infty(\Omega)$ is equivalent to the space of Lipschitz continuous functions under certain restrictions on the domain Ω[3]. This implies that $W^1_\infty(\Omega) \subset C(\Omega)$, and that any $f \in W^1_\infty(\Omega)$ can be evaluated pointwise.*

[3] Specifically, we require either a C^1 boundary (see Evans, section 5.8.2) or a property called *quasiconvexity*: there exists a constant $M < \infty$ such that two points a and b in the domain can be joined by a curve of length at most $M\|a - b\|$. All of the domains we'll see are quasiconvex. For a guided proof in the case where Ω is convex, see Brenner & Scott exercises 1.x.14 and 1.x.15.

Example 3.3.7. *Suppose $l \geq k$. Then $W^l_p(\Omega) \subset W^k_p(\Omega)$; we're just asking for fewer derivatives.*

Example 3.3.8. *Suppose $1 \leq p \leq q \leq \infty$ and that Ω is bounded. Then $W^k_q(\Omega) \subset W^k_p(\Omega)$. This follows from the Lebesgue inclusion theorem 2.5.10.*

Example 3.3.9. *If $f \in W^k_p(\Omega)$, and α is a multi-index with $|\alpha| = n \leq k$, then $D^\alpha f \in W^{k-n}_p(\Omega)$.*

3.4 Sobolev embeddings

There are other inclusions between Sobolev spaces that are less obvious. These will be encoded in *Sobolev's inequality*. However, in order for the result to be true, we will need an additional regularity requirement on the domain Ω.

Definition 3.4.1 (Lipschitz domain, informal)**.** *We say* Ω *is a* Lipschitz domain, *or has Lipschitz boundary, if* $\partial\Omega$ *is everywhere locally the graph of a Lipschitz continuous function*[4].

[4] For a formal definition, see Brenner & Scott, definition 1.4.4.

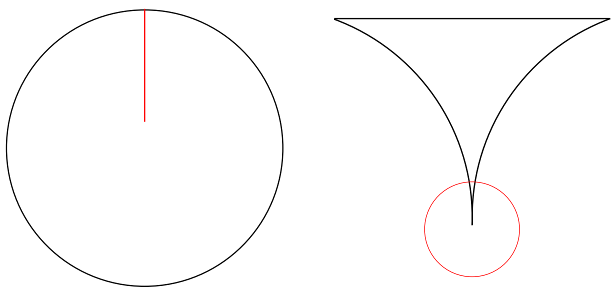

Figure 3.1: Some non-Lipschitz domains, with the geometric feature at fault highlighted.

For some examples of non-Lipschitz domains, see figure 3.1, reproduced from Hiptmair's notes[5].

[5] R. Hiptmair and C. Schwab. Numerical methods for elliptic and parabolic boundary value problems, 2008. http://www.sam.math.ethz.ch/~hiptmair/tmp/NAPDE_08.pdf

This regularity condition is important: without it, the Sobolev inequality *is not true*. Henceforth, we assume that Ω is a Lipschitz domain.

There are three numbers describing a Sobolev space: n, the dimension of the domain, k, the number of weak derivatives possessed, and p, the integrability of those derivatives. Sobolev's inequality tells us that if you possess enough weak derivatives that are integrable enough, then your function is continuous and bounded.

Theorem 3.4.2 (Sobolev's inequality)**.** *Let* $\Omega \subset \mathbb{R}^n$ *be a Lipschitz domain. Let* k *be a positive integer and let* $p \in [1, \infty)$. *Suppose*

$$k \geq n \text{ when } p = 1 \tag{3.4.1}$$

$$k > n/p \text{ when } p > 1. \tag{3.4.2}$$

Then there is a constant C *such that for all* $u \in W_p^k(\Omega)$,

$$\|u\|_{L^\infty(\Omega)} \leq C\|u\|_{W_p^k(\Omega)}, \tag{3.4.3}$$

and moreover there is a continuous function in the equivalence class of u.

Note that we don't need the case $p = \infty$, because if $k \geq 1$ and $p = \infty$ you are already Lipschitz continuous on a quasiconvex domain and hence continuous.

Let's look at some consequences of this.

Example 3.4.3. *In one dimension ($n = 1$), we have that the existence of* a single *weak derivative of any integrability is enough to ensure continuity. This is why the discontinuous sign function of example 3.1.10 could not have a weak derivative. In fact, a piecewise smooth function on a bounded domain Ω is in $H^1(\Omega)$ if and only if it is continuous. For a proof, see Braess*[6]*, theorem 5.2.*

[6] D. Braess. *Finite Elements: theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, third edition, 2007

Example 3.4.4. *In two dimensions ($n = 2$), we have $W_1^1(\Omega) \not\subset C(\Omega)$, but $W_1^2(\Omega) \subset C(\Omega)$.*

Example 3.4.5. *In three dimensions ($n = 3$), we have $W_1^2(\Omega) \not\subset C(\Omega)$, but $W_1^3(\Omega) \subset C(\Omega)$.*

Example 3.4.6. *Let's look at the continuity properties of the Hilbert spaces $H^k(\Omega)$, i.e. $p = 2$. With $p = 2$, Sobolev's inequality tells us that we need*

$$k > n/2. \tag{3.4.4}$$

In one dimension,

$$H^1(\Omega) \subset C(\Omega). \tag{3.4.5}$$

For $n = 2$, Sobolev's inequality tells us we need $k > 1$, i.e. $k \geq 2$, so in two dimensions

$$H^1(\Omega) \not\subset C(\Omega), \quad H^2(\Omega) \subset C(\Omega). \tag{3.4.6}$$

For $n = 3$, Sobolev's inequality tells us we need $k > 1.5$, so $k \geq 2$ is again sufficient.

By applying theorem 3.4.2 to the derivatives of functions in Sobolev spaces, one derives the following corollary:

Corollary 3.4.7. *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain. Let k and m be positive integers satisfying $m < k$ and let $p \in [1, \infty)$. Suppose*

$$k - m \geq n \text{ when } p = 1 \tag{3.4.7}$$

$$k - m > n/p \text{ when } p > 1. \tag{3.4.8}$$

Then there is a constant C such that for all $u \in W_p^k(\Omega)$,

$$\|u\|_{W_\infty^m(\Omega)} \leq C\|u\|_{W_p^k(\Omega)}, \tag{3.4.9}$$

and moreover there is a $C^m(\Omega)$ function in the equivalence class of u.

There are entire books written about Sobolev spaces and embedding theorems, e.g. Maz'ya[7]; we have only scratched the surface here. However, there is another (nonexaminable!) embedding theorem that is quite interesting that I wish to quote for our education. The following summarises various results from Evans, chapter 5[8].

[7] V. Maz'ya. *Sobolev Spaces*, volume 342 of *A Series of Comprehensive Studies in Mathematics*. Springer, 2011

[8] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010

Theorem 3.4.8. *Let*

$$p^* = \begin{cases} \frac{np}{n-kp} & \text{if } kp < n \\ \infty & \text{if } kp \geq n. \end{cases} \tag{3.4.10}$$

Let Ω be a bounded domain with Lipschitz boundary. Then

$$W_p^k(\Omega) \subset L^q(\Omega), \tag{3.4.11}$$

where $q \in [1, p^*]$ if $kp \neq n$ and $q \in [1, \infty)$ if $kp = n$.

Example 3.4.9. *Consider $H^1(\Omega) = W_2^1(\Omega)$. In one dimension, $kp = 2 > 1 = n$ so $p^* = \infty$ and $H^1(\Omega) \subset L^\infty(\Omega)$. In two dimensions, $kp = n$ so $p^* = \infty$ and $H^1(\Omega) \subset L^q(\Omega)$ for any $q \in [1, \infty)$. In three dimensions, $kp < 3 = n$ so $p^* = 6$ and $H^1(\Omega) \subset L^q(\Omega)$ for any $q \in [1, 6]$*[9].

[9] The embedding into $L^q(\Omega)$ is compact for $q \in [1, 6)$, but is not compact for $q = 6$.

Example 3.4.10. *Consider $H^2(\Omega) = W_2^2(\Omega)$. Here $kp = 4 > n$ for $n = 1, 2, 3$, so $p^* = \infty$ and $H^2(\Omega) \subset L^\infty(\Omega)$.*

3.5 Review of our variational formulation

We are now in a position to see why the space

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : u|_\Gamma = 0\} \tag{3.5.1}$$

is the "right" one for the problem (Q).

1. We want $v \in L^2(\Omega)$ and $f \in L^2(\Omega)$ to ensure that $F(v)$ is a bounded linear functional.
2. We want $u|_\Gamma = 0$ to satisfy the strongly-imposed boundary conditions.
3. We need the first weak derivatives to exist to talk about ∇u and ∇v.
4. We want u and v to have square-integrable weak derivatives, as this guarantees $a(u, v) < \infty$ (by Cauchy-Schwarz).

4 The Lax–Milgram Theorem

The Lax–Milgram Theorem gives sufficient conditions for a variational problem like (Q) to be well-posed. Now that we understand some aspects of Hilbert, Lebesgue and Sobolev spaces, we are in a position to state and prove the theorem. First, we will define some conditions on the forms arising in problem (Q).

Definition 4.0.1 (Bounded bilinear form)**.** *A bilinear form* $a : H \times H \to \mathbb{R}$ *is said be to* continuous *if there exists* $C \in [0, \infty)$ *such that*

$$|a(v, w)| \leq C\|v\|_H\|w\|_H \quad \textit{for all } v, w \in H. \tag{4.0.1}$$

This property is also called continuity, for the same reason that a linear form is bounded if and only if it is continuous. Most forms one meets in practice are continuous, and it is usually fairly straightforward to prove continuity.

Definition 4.0.2 (Coercive bilinear form)**.** *A bilinear form* $a : H \times H \to \mathbb{R}$ *is said be to* coercive *on* $V \subset H$ *if there exists* $\alpha > 0$ *such that*

$$a(v, v) \geq \alpha\|v\|_H^2 \quad \textit{for all } v \in V. \tag{4.0.2}$$

Coercivity is a much stronger property. For example, the weak formulation of the Stokes equations in fluid mechanics do not enjoy this property, but are still well-posed (when the right function spaces are chosen). Nevertheless, it holds for many important equations of mathematical physics, including the Poisson equation, the sign-positive Helmholtz equation and linear elasticity.

Proving coercivity is nontrivial. In particular, the bilinear form (1.2.9) is indeed coercive, but this is far from obvious. We will discuss this later, after we have reviewed the abstract variational theory.

First, let's examine a simple case, before discussing the situation in general.

4.1 *Symmetric coercive continuous problems*

Coercivity and boundedness are almost enough to yield that a is an inner product. We just need one more ingredient — symmetry.

Theorem 4.1.1. *Let H be a Hilbert space, and suppose $a : H \times H \to \mathbb{R}$ is a symmetric bilinear form that is continuous on H and coercive on a closed subspace $V \subset H$. Then $(V, a(\cdot,\cdot))$ is a Hilbert space.*

Proof. An immediate consequence of coercivity is that if $v \in V$ and $a(v,v) = 0$, then $v = 0$. Symmetry and linearity are assumed, so $a(\cdot,\cdot)$ is an inner product on V.

Denote

$$\|v\|_a = \sqrt{a(v,v)}. \tag{4.1.1}$$

It remains to show that $(V, \|\cdot\|_a)$ is complete. Suppose that $\{v_n\}$ is a Cauchy sequence in $(V, \|\cdot\|_a)$. By coercivity, $\{v_n\}$ is also Cauchy in $(H, \|\cdot\|_H)$. Since H is complete, there exists $v \in H$ such that $v_n \to v$ in the $\|\cdot\|_H$ norm. Since V is closed in H, $v \in V$. Now observe that as a is bounded

$$\|v - v_n\|_a = \sqrt{a(v - v_n, v - v_n)} \leq \sqrt{C\|v - v_n\|_H^2} = \sqrt{C}\|v - v_n\|_H \tag{4.1.2}$$

where C is the boundedness constant for a. Hence $v_n \to v$ in the $\|\cdot\|_a$ norm too, so V is complete with respect to this norm. $\square$

This is a very powerful result. The well-posedness of (Q) follows immediately.

Theorem 4.1.2 (Well-posedness of symmetric continuous coercive variational problems)**.** *Let V be a closed subspace of a Hilbert space H. Let $a : H \times H \to \mathbb{R}$ be a symmetric continuous V-coercive bilinear form, and let $F \in V^*$. Consider the variational problem:*

$$\text{find } u \in V \text{ such that } a(u,v) = F(v) \quad \text{for all } v \in V. \tag{4.1.3}$$

This problem has a unique stable solution.

Proof. Theorem 4.1.1 implies that $a(\cdot,\cdot)$ is an inner product on V, and that (V, a) is a Hilbert space. Apply the Riesz Representation Theorem, theorem 2.4.1.

Stability means that we can find a constant C such that

$$\|u\|_V \leq C\|F\|_{V^*}. \tag{4.1.4}$$

By the Riesz representation theorem, the Riesz map is an isomorphism, so this follows for the norms generated by the inner product with $C = 1$. $\square$

Example 4.1.3. *The variational problem*

$$\text{find } u \in H_0^1(\Omega) \text{ such that } \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x \text{ for all } v \in H_0^1(\Omega) \tag{4.1.5}$$

is well-posed, as $H_0^1(\Omega)$ is a closed subspace of $H^1(\Omega)$, and the bilinear form is coercive, symmetric and bounded[1].

[1] We will discuss the proofs of these properties later.

4.2 *Nonsymmetric coercive continuous problems*

Now let us drop the assumption that $a(u,v) = a(v,u)$.

Theorem 4.2.1 (Lax–Milgram)**.** *Let V be a closed subspace of a Hilbert space H. Let $a : H \times H \to \mathbb{R}$ be a (not necessarily symmetric) continuous V-coercive bilinear form, and let $F \in V^*$. Consider the variational problem:*

$$\text{find } u \in V \text{ such that } a(u,v) = F(v) \quad \text{for all } v \in V. \tag{4.2.1}$$

This problem has a unique stable solution.

Let us discuss some preliminary manipulations before proving this theorem. We now show that it is possible to regard a variational problem (Q) with $a(\cdot,\cdot)$ continuous *as an equality in the dual space*.

Lemma 4.2.2 (Converting a variational problem to an equation in the dual space)**.** *Let $a : V \times V \to \mathbb{R}$ be linear in its second argument and bounded. For any $u \in V$, define a functional via $A : u \mapsto Au$*

$$Au(v) \equiv a(u,v) \quad \text{for all } v \in V. \tag{4.2.2}$$

Then $Au \in V^*$, i.e. $A : V \to V^*$. Furthermore, A is itself linear if a is linear in its first argument.

Proof. Au is a linear operator on V, since

$$Au(\beta v_1 + \gamma v_2) = a(u, \beta v_1 + \gamma v_2) \tag{4.2.3}$$
$$= \beta a(u, v_1) + \gamma a(u, v_2) \tag{4.2.4}$$
$$= \beta Au(v_1) + \gamma Au(v_2). \tag{4.2.5}$$

Au is also continuous, since

$$|Au(v)| = |a(u,v)| \leq C\|u\|_H \|v\|_H \tag{4.2.6}$$

so

$$\|Au\|_{V^*} = \sup_{v \neq 0} \frac{|Au(v)|}{\|v\|_H} \leq C\|u\|_H < \infty. \tag{4.2.7}$$

Thus $Au \in V^*$.

If a is linear in its first argument, linearity of A follows with a similar argument. □

Thus, the variational problem

$$\text{find } u \in V \text{ such that } a(u,v) = F(v) \quad \text{for all } v \in V \tag{4.2.8}$$

is equivalent to

$$\text{find } u \in V \text{ such that } \langle Au, v \rangle = \langle F, v \rangle \quad \text{for all } v \in V. \tag{4.2.9}$$

And since equality of two dual objects means exactly that they have the same output on all possible inputs, this is equivalent to

$$\text{find } u \in V \text{ such that } Au = F, \tag{4.2.10}$$

where the equality is between dual objects, $Au \in V^*$ and $F \in V^*$.

Example 4.2.3. *In the specific case of the homogeneous Dirichlet Laplacian (P), we have $A : H_0^1(\Omega) \to \left(H_0^1(\Omega)\right)^*$. We could symbolically write $A = -\nabla^2$ and interpret*

$$-\nabla^2 u = f \tag{4.2.11}$$

as an equation in the dual of $H_0^1(\Omega)$[2]. We define

$$H^{-1}(\Omega) := \left(H_0^1(\Omega)\right)^* \tag{4.2.12}$$

and can regard the Laplacian as a map $H_0^1(\Omega) \to H^{-1}(\Omega)$.

Now note that it immediately follows that we can write variational problems *as equations in the primal space*. We know from the Riesz Representation Theorem (theorem 2.4.1) that there is an isometric isomorphism $R : V^* \to V$ from the dual of a Hilbert space V^* back to V. By composing these operators, we have the problem

$$\text{find } u \in V \text{ such that } RAu = RF, \tag{4.2.13}$$

where the equality is between *primal* objects, $RAu \in V$ and $RF \in V$.

Our proof will proceed as follows: we will define a map $T : V \to V$ whose fixed point is the solution of our variational problem, and then show it is a contraction, and invoke the Banach contraction mapping theorem. We first recall the contraction mapping theorem[3].

Theorem 4.2.4 (Contraction mapping theorem)**.** *Given a nonempty Banach space V and a mapping $T : V \to V$ satisfying*

$$\|Tv_1 - Tv_2\| \leq M\|v_1 - v_2\| \tag{4.2.14}$$

for all $v_1, v_2 \in V$ and fixed M, $0 \leq M < 1$, there exists a unique $u \in V$ such that

$$u = Tu. \tag{4.2.15}$$

That is, the contraction T has a unique fixed point u.

We now prove the Lax–Milgram Theorem.

Proof. Cast the variational problem

$$\text{find } u \in V \text{ such that } a(u, v) = F(v) \quad \text{for all } v \in V \tag{4.2.16}$$

as the primal equality

$$\text{find } u \in V \text{ such that } RAu = RF \tag{4.2.17}$$

[2] For a characterisation of this space, see Evans, section 5.9.1.

[3] Note that the theorem does not require T to be linear.

as discussed. For a fixed $\rho \in (0, \infty)$, define the affine map $T : V \to V$

$$Tv = v - \rho\,(RAv - RF)\,. \tag{4.2.18}$$

If T is a contraction for some ρ, then there exists a unique fixed point $u \in V$ such that

$$Tu = u - \rho(RAu - RF) = u, \tag{4.2.19}$$

i.e. that $RAu = RF$. We now show that such a ρ exists. For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(RAv_1 - RAv_2)\|_H^2 \\
&= \|v - \rho(RAv)\|_H^2 && \text{(by linearity of } R, A\text{)} \\
&= \|v\|_H^2 - 2\rho(RAv, v) + \rho^2\|RAv\|_H^2 && \text{(by linearity of inner product)} \\
&= \|v\|_H^2 - 2\rho Av(v) + \rho^2 Av(RAv) && \text{(by definition of } R\text{)} \\
&= \|v\|_H^2 - 2\rho a(v, v) + \rho^2 a(v, RAv) && \text{(by definition of } A\text{)} \\
&\le \|v\|_H^2 - 2\rho\alpha\|v\|_H^2 + \rho^2 C\|v\|_H\|RAv\|_H && \text{(coercivity and continuity)} \\
&\le (1 - 2\rho\alpha + \rho^2 C^2)\|v\|_H^2 && \text{(}A\text{ bounded, } R\text{ isometric)} \\
&= (1 - 2\rho\alpha + \rho^2 C^2)\|v_1 - v_2\|_H^2.
\end{aligned}$$

Thus, if we can find a ρ such that

$$1 - 2\rho\alpha + \rho^2 C^2 < 1, \tag{4.2.20}$$

i.e. that

$$\rho(\rho C^2 - 2\alpha) < 0. \tag{4.2.21}$$

If we choose $\rho \in (0, 2\alpha/C^2)$ then T is a contraction and the proof is complete. □

We have proven existence and uniqueness, but one further result is required to prove well-posedness: stability. This is captured in the following theorem.

Theorem 4.2.5. *Under the previous assumptions, the solution of the variational problem satisfies*

$$\|u\|_H \le \frac{1}{\alpha}\|F\|_{V^*}, \tag{4.2.22}$$

i.e. the problem is stable.

Proof.

$$\|u\|_H^2 \le \frac{1}{\alpha}a(u, u) = \frac{1}{\alpha}F(u) \le \frac{1}{\alpha}\|F\|_{V^*}\|u\|_H, \tag{4.2.23}$$

and the result follows. □

We close with an important remark about the practical importance of the Riesz map.

Remark 4.2.6 (The Riesz map and the solution of linear systems)**.**
This trick of casting the variational problem as an equality in the primal space might seem like an artifice of the proof, but in fact it is essential for the computational solution of the linear system arising from discretisation. Let

$$Kx = b \tag{4.2.24}$$

be the linear system arising from the discretisation. K *is a map* $K : \mathbb{R}^n \to \mathbb{R}^n$*, but they are two* different *copies of* $\mathbb{R}^n$*; just as* A *maps from primal space to dual space, so does* K*. Confusing the two is a recipe for disaster.*

For small problem sizes (i.e. where the number of degrees of freedom numbers less than a few million), factorisation-based solvers such as Cholesky ($K = LL^T$) *and Gaussian elimination* ($K = LU$) *are feasible; but for larger and larger problem sizes, the only feasible algorithms are those that use only the* action *of the linear system* $x \mapsto Kx$*. The Riesz map is essential as any iterative algorithm will be something like*

$$x^{k+1} = x^k + \nu_k(Kx^k) \tag{4.2.25}$$

The Riesz map is essential for constructing a primal update $\Delta x^k = \nu_k(Kx^k)$ *from the dual vector* Kx^k*. This is the heart of the subject of* preconditioning*, a subdiscipline of numerical linear algebra. For more details, see the review article of Kirby*[4]*.*

[4] R. C. Kirby. From functional analysis to iterative methods. *SIAM Review*, 52(2):269–293, 2010

5 More on variational formulations

5.1 The Laplacian with a Dirichlet boundary condition

5.1.1 The one-dimensional homogeneous case

Let $\Omega = (0, 1)$. Consider the two-point boundary value problem

$$-u'' = f, \quad u(0) = 0, u'(1) = g. \tag{5.1.1}$$

The solution can be determined from f via two integrations. First of all, by integrating both sides from t to 1, we can write

$$u'(t) = \int_t^1 f(s)\, \mathrm{d}s + g, \tag{5.1.2}$$

and integrating again from 0 to x yields

$$u(x) = \int_0^x \int_t^1 f(s)\, \mathrm{d}s\, \mathrm{d}t + gx. \tag{5.1.3}$$

This shows that the equation is well-posed.

We will prove that this equation is well-posed using Lax–Milgram. First, let us cast it in variational form. Define the space

$$V = \{v \in H^1(0, 1) : v(0) = 0\}. \tag{5.1.4}$$

This definition makes sense, because we know $H^1(\Omega)$ functions in one dimension are continuous and we can thus evaluate v at the left endpoint. We encode the Dirichlet conditions in the definition of the spaces; such conditions are called *essential* boundary conditions, or strongly imposed boundary conditions. We will encode the Neumann condition in the variational form itself; these are called weakly imposed boundary conditions.

Multiplying the equation by $v \in V$ and integrating, we find

$$\int_0^1 -u''v\, \mathrm{d}x = \int_0^1 fv\, \mathrm{d}x. \tag{5.1.5}$$

We next integrate by parts:

$$\int_0^1 u'v'\, \mathrm{d}x - u'(1)v(1) + u'(0)v(0) = \int_0^1 fv\, \mathrm{d}x. \tag{5.1.6}$$

The surface integral term on the left disappears as $v(0) = 0$. On the right, we know that $u'(1) = g$, and so we have

$$\int_0^1 u'v' \, \mathrm{d}x = \int_0^1 fv \, \mathrm{d}x + gv(1). \tag{5.1.7}$$

Thus, we have the variational problem (Q) with

$$a(u,v) = \int_0^1 u'v' \, \mathrm{d}x, \tag{5.1.8}$$

and

$$F(v) = \int_0^1 fv \, \mathrm{d}x + gv(1). \tag{5.1.9}$$

Theorem 5.1.1 (Coercivity of this bilinear form)**.** *The bilinear form defined in* (5.1.8) *is coercive on V.*

Proof. Recall that the norm on $H^1(0,1)$ is

$$\|v\|^2_{H^1(0,1)} = \|v\|^2_{L^2(0,1)} + \|v'\|^2_{L^2(0,1)}. \tag{5.1.10}$$

We wish to show that there exists a constant $\alpha > 0$ such that

$$a(v,v) \geq \alpha \|v\|^2_{H^1(0,1)} \quad \text{for all } v \in V. \tag{5.1.11}$$

Expanding definitions, we want to find an α such that

$$a(v,v) = \|v'\|^2_{L^2(0,1)} \geq \alpha \left(\|v\|^2_{L^2(0,1)} + \|v'\|^2_{L^2(0,1)} \right). \tag{5.1.12}$$

If we can prove that there exists an α' such that

$$\|v'\|^2_{L^2(0,1)} \geq \alpha' \|v\|^2_{L^2(0,1)} \tag{5.1.13}$$

then we are done with $\alpha = \frac{\alpha'}{\alpha'+1}$.

Let us write

$$v(t) = \int_0^t v'(x) \, \mathrm{d}x = \int_0^1 v'(x) w_t'(x) \, \mathrm{d}x = a(v, w_t), \tag{5.1.14}$$

where the function $w_t \in V$ is defined by

$$w_t(x) = \begin{cases} x & 0 \leq x \leq t, \\ t & x > t. \end{cases} \tag{5.1.15}$$

This function is not strongly differentiable, but has weak derivative

$$w_t'(x) = \begin{cases} 1 & 0 \leq x \leq t, \\ 0 & x > t, \end{cases} \tag{5.1.16}$$

ensuring the correctness of (5.1.14). The function $w_t(x)$ is the a-Riesz representation of the functional $j : v \mapsto v(t)$.

We can invoke Cauchy–Schwarz on $L^2(0,1)$ to get

$$|v(t)| = |a(v, w_t)| \leq \|v'\|_{L^2(0,1)} \|w_t'\|_{L^2(0,1)} = \sqrt{t}\|v'\|_{L^2(0,1)}. \tag{5.1.17}$$

Thus,

$$\|v\|^2_{L^2(0,1)} = \int_0^1 v^2(x)\,\mathrm{d}x \leq \int_0^1 x\|v'\|^2_{L^2(0,1)}\,\mathrm{d}x = \frac{1}{2}\|v'\|^2_{L^2(0,1)} \tag{5.1.18}$$

so in this case we can take $\alpha' = 2$ and thus $\alpha = \frac{2}{3}$. □

Note that if we consider a over the whole of $H^1(0,1)$, it is *not* coercive: $v(x) \equiv 1$ is in $H^1(0,1)$ with $a(v,v) = 0$ but $\|v\| > 0$. The boundary condition $v(0) = 0$ is essential to the well-posedness of the equation.

Note also that the coercivity constant will depend on the length of the domain: for an interval of length L, $\alpha' = \frac{2}{L^2}$ and $\alpha = \frac{2}{L^2+2}$.

Theorem 5.1.2 (Continuity of this bilinear form)**.** *The bilinear form defined in* (5.1.8) *is continuous on* $H^1(0,1)$.

Proof.

$$\begin{aligned}
|a(u,v)| = |(u',v')_{L^2(0,1)}| &\leq \|u'\|_{L^2(0,1)}\|v'\|_{L^2(0,1)} \\
&\leq \left(\|u\|^2_{L^2(0,1)} + \|u'\|^2_{L^2(0,1)}\right)^{\frac{1}{2}} \left(\|v\|^2_{L^2(0,1)} + \|v'\|^2_{L^2(0,1)}\right)^{\frac{1}{2}} \\
&= \|u\|_{H^1(0,1)}\|v\|_{H^1(0,1)}.
\end{aligned}$$

That is, the bilinear form is continuous with $C = 1$. □

It now remains to prove continuity of the right-hand side F.

Theorem 5.1.3 (Continuity of the linear form)**.** *The linear form defined in* (5.1.9) *is bounded.*

Proof. The boundedness of the first term with respect to the $L^2(\Omega)$ norm follows from Hölder's inequality, as $f \in L^2(0,1)$ and $v \in L^2(0,1)$[1]. The boundedness of the second term follows from the continuity of H^1-functions in one dimension. □

[1] Fleshing out the proof for the boundedness of the first term will require the concept of the equivalence of norms and the Poincaré–Friedrichs inequality, both of which will be discussed later.

Thus, by theorem 4.1.2 or 4.2.1, the variational problem is well-posed.

5.1.2 *The higher-dimensional homogeneous case*

Now let us break up Γ into two disjoint components Γ_D and Γ_N, where both Γ_D and Γ_N have nonzero measure, and consider the higher-dimensional problem

$$\begin{aligned}
-\nabla^2 u &= f \text{ in } \Omega, \\
u &= 0 \text{ on } \Gamma_D, && (5.1.19)\\
\nabla u \cdot n &= g \text{ on } \Gamma_N. && (5.1.20)
\end{aligned}$$

Define the space

$$V = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}. \tag{5.1.21}$$

Multiplying by $v \in V$, integrating and integrating by parts, we get

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x + \int_{\Gamma_N} g v \, \mathrm{d}s. \tag{5.1.22}$$

Let us consider the conditions of Lax–Milgram.

Linearity and boundedness of the right-hand side is not straightforward, but true[2]. The earlier proof of theorem 5.1.2 that the bilinear form

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x \tag{5.1.23}$$

is continuous is true in general for any dimension. However, our coercivity proof was specific to one-dimensional problems. The result that gives coercivity in higher dimensions is known as the Poincaré–Friedrichs inequality; see Brenner and Scott, proposition 5.3.4.

[2] It follows from the boundedness of the *Dirichlet trace operator*, an object that evaluates functions on the boundary. We have not done the functional analysis required in this course to understand this.

Theorem 5.1.4 (Poincaré–Friedrichs inequality)**.** *Let Ω be a bounded Lipschitz domain, and suppose $\Gamma_D \subset \partial\Omega$ is closed and has nonzero measure. Let*

$$V = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}. \tag{5.1.24}$$

Then there is a constant $K < \infty$ depending only on Ω and Γ_D such that

$$\int_\Omega u^2 \, \mathrm{d}x \le K \int_\Omega |\nabla u|^2 \, \mathrm{d}x \tag{5.1.25}$$

for all $u \in V$. The constant $K(\Omega, \Gamma_D)$ is called the Poincaré constant for the domain and boundary.

With this coercivity result, we can invoke Lax–Milgram and assert the well-posedness of the variational problem.

5.1.3 *The inhomogeneous case*

Now consider

$$\begin{aligned} -\nabla^2 u &= f \text{ in } \Omega \\ u &= h \text{ on } \Gamma_D \\ \nabla u \cdot n &= g \text{ on } \Gamma_N. \end{aligned} \tag{5.1.26}$$

i.e. we do not assume that the Dirichlet condition is homogeneous. For simplicity, suppose h is defined on all of Ω such that $h \in H^1(\Omega)$[3]. Define

$$\hat{u} = u - h. \tag{5.1.27}$$

[3] Actually, we require that h lives in a function space called $H^{1/2}(\Gamma_D)$; when you evaluate a H^1 function on the boundary you "lose half an order of regularity". Such a function can always be extended to a function in $H^1(\Omega)$ by solving an elliptic PDE, so this supposition is justified.

Then $\hat{u}$ satisfies

$$\begin{aligned}
-\nabla^2 \hat{u} &= f + \nabla^2 h \quad \text{in } \Omega \\
\hat{u} &= 0 \quad \text{on } \Gamma_D \\
\nabla \hat{u} \cdot n &= g - \nabla h \cdot n \text{ on } \Gamma_N,
\end{aligned} \tag{5.1.28}$$

where $\nabla^2 h$ is to be understood weakly, as described in example 4.2.3. That is, we can solve for $\hat{u}$ as in the previous section, and then set $u = \hat{u} + h$ as a post-processing step. Alternatively, stated in variational form, if u satisfies $a(u,v) = F(v)$ for all $v \in V$, then $\hat{u}$ satisfies

$$a(\hat{u}, v) = a(u - h, v) = a(u, v) - a(h, v) = F(v) - a(h, v) \tag{5.1.29}$$

for all $v \in V$.

5.2 Pure Neumann boundary conditions

Consider the pure Neumann problem

$$\begin{aligned}
-\nabla^2 u &= f \text{ in } \Omega \\
\nabla u \cdot n &= g \text{ on } \Gamma.
\end{aligned} \tag{5.2.1}$$

Clearly, this problem does not have a unique solution, for if u satisfies the equations then so does $u + c$ for any constant c.

In order for this to have any solution, we have to impose additional restrictions on the data. Observe that

$$\int_\Omega f \, \mathrm{d}x = \int_\Omega -\nabla^2 u \, \mathrm{d}x = \int_\Omega \nabla u \cdot \nabla 1 \, \mathrm{d}x - \int_\Gamma \nabla u \cdot n \, \mathrm{d}s = -\int_\Gamma g \, \mathrm{d}s$$

and thus f and g have to satisfy a *compatibility condition*. This is typical of linear problems: you either have a unique solution, no solutions (when the compatibility condition is not satisfied, i.e. when the right-hand side of our problem is not in the range of the operator), or an infinite number of solutions (when it is).

If we pose this in variational form with $v \in V = H^1(\Omega)$, we get

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x - \int_\Gamma g v \, \mathrm{d}s, \tag{5.2.2}$$

and we cannot apply our theorems because the bilinear form is not coercive on $H^1(\Omega)$. However, if we restrict ourselves to the function space

$$V = \{v \in H^1(\Omega) : \int_\Omega v \, \mathrm{d}x = 0\} \tag{5.2.3}$$

then the problem is well-posed under the compatibility condition (5.2). (The proof relies on a similar Poincaré inequality, the Poincaré–Neumann inequality.)

5.3 A nonsymmetric problem

Of course, all of the problems we have considered heretofore are symmetric, and so strictly speaking we haven't needed Lax–Milgram. Assume $f \in L^2(0,1)$, and consider a problem which is coercive and continuous, but not symmetric:

$$-u'' + u' + u = f, \quad u'(0) = 0 = u'(1). \tag{5.3.1}$$

As we have no Dirichlet boundary conditions, we will use the space $V = H^1(0,1)$. Testing against $v \in V$ and integrating by parts, we find

$$\int_0^1 u'v' \,\mathrm{d}x + \int_0^1 u'v \,\mathrm{d}x + \int_0^1 uv \,\mathrm{d}x = \int_0^1 fv \,\mathrm{d}x. \tag{5.3.2}$$

Thus, our standard variational problem has

$$a(u,v) = \int_0^1 u'v' + u'v + uv \,\mathrm{d}x \tag{5.3.3}$$

and

$$F(v) = \int_0^1 fv \,\mathrm{d}x. \tag{5.3.4}$$

Boundedness of F follows from Hölder's inequality as $f \in L^2(0,1)$. To invoke Lax–Milgram, we need to prove continuity and coercivity of a. To prove continuity, observe that

$$\begin{aligned}
|a(u,v)| &\le |(u,v)_{H^1(0,1)}| + \left|\int_0^1 u'v \,\mathrm{d}x\right| && (5.3.5)\\
&\le \|u\|_{H^1(0,1)}\|v\|_{H^1(0,1)} + \|u'\|_{L^2(0,1)}\|v\|_{L^2(0,1)} && (5.3.6)\\
&\le 2\|u\|_{H^1(0,1)}\|v\|_{H^1(0,1)} && (5.3.7)
\end{aligned}$$

so we can take our continuity constant $C = 2$. To prove coercivity, observe that

$$\begin{aligned}
a(v,v) &= \int_0^1 v'^2 + v'v + v^2 \,\mathrm{d}x && (5.3.8)\\
&= \frac{1}{2}\int_0^1 (v'^2 + v^2) \,\mathrm{d}x + \frac{1}{2}\int_0^1 (v' + v)^2 \,\mathrm{d}x && (5.3.9)\\
&\ge \frac{1}{2}\|v\|^2_{H^1(0,1)}. && (5.3.10)
\end{aligned}$$

We can thus invoke the Lax–Milgram Theorem.

6 Differentiation in Banach spaces and energy

There is a fundamental connection between symmetric coercive problems and optimisation of convex energy functionals in function spaces. In this setting, the method of Galerkin projection possesses a beautiful optimality property, which accounts for a large part of the popularity of finite element methods.

For more details on differentiation in Banach spaces, see Hinze et al.[1]

[1] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009

6.1 Differentiation between Banach spaces

In this section we discuss how to take directional derivatives of (possibly nonlinear) functionals defined on Banach spaces.

Let J be a nonlinear functional $J : V \to \mathbb{R}$. As a concrete example, consider the nonlinear functional $J : H_0^1(\Omega) \to \mathbb{R}$ defined by

$$J(u) = \frac{1}{2} \int_\Omega |\nabla u|^2 \, \mathrm{d}x. \tag{6.1.1}$$

How will this functional value change if we make a small perturbation v to the input argument?

Definition 6.1.1 (Directional derivative)**.** *Let $J : V \to W$, where V and W are Banach spaces. The directional derivative of J evaluated at $u \in V$ in the direction $v \in V$ is*

$$J'(u; v) = \lim_{\varepsilon \to 0^+} \frac{J(u + \varepsilon v) - J(u)}{\varepsilon}, \tag{6.1.2}$$

if the limit exists.

Definition 6.1.2 (Directionally differentiable)**.** *If the directional derivative of J at u in the direction v exists for all v, then J is directionally differentiable at u.*

Definition 6.1.3 (Gâteaux differentiable)**.** *If J is directionally differentiable at u, and for fixed u the map $J'(u) : V \to W$ is linear and bounded, then J is Gâteaux differentiable at u with derivative $J'(u)$.*

Note that we define derivatives of maps between arbitrary Banach spaces, but in practice we apply it to functionals on a Hilbert space, i.e. functions mapping to $\mathbb{R}$.

The strongest definition of differentiability we will require is *Fréchet* differentiability. This asserts that the derivative is a *good approximation* to J near a point u.

Definition 6.1.4 (Fréchet differentiable)**.** *Suppose J is Gâteaux differentiable at a point $u \in V$ and that the derivative J' satisfies*

$$\lim_{v \to 0} \frac{\|J(u+v) - J(u) - J'(u)v\|}{\|v\|} = 0 \quad \textit{for all } v \in V. \tag{6.1.3}$$

Then J is Fréchet differentiable at u.

If it exists, the Fréchet derivative is unique.

Do not concern yourself too much with the subtle distinctions between senses of differentiability; we don't need these fine shades of grey. If anything arises on this course that is not Fréchet differentiable, it will be surrounded by flashing lights and hazard signs.

Let us work this derivative out for the example (6.1.1):

$$\begin{aligned} J'(u;v) &= \lim_{\varepsilon \to 0^+} \frac{1}{2\varepsilon} \int_\Omega |\nabla u + \varepsilon \nabla v|^2 - |\nabla u|^2 \, \mathrm{d}x && (6.1.4) \\ &= \lim_{\varepsilon \to 0^+} \frac{1}{2\varepsilon} \int_\Omega 2\varepsilon \nabla u \cdot \nabla v + \varepsilon^2 |\nabla v|^2 \, \mathrm{d}x && (6.1.5) \\ &= \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x, && (6.1.6) \end{aligned}$$

exactly the bilinear form (1.2.9). It is no accident that the bilinear form in our example PDE is the derivative of something else. In fact, we shall see later that our running Laplacian example (Q) encodes the optimality conditions for the minimisation of an associated energy functional.

Before moving on, let us look at one reason why the derivative is useful.

Theorem 6.1.5 (Analogue of the fundamental theorem of calculus)**.** *Let $u, v \in V$. Suppose $J : V \to W$ is Fréchet differentiable on the line segment $\{u + tv : t \in [0,1]\}$. Then the following holds:*

$$J(u+v) - J(u) = \int_0^1 J'(u+tv; v) \, \mathrm{d}t. \tag{6.1.7}$$

There is also an analogue of the chain rule. For details, see Hinze et al.

6.2 Symmetric coercive problems and energy

Let us take our favourite homogeneous Dirichlet Laplacian case once more:

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = F(v) \text{ for all } v \in H_0^1(\Omega), \tag{6.2.1}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x \tag{6.2.2}$$

and

$$F(v) = \int_\Omega fv \, \mathrm{d}x. \tag{6.2.3}$$

It turns out that this problem can be recast as an energy minimisation problem. Consider the problem

$$u = \operatorname*{argmin}_{v \in H_0^1(\Omega)} J(v) = \frac{1}{2}\int_\Omega \nabla v \cdot \nabla v \, \mathrm{d}x - \int_\Omega fv \, \mathrm{d}x. \tag{6.2.4}$$

Note that $J(v) = \frac{1}{2}a(v,v) - F(v)$.

Theorem 6.2.1 (Energy minimisation)**.** *Let u be the (unique) solution to (6.2.1) in $H_0^1(\Omega)$. Then u is the unique minimiser of J over $H_0^1(\Omega)$.*

Proof. Let u be the unique solution to (6.2.1). Let $v \in H_0^1(\Omega)$. Then

$$\begin{aligned}
J(v) - J(u) &= \frac{1}{2}a(v,v) - F(v) - \frac{1}{2}a(u,u) + F(u) && (6.2.5)\\
&= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - F(v-u) && (6.2.6)\\
&= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - a(u,v-u) && (6.2.7)\\
&= \frac{1}{2}\left(a(v,v) - 2a(u,v) + a(u,u)\right) && (6.2.8)\\
&= \frac{1}{2}a(v-u,v-u). && (6.2.9)
\end{aligned}$$

Because a is coercive,

$$J(v) - J(u) \geq \frac{\alpha}{2}\|v-u\|^2_{H^1(\Omega)} \geq 0 \text{ for all } v \in H_0^1(\Omega). \tag{6.2.10}$$

Thus,

$$J(v) \geq J(u) \text{ for all } v \in H_0^1(\Omega), \tag{6.2.11}$$

i.e. u minimises J over $H_0^1(\Omega)$.

The minimiser u is unique, because if $\tilde{u}$ also minimises J, then

$$J(\tilde{u}) - J(u) = 0 \geq \frac{\alpha}{2}\|\tilde{u}-u\|^2_{H^1(\Omega)} \geq 0 \tag{6.2.12}$$

and hence $\tilde{u} = u$. □

Now suppose u is the unique minimiser of J. Then $J(u) \leq J(u + \varepsilon v)$ for all $v \in H_0^1(\Omega)$ and $\varepsilon > 0$. This implies

$$\frac{J(u+\varepsilon v) - J(u)}{\varepsilon} = a(u,v) + \frac{1}{2}\varepsilon a(v,v) - F(v) \geq 0, \tag{6.2.13}$$

and passing to the limit $\varepsilon \to 0^+$ yields that $a(u,v) \geq F(v)$ for all $v \in H_0^1(\Omega)$. Now replace v by $-v$, which is also in $H_0^1(\Omega)$, to find

$$\frac{J(u-\varepsilon v) - J(u)}{\varepsilon} = -a(u,v) + \frac{1}{2}\varepsilon a(v,v) + F(v) \geq 0, \tag{6.2.14}$$

to conclude that $a(u,v) \leq F(v)$ for all $v \in H_0^1(\Omega)$.

This proves the following result.

Theorem 6.2.2. *Let $u \in H_0^1(\Omega)$ be the unique minimiser of $J(\cdot)$ over $H_0^1(\Omega)$. Then u is the unique solution of* (6.2.1). *The variational problem* (6.2.1) *is called the* Euler–Lagrange *equation for this minimisation problem.*

This theorem is the converse of the previous result, and the two results together express the equivalence of the weak formulation and the minimisation problem.

Another way to see this is the following. If u minimises J, then J has a stationary point at u[2]. That is, $J'(u) = 0$, or

$$J'(u;v) = 0 \text{ for all } v \in H_0^1(\Omega). \tag{6.2.15}$$

[2] The converse is not true! A nonconvex functional $J : V \to \mathbb{R}$ might have a stationary point that is a saddle point.

Calculating $J'(u;v)$, we find

$$\begin{aligned}
J'(u;v) &= \lim_{\varepsilon \to 0} \frac{J(u+\varepsilon v) - J(u)}{\varepsilon} && (6.2.16)\\
&= \lim_{\varepsilon \to 0} \frac{\frac{1}{2}a(u+\varepsilon v, u+\varepsilon v) - F(u+\varepsilon v) - \frac{1}{2}a(u,u) + F(u)}{\varepsilon} && (6.2.17)\\
&= \lim_{\varepsilon \to 0} a(u,v) - F(v) + \frac{\varepsilon}{2}a(v,v) && (6.2.18)\\
&= a(u,v) - F(v) && (6.2.19)
\end{aligned}$$

and hence

$$a(u,v) - F(v) = 0 \text{ for all } v \in H_0^1(\Omega). \tag{6.2.20}$$

6.3 *Galerkin approximation and energy minimisation*

Consider the Galerkin approximation to (6.2.1) over a subspace $V_h \subset H_0^1(\Omega)$. For exactly the same reason (replace $H_0^1(\Omega)$ with V_h throughout the proof), the Galerkin approximation satisfies the following crucial property.

Theorem 6.3.1. *Let u_h be the Galerkin approximation of a symmetric linear continuous coercive variational problem in V_h. Then*

$$u_h = \underset{v_h \in V_h}{argmin} J(v_h). \tag{6.3.1}$$

That is, we have

$$J(u) \leq J(u_h) \leq J(v_h) \text{ for all } v_h \in V_h. \tag{6.3.2}$$

In other words, *the Galerkin projection gives you the minimal energy solution in your trial space*. This optimality property is one of the reasons why Galerkin approximation is fundamentally a good idea. It also explains why the finite element method is especially popular in engineering; the equations of linear elasticity that describe small deformations to structures are symmetric and coercive, and so the finite element method yields an optimal approximation on any mesh.

6.4 The Helmholtz equation

As a final example, let $V = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}$, and consider the energy $J : V \to \mathbb{R}$

$$J(u) = \frac{1}{2}\int_\Omega \nabla u \cdot \nabla u \, \mathrm{d}x + \frac{1}{2}\int_\Omega u^2 \, \mathrm{d}x - \int_\Omega f u \, \mathrm{d}x - \int_{\Gamma_N} g u \, \mathrm{d}s, \tag{6.4.1}$$

where $f \in L^2(\Omega)$, Γ_N is a part (possibly all) of the boundary with nonzero measure, $\Gamma_D = \partial\Omega \setminus \Gamma_N$, and $g \in H^{-1/2}(\Gamma_N) = \left(H^{1/2}(\Gamma_N)\right)^*$. We seek stationary points of this energy, so set its derivative to zero:

$$\begin{aligned} J'(u; v) &= \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \int_\Omega uv \, \mathrm{d}x - \int_\Omega f v \, \mathrm{d}x - \int_{\Gamma_N} g v \, \mathrm{d}s && (6.4.2) \\ &= 0 \text{ for all } v \in V. && (6.4.3) \end{aligned}$$

We recognise this as the weak form for the equation

$$\begin{aligned} -\nabla^2 u + u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma_D, \\ \nabla u \cdot n &= g \text{ on } \Gamma_N. \end{aligned} \tag{6.4.4}$$

This is the Helmholtz equation, or more specifically the "good" Helmholtz equation. Let us consider the conditions of Lax–Milgram.

First, consider boundedness of

$$F(v) = \int_\Omega f v \, \mathrm{d}x + \int_{\Gamma_N} g v ds. \tag{6.4.5}$$

As $f \in L^2(\Omega)$, the first term is bounded. Similarly, if $v \in H^1(\Omega)$, then $v|_{\Gamma_N} \in H^{1/2}(\Gamma_N)$, and so $g \in H^{-1/2}(\Gamma_N)$ ensures that the second term is bounded[3]. As F is obviously linear, the right-hand side is in the dual space of $H^1(\Omega)$.

Observe that

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \int_\Omega uv \, \mathrm{d}x \tag{6.4.6}$$

is the $H^1(\Omega)$ inner product, and thus continuity follows immediately from Cauchy–Schwarz with $C = 1$. Similarly, as

$$a(v,v) = (v,v)_{H^1(\Omega)} = \|v\|^2_{H^1(\Omega)}, \tag{6.4.7}$$

the problem is coercive with $\alpha = 1$. (These properties are generally true for bilinear forms that happen to be inner products.)

For our final word, let us consider the slightly modified problem

$$\begin{aligned} -\nabla^2 u - k^2 u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma_D = \Gamma \setminus \Gamma_N, \\ \nabla u \cdot n &= g \text{ on } \Gamma_N. \end{aligned} \tag{6.4.8}$$

This is known as the "bad" Helmholtz equation (in contrast to the "good" Helmholtz above), as the problem is no longer coercive for sufficiently large k. In fact, the development of good discretisations and fast solvers for this equation is an outstanding open problem in numerical analysis of major importance in wave propagation, such as in seismology, acoustics, and electromagnetism.

[3] This relies on the boundedness of the Dirichlet trace operator, something we won't discuss.

6.5 A remark on optimisation in Hilbert spaces

Suppose we wish to minimise a functional $J : \mathbb{R}^n \to \mathbb{R}$. One of the most fundamental algorithms for this task is *steepest descent*, which consists of setting

$$u^{k+1} = u^k - \alpha_k \nabla J(u^k), \tag{6.5.1}$$

where u^k is the previous iterate, $\nabla J(u^k)$ is the gradient of the functional, and α_k is typically chosen by line search. In this finite dimensional setting, $u^k \in \mathbb{R}^n$, and $\nabla J(u^k) \in \mathbb{R}^n$, and the iteration is well-defined[4].

Now consider the case where $J : V \to \mathbb{R}$, with V an infinite-dimensional Hilbert space. What is the analogue of ∇J that will give us a descent direction for the functional?

The answer involves the Fréchet derivative J', but it is not quite the complete story. Consider the following iteration:

$$u^{k+1} = u^k - \alpha_k J'(u^k). \tag{6.5.2}$$

This iteration *is not well-defined*: on the right-hand side, we are instructed to add a quantity in V to a quantity in V^*. These are different spaces, and must be carefully distinguished.

[4] There are much faster algorithms, of course, such as Newton's method or a quasi-Newton method.

The key ingredient is the Riesz map $R : V^* \to V$. The correct definition of steepest descent in infinite dimensions is the iteration

$$u^{k+1} = u^k - \alpha_k R J'(u^k). \tag{6.5.3}$$

This is why the definition of the gradient in $\mathbb{R}^n$ involves the transpose: transposition is the Riesz map on $\mathbb{R}^n$.

We *cannot* just discretise the problem to $\mathbb{R}^n$ and apply our standard Euclidean algorithms; we would be "exploring the space with the wrong metric", and the computational performance would be disastrous. In particular, the number of iterations required for convergence of the algorithm would blow up as the mesh size $h \to 0$[5]. To achieve mesh independent convergence, we must formulate the optimisation algorithm *in function spaces*, and defer discretisation to the very last possible moment.

[5] T. Schwedes, D. A. Ham, S. W. Funke, and M. D. Piggott. *Mesh Dependence in PDE-Constrained Optimisation*. Springer International Publishing, 2017

7 Galerkin approximation

Given a linear variational problem

$$\text{find } u \in V \text{ such that } a(u, v) = F(v) \text{ for all } v \in V, \tag{7.0.1}$$

we form its Galerkin approximation over a subspace $V_h \subset V$

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = F(v_h) \text{ for all } v_h \in V_h. \tag{7.0.2}$$

We first consider its approximation properties over *arbitrary* subspaces V_h, i.e. not specific to finite elements. We will then specialise these generic error estimates to function spaces constructed via finite elements.

7.1 Elementary properties of the approximation

Corollary 7.1.1. *Let a and F satisfy the hypothesis of the Lax–Milgram Theorem. Then the Galerkin approximation is well-posed for any closed subspace $V_h \subset V$.*

Proof. This follows directly from Lax–Milgram. As $V_h \subset V$, $a : V_h \times V_h \to \mathbb{R}$ is bounded and coercive on V_h, and $F : V_h \to \mathbb{R}$ is linear and bounded. Thus, by Lax–Milgram, the variational problem defining the Galerkin approximation is well-posed. □

Recall that when a basis $\{\phi_i\}$ is chosen, the Galerkin approximation can be written as

$$Ax = b, \tag{7.1.1}$$

where x_i are the basis function coefficients for the solution u_h,

$$b_i = F(\phi_i) \tag{7.1.2}$$

and

$$A_{ji} = a(\phi_i, \phi_j). \tag{7.1.3}$$

One useful fact about the finite element method is that *the linear system inherits the structural properties of the variational problem.* For example:

Lemma 7.1.2. *If a is a symmetric bilinear form, A is a symmetric matrix.*

Proof.

$$A_{ji} = a(\phi_i, \phi_j) = a(\phi_j, \phi_i) = A_{ij}. \tag{7.1.4}$$

□

Another important property is positive-definiteness:

Lemma 7.1.3. *If a is coercive, then A is positive-definite.*

Proof. For positive-definiteness, we require

$$c^T A c > 0 \text{ for } c \neq 0. \tag{7.1.5}$$

Let $v \in V_h$ be the function with basis function coefficients c, i.e.

$$v(x) = \sum_i c_i \phi_i(x). \tag{7.1.6}$$

Then

$$c^T A c = a(v, v) \geq \alpha \|v\|_V^2 > 0 \text{ if } c \neq 0. \tag{7.1.7}$$

□

7.2 Galerkin orthogonality

We know that the solution u satisfies

$$a(u, v) = F(v) \quad \text{for all } v \in V, \tag{7.2.1}$$

and thus in particular

$$a(u, v_h) = F(v_h) \quad \text{for all } v_h \in V_h \subset V. \tag{7.2.2}$$

The Galerkin approximation $u_h \in V_h$ satisfies

$$a(u_h, v_h) = F(v_h) \quad \text{for all } v_h \in V_h \subset V. \tag{7.2.3}$$

Subtracting (7.2.3) from (7.2.2), we find

$$a(u - u_h, v_h) = 0 \quad \text{for all } v_h \in V_h. \tag{7.2.4}$$

In words, *the error $e_h = u - u_h$ is a-orthogonal to the test space V_h*. More informally, the error is "zero when viewed on our mesh". This crisp characterisation of the error is a distinctive feature of Galerkin methods, and is crucial for the error estimates we will derive.

7.3 Asymmetric case: quasi-optimality in V-norm

Let us first consider the general case, where we do not assume a is symmetric. It turns out that the error in our Galerkin approximation is *quasi-optimal*: it is optimal, up to some problem-specific constants.

Suppose a is V-coercive and continuous, but not necessarily symmetric. For any $v_h \in V_h$,

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

Dividing by α and minimising over $v_h \in V$, we obtain the following result.

Lemma 7.3.1 (Céa's Lemma)**.** *The Galerkin approximation $u_h \in V_h$ to $u \in V$ is quasi-optimal, in that it satisfies*

$$\|u-u_h\|_V \leq \frac{C}{\alpha}\min_{v_h\in V_h}\|u-v_h\|_V.$$

The minimum is achieved because V_h is closed by assumption (otherwise we would have an infimum).

This result is very useful because it allows us to intuitively characterise the quality of the approximation. Before seeing how we can use this to bound the approximation error in terms of the mesh size h, let us consider this quasi-optimality result in the context of symmetric problems.

7.4 Symmetric case: optimality in energy norm

For this section, assume a is symmetric.

If a is symmetric, then a defines an inner product on V, and $(V, a(\cdot,\cdot))$ is a Hilbert space, by theorem 4.1.1. When measured with respect to the $\|\cdot\|_a$-norm, the problem is continuous with $C = 1$ (by Cauchy–Schwarz) and is coercive with $\alpha = 1$ (by definition of the norm).

Thus, if we restate Céa's Lemma with respect to the $(V, a(\cdot,\cdot))$ Hilbert space, we have

$$\|u-u_h\|_a \leq \frac{C}{\alpha}\min_{v_h\in V_h}\|u-v_h\|_a \tag{7.4.1}$$

$$= \min_{v_h\in V_h}\|u-v_h\|_a. \tag{7.4.2}$$

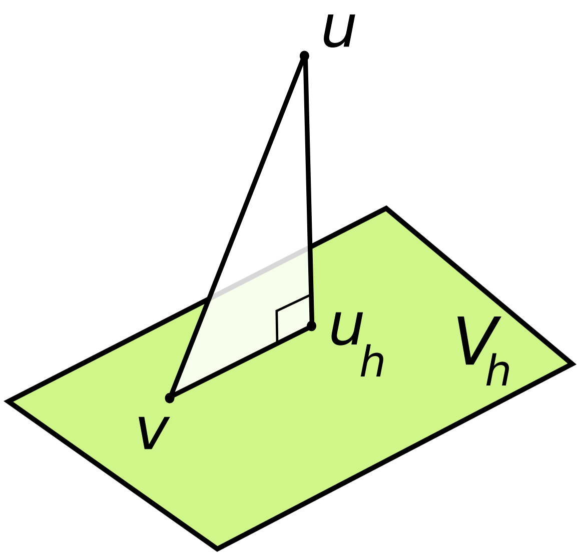

Figure 7.1: For symmetric problems, the Galerkin approximation is the orthogonal projection of the solution onto the trial space when measured in the energy norm. Credit: Oleg Alexandrov, Wikipedia.

Since $u_h \in V_h$, we have to have equality, and thus *the error is optimal in the norm induced by the problem*:

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a. \tag{7.4.3}$$

This is closely related to the energy minimisation property discussed in the previous lecture.

7.5 *Quasioptimality, interpolation, and regularity*

Céa's Lemma allows us to *bound the error in terms of an interpolation operator mapping to the space V_h*. That is, if we can find *any* specific $v_h \in V_h$ so that $\|u - v_h\|$ is bounded, then this in turn bounds the approximation error. Ideally, this bound would be given in terms of quantities under our control, such as the mesh size and approximation order.

Such bounds are given by interpolation estimates. We construct an interpolation operator $\mathcal{I}_h$ mapping from some function space to V_h. A major objective of this course will be to derive the following theorem, which quantifies the error introduced by interpolation, over the next few weeks.

Theorem 7.5.1. *Let Ω be a polygonal (in two dimensions) or polyhedral (in three dimensions) domain. Suppose we have a sequence of meshes $\{\mathcal{M}_h\}$ indexed by mesh size h with $h \to 0$. Let V_h be the discrete function space constructed by equipping mesh $\mathcal{M}_h$ with Lagrange finite elements of degree p. Let $u \in H^{p+1}(\Omega)$. Let $\mathcal{I}_h u$ denote the interpolant of u onto V_h. Then,* *under certain regularity conditions on the mesh, there exists a constant c such that*

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq ch^p \sum_{|\alpha|=p+1} \|D^\alpha u\|_{L^2(\Omega)}. \tag{IE}$$

Note that the interpolation operator bound requires $u \in H^2(\Omega)$ at least; this is because Lagrange elements require the evaluation of the solution, and so we must have $u \in C(\Omega)$. Furthermore, the smoother the solution, the faster the convergence: if $u \in H^s(\Omega)$, then our bound is $O(h^{s-1})$.

Consider our standard model problem (Q). To derive a bound on the finite element approximation error, we must employ three ingredients. The first is a quasi-optimality result, such as Céa's Lemma, which bounds the finite element error in terms of the best approximation in V_h:

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}. \tag{7.5.1}$$

The second is an interpolation error result, which places an upper bound on the error of the best approximation:

$$\min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq \|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq O(h^{s-1}), \tag{7.5.2}$$

so long as $u \in H^2(\Omega)$ or better. The third is an elliptic regularity result, which guarantees that $u \in H^s(\Omega)$ for some $s \geq 2$[1]. Combining these three results, we have

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{C}{\alpha} \cdot O(h^{s-1}). \tag{7.5.3}$$

This bound is called an *a priori* bound, in that it can be derived before one actually computes u_h[2]

[1] For example, this is true if Ω is Lipschitz and convex, or if $\partial\Omega$ is smooth.

[2] This contrasts with *a posteriori* error bounds which make explicit use of the computed approximation u_h to provide computable bounds on the global error.

7.6 *A success: linear elasticity*

One of the most important early applications of the finite element method was to the equations of linear elasticity, which describe how a solid object deforms and becomes internally stressed due to loading conditions. No bridge is built or rocket launched without a finite element approximation of linear elasticity somewhere along the way. In fact, the finite element method was largely invented by engineers seeking to solve linear elasticity; its mathematical foundations came much later. This discussion is drawn from Ciarlet[3].

[3] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1978. Reprinted by SIAM in 2002

Let $\Omega \subset \mathbb{R}^3$ be an open bounded Lipschitz domain; its closure $\bar{\Omega}$ is referred to as the reference configuration. We seek to characterise its shape upon loading via a mapping $\phi : \bar{\Omega} \to \mathbb{R}^3$ via

$$\widetilde{\Omega} = \phi(\bar{\Omega}). \tag{7.6.1}$$

It is useful to write the deformation ϕ as the sum of the identity map, plus a displacement:

$$\phi(x) = x + u(x), \tag{7.6.2}$$

where $u(x) : \bar{\Omega} \to \mathbb{R}^3$.

The equations that describe the static configuration of Ω under a body force $f : \Omega \to \mathbb{R}^3$ are the following. First, we have the equation of motion, a form of Newton's second law,

$$\nabla \cdot \sigma + f = \rho \ddot{u} = 0, \tag{7.6.3}$$

where $\sigma : \Omega \to \mathbb{R}^{3\times 3}_{\text{sym}}$ is the stress tensor, f is the body force, ρ is the density and $\ddot{u}$ refers to the second derivative of the displacement in time. Since we only seek the static configuration, we set the time derivatives to zero. For an isotropic material[4] the material is described by two parameters $\mu > 0$ (the shear modulus) and $\lambda > 0$ (the first Lamé parameter), and the stress is given by

[4] In general the material can be anisotropic, and is defined by a symmetric fourth-order stiffness tensor, described by 21 components. In turn, each of those components might be a function of space.

$$\sigma = 2\mu\varepsilon(u) + \lambda \text{tr}\varepsilon(u) I, \tag{7.6.4}$$

where $\varepsilon : \Omega \to \mathbb{R}^{3\times 3}_{\text{sym}}$ denotes the strain. This is a more complicated form of Hooke's law, relating the strain and the stress in a linear way. In turn, the strain is described in terms of the deformation by

$$\begin{aligned} \varepsilon(u) &= \frac{1}{2}\left(\nabla u^T + \nabla u + \nabla u^T \nabla u\right), && (7.6.5)\\ &\approx \frac{1}{2}\left(\nabla u^T + \nabla u\right), && (7.6.6) \end{aligned}$$

where the last quadratic term has been dropped because we assume the displacements to be small[5].

[5] If we do not make this assumption, or the stress-strain relationship is nonlinear, then this gives rise to a nonlinear equation instead. This is necessary for studying large deformations or more complicated materials, such as rubber or the human heart.

We close the system with boundary conditions. Decomposing the boundary $\Gamma = \Gamma_D \cup \Gamma_N$ into two disjoint parts of nonzero measure, we set

$$\begin{aligned} u &= 0 \quad \text{on } \Gamma_D, && (7.6.7)\\ \sigma \cdot n &= g \quad \text{on } \Gamma_N, && (7.6.8) \end{aligned}$$

where n is the outward-facing unit normal, and g are prescribed boundary tractions.

When cast in variational form[6], the problem is the familiar

[6] We won't go through the details of the integration by parts. For details, see Ciarlet, pg. 25.

$$\text{find } u \in V \text{ such that } a(u, v) = F(v) \text{ for all } v \in V, \tag{7.6.9}$$

for certain a, F and V we will now describe. Let

$$\left(H^1(\Omega)\right)^3 = H^1(\Omega; \mathbb{R}^3) = H^1(\Omega) \times H^1(\Omega) \times H^1(\Omega) \tag{7.6.10}$$

be the set of vector-valued functions where each component is in $H^1(\Omega)$. As before, we must encode our Dirichlet conditions in the space, so we choose

$$V = \{v \in H^1(\Omega; \mathbb{R}^3) : v|_{\Gamma_D} = 0\}. \tag{7.6.11}$$

The bilinear form $a : H^1(\Omega; \mathbb{R}^3) \times H^1(\Omega; \mathbb{R}^3) \to \mathbb{R}$ is given by

$$\begin{aligned} a(u, v) &= \int_\Omega \sigma(u) : \varepsilon(v) \, \mathrm{d}x & (7.6.12) \\ &= \int_\Omega \sum_{i,j=1}^{3} \sigma_{ij}(u)\varepsilon_{ij}(v) \, \mathrm{d}x & (7.6.13) \\ &= \int_\Omega \lambda \, \mathrm{div} \, u \, \mathrm{div} \, v + 2\mu\varepsilon(u) : \varepsilon(v) \, \mathrm{d}x. & (7.6.14) \end{aligned}$$

This is a symmetric bilinear form that is coercive on V[7]. The linear form $F : H^1(\Omega; \mathbb{R}^3) \to \mathbb{R}$ is given by

[7] Its coercivity is nontrivial, and is guaranteed by a result known as Korn's inequality. As in the Laplacian case, this result relies on the measure of Γ_D being nonzero. See Ciarlet, pg. 24.

$$F(v) = \int_\Omega f \cdot v \, \mathrm{d}x + \int_{\Gamma_N} g \cdot v \, \mathrm{d}s. \tag{7.6.15}$$

From the theory we have already studied, we know the following facts. First, the equations of linear elasticity are well-posed if Γ_D has nonzero measure, $f \in L^2(\Omega; \mathbb{R}^3)$, and $g \in H^{-1/2}(\Omega; \mathbb{R}^3)$, by the Riesz Representation Theorem. Second, the Galerkin approximation over any $V_h \subset V$ will be optimal when measured in the energy norm induced by the problem. Third, the Galerkin approximation minimises the potential energy, given by the sum of the strain energy and the potential energy of the exterior forces,

$$J(v) = \frac{1}{2} \int_\Omega \lambda(\mathrm{div} \, v)^2 + 2\mu\varepsilon(v) : \varepsilon(v) \, \mathrm{d}x - \int_\Omega f \cdot v \, \mathrm{d}x - \int_{\Gamma_N} g \cdot v \, \mathrm{d}s. \tag{7.6.16}$$

These approximation properties strongly motivate the use of a Galerkin approximation. The finite element method is the favoured means of constructing V_h because of its geometric flexibility, to be described in the next lectures.

7.7 *A warning: advection-dominated problems*

Consider again the error bound (7.5.3)

$$\|u - u_h\|_{H^1(\Omega)} \le \frac{C}{\alpha} \cdot O(h^{s-1}). \tag{7.7.1}$$

This result is theoretically reassuring, but note that in practice it is very difficult to bound the constants involved, and for certain elliptic problems the coercivity and boundedness coefficients might be such that C/α is very large. The approximation does indeed converge as

$h \to 0$, but one might have to take h unaffordably small before any reasonable error is achieved. We now present such an example, from Prof. Süli's notes[8].

[8] E. Süli. Lecture notes on Finite Element Methods for Partial Differential Equations, 2012. http://people.maths.ox.ac.uk/suli/fem.pdf

Consider the following advection-diffusion equation:

$$-\varepsilon\nabla^2 u + \mathbf{b}\cdot\nabla u = f \text{ in } \Omega, \tag{7.7.2}$$

$$u = 0 \text{ on } \Gamma, \tag{7.7.3}$$

where $\varepsilon > 0$, $\mathbf{b} = (b_1, \ldots, b_n)^T$ with $b_i \in W^1_\infty(\Omega)$ for $i = 1, \ldots, n$. For simplicity assume that $\operatorname{div}(\mathbf{b}) \leq 0$ almost everywhere. When advection dominates diffusion, the *Péclet number*

$$\mathrm{Pe} = \frac{\sqrt{\sum_i \|b_i\|^2_{L^\infty(\Omega)}}}{\varepsilon} \tag{7.7.4}$$

is very large, e.g. 10^6–10^8.

The continuity constant for this problem is

$$C = \sqrt{\varepsilon^2 + \sum_i \|b_i\|^2_{L^\infty(\Omega)}} \tag{7.7.5}$$

and the coercivity constant is

$$\alpha = \frac{\varepsilon}{\sqrt{1 + K^2(\Omega, \Gamma)}} \tag{7.7.6}$$

where $K(\Omega, \Gamma)$ is the associated Poincaré constant. Thus, the constant in front of the error bound is

$$\frac{C}{\alpha} = \sqrt{1 + K^2}\sqrt{(1 + \mathrm{Pe}^2)}. \tag{7.7.7}$$

Thus, when $\varepsilon \ll 1$, the constant on the right-hand side of the error bound is extremely large through the presence of the Péclet number[9]. Thus, on coarse meshes, conventional finite element methods can be very badly behaved; the solution typically exhibits large nonphysical oscillations which can only be eliminated by drastically reducing the mesh size h.

[9] In fact, things are even worse, as $B(u)$ will also depend on ε via u.

There are finite element techniques to ameliorate this problem, but they are beyond the scope of the course. For more details on this, see Elman, Silvester & Wathen, chapter 6[10].

[10] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers: with applications in incompressible fluid dynamics*. Oxford University Press, 2014

8 Function spaces constructed via finite elements

We now specialise our previous error analysis to the particular choice of discrete function spaces we will choose. These discrete function spaces are constructed via *finite elements* (as the name suggests!).

The beautiful images of finite elements depicted in this chapter are taken from Logg et al.[1]

[1] A. Logg, K. A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2011

8.1 Finite elements

Definition 8.1.1 (Finite element)**.** *A finite element is a triple $(K, \mathcal{V}, \mathcal{L})$ where*

- *The cell K is a bounded, closed subset of $\mathbb{R}^n$ with nonempty connected interior and piecewise smooth boundary;*
- *The space $\mathcal{V} = \mathcal{V}(K)$ is a finite dimensional function space on K of dimension d;*
- *The set of degrees of freedom $\mathcal{L} = \{\ell_1, \ldots, \ell_d\}$ is a basis for $\mathcal{V}^*$, the dual space of $\mathcal{V}$.*

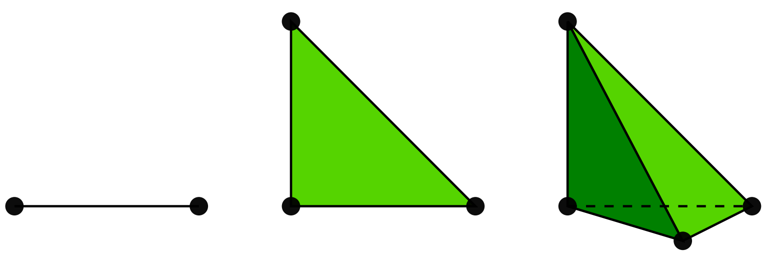

Figure 8.1: The linear Lagrange finite element in one, two and three dimensions. The black circles denote pointwise evaluation.

Example 8.1.2. *Consider the standard linear Lagrange finite element on the triangle, figure 8.1. The cell K is given by the triangle and $\mathcal{V}$ is the space of first degree polynomials on K (a space of dimension $d = 3$). As a basis for*

$\mathcal{V}^*$, we may take point evaluation at the three vertices of K, that is

$$\begin{aligned} \ell_i &: \mathcal{V} \to \mathbb{R} \\ \ell_i(v) &= v(x_i), \end{aligned} \tag{8.1.1}$$

where x_i denotes the coordinates of the i^{th} vertex.

Before we move on, a small piece of notation.

Definition 8.1.3 (Polynomial spaces)**.** *Denote the space of q-degree polynomials on a geometric object $K \subset \mathbb{R}^n$ via $\mathcal{P}_q(K)$:*

$$\mathcal{P}_q(K) = \text{span}\{\, x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}\big|_K : \sum_{i=1}^{n} \alpha_i \le q, \alpha_i \ge 0 \textit{ for all } i = 1, \ldots, n\}. \tag{8.1.2}$$

For example, $\mathcal{P}_3(K)$ on a triangle K is

$$\mathcal{P}_3(K) = \text{span}\{1, x, x^2, x^3, y, y^2, y^3, xy, x^2y, xy^2\} \tag{8.1.3}$$

where each of the functions is restricted to K.

The significance of these objects is the following. We will break up the domain Ω into the non-overlapping union of cells that can be easily mapped to K (e.g. triangles in various configurations and orientations). On each cell, we will approximate the solution with a function in $\mathcal{V}$. The degrees of freedom encode *what we need to store* to uniquely specify a function in $\mathcal{V}$: for example, on the linear Lagrange element, knowing the values $\ell_i(v)$ is sufficient to uniquely determine a specific $v_K \in \mathcal{V}$. In other words, we will solve for the values of $\ell_i(v)$ on each element to determine our overall (global) approximation $v_h \in V_h$.

This can be concretely seen by the following argument. Suppose $\mathcal{V}$ consists of continuous functions (it always does in practice) and for fixed x consider the functional

$$\ell_x(v) = v(x). \tag{8.1.4}$$

Since ℓ_x is a continuous linear functional, and ℓ_i form a basis for V^*, we can therefore express

$$\ell_x = \alpha_1 \ell_1 + \cdots + \alpha_d \ell_d \tag{8.1.5}$$

for some coefficients α. Thus, if we know the values of $\ell_i(v)$, we know the value of the function v at every point $x \in K$, and the function is uniquely specified.

The main work in verifying that something is a finite element is in checking that $\mathcal{L}$ is indeed a basis for $\mathcal{V}^*$. This is simplified by the following lemma.

Lemma 8.1.4 (Verifying finite elements)**.** *Let $\mathcal{V}$ be a d-dimensional vector space and let $\mathcal{L} = \{\ell_1, \ldots, \ell_d\}$ be a subset of the dual space $\mathcal{V}^*$. Then the following two statements are equivalent:*

(a) $\mathcal{L}$ is a basis for $\mathcal{V}^*$;

(b) Given $v \in \mathcal{V}$ with $\ell_i(v) = 0$ for $i = 1, \ldots, d$, then $v \equiv 0$.

This means that we just need to verify condition (b), which is much easier; we set the degrees of freedom to be zero and show that the only element of $\mathcal{V}$ is the zero element.

Proof. Let $\{\phi_1, \ldots, \phi_d\}$ be some basis for $\mathcal{V}$. $\mathcal{L}$ is a basis for $\mathcal{V}^*$ iff given any $\ell \in \mathcal{V}^*$, there exists a set of coefficients $\{\alpha_1, \ldots, \alpha_d\}$ such that

$$\ell = \alpha_1 \ell_1 + \cdots + \alpha_d \ell_d \tag{8.1.6}$$

because $\dim(\mathcal{V}^*) = \dim(\mathcal{V}) = d$. Given an $\ell \in \mathcal{V}^*$, denote

$$y_i = \ell(\phi_i) \stackrel{?}{=} \alpha_1 \ell_1(\phi_i) + \cdots + \alpha_d \ell_d(\phi_i). \tag{8.1.7}$$

If we define the matrix $B_{ij} = \ell_j(\phi_i)$, then (a) is equivalent to saying that the system

$$B\alpha = y \tag{8.1.8}$$

is always solvable, which is the same thing as B being invertible.

Given any $v \in \mathcal{V}$, we can write

$$v = \beta_1 \phi_1 + \cdots + \beta_d \phi_d. \tag{8.1.9}$$

The condition $\ell_i(v) = 0$ means that

$$\beta_1 \ell_i(\phi_1) + \cdots + \beta_d \ell_i(\phi_d) = 0 \tag{8.1.10}$$

and so (b) is equivalent to

$$\beta_1 \ell_i(\phi_1) + \cdots + \beta_d \ell_i(\phi_d) = 0 \; \forall i = 1, \ldots, d \implies \beta_1 = \cdots = \beta_d = 0. \tag{8.1.11}$$

Define the matrix C via $C_{ij} = \ell_i(\phi_j)$. Then (b) is equivalent to $C\beta = 0$ only has trivial solutions, which means C is invertible. But $C = B^T$, so (a) is equivalent to (b). □

Definition 8.1.5. *We say that $\mathcal{L}$* determines *$\mathcal{V}$ if given $v \in \mathcal{V}$, $\ell_i(v) = 0 \; \forall i \implies v = 0$. We also say that $\mathcal{L}$ is* unisolvent.

Example 8.1.6. *For the linear Lagrange triangle, if v is zero at each vertex, then v must be zero everywhere as a plane is uniquely determined by its values at three non-collinear points. Thus, the linear Lagrange element on a triangle is indeed a finite element.*

Having fixed $\mathcal{L}$, the usual choice for a basis of $\mathcal{V}$ is the *nodal basis*.

Definition 8.1.7 (nodal basis)**.** *The basis $(\phi_1, \ldots, \phi_d)$ of $\mathcal{V}$ dual to $\mathcal{L}$, i.e. with the property that*

$$\ell_i(\phi_j) = \delta_{ij} \tag{8.1.12}$$

is called the nodal basis *for $\mathcal{V}$.*

Example 8.1.8 (nodal basis for the linear Lagrange element in one dimension)**.** *Let $K = [0, 1]$, $\mathcal{V}$ be the set of linear functions on K, and $\mathcal{L}$ be pointwise evaluation at the endpoints. Then the nodal basis is given by*

$$\phi_1(x) = 1 - x, \quad \phi_2(x) = x. \tag{8.1.13}$$

Example 8.1.9 (nodal basis for the linear Lagrange element in two dimensions)**.** *Let K be the triangle with vertices at $(0, 0)$, $(0, 1)$, $(1, 0)$. Let $\mathcal{V}$ be the set of linear functions on K, and $\mathcal{L}$ be pointwise evaluation at the vertices. Then the nodal basis is given by*

$$\phi_1(x) = 1 - x_1 - x_2, \quad \phi_2(x) = x_1, \quad \phi_3(x) = x_2. \tag{8.1.14}$$

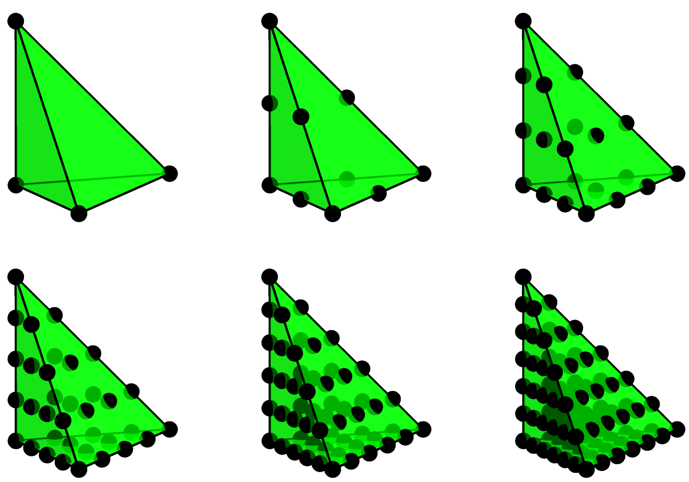

Figure 8.2: The Lagrange CG_q elements for $q = 1, \ldots, 6$.

Of course, there are other finite elements. Let us generalise the Lagrange finite element, before looking at others.

Definition 8.1.10 (Lagrange element)**.** *The Lagrange element of spatial dimension n and degree $q \geq 1$ (sometimes called CG_q) is defined by*

- *K is an n-dimensional simplex (interval, triangle, tetrahedron),*
- *$\mathcal{V} = \mathcal{P}_q(K)$,*
- *$\ell_i : v \mapsto v(x_i)$, $i = 1, \ldots, f(q)$,*

where x_i, $i = 1, \ldots, f(q)$ is an enumeration of points in the element. (See Logg et al.[2] for details of $f(q)$ and the enumeration of points.)

[2] A. Logg, K. A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method.* Springer, 2011

For an illustration of the elements for $q = 1, \ldots 6$ in three dimensions, see figure 8.2. As we enrich the approximation space on each element, we expect that the interpolation error of the true solution to our variational problem will decrease, and thus the error bound of (7.5.3) will improve. We will see later that this will depend on both the interpolation order q and the smoothness of the exact solution.

Before we introduce any other elements, let us discuss how elements are used for interpolation.

8.2 *Local interpolation operator*

One of the main things we will do with finite elements is interpolate functions onto them.

Definition 8.2.1 (Interpolant on an element)**.** *Let $(K, \mathcal{V}, \mathcal{L})$ be a finite element. For a suitable[3] function space H, define the interpolant $\mathcal{I}_K : H \to \mathcal{V}$ via*

[3] Suitable means that that $\ell_i : H \to \mathbb{R}$ is well defined on H for all $\ell_i \in \mathcal{L}$.

$$\begin{aligned} \mathcal{I}_K &: u \mapsto \mathcal{I}_K u \\ \ell_i(\mathcal{I}_K u) &= \ell_i(u) \quad \textit{for all } \ell_i \in \mathcal{L}. \end{aligned} \tag{8.2.1}$$

That is, the interpolant matches the function being interpolated at the degrees of freedom.

In the nodal basis, the interpolation operator is particularly simple. It is a straightforward exercise to verify that

$$\mathcal{I}_K u = \sum_{i=1}^{d} \ell_i(u)\phi_i \tag{8.2.2}$$

satisfies the condition (8.2.1).

8.3 *Meshes and the local-to-global mapping*

To define a global function space

$$V_h = \mathrm{span}\{\phi_1, \ldots, \phi_N\}, \tag{8.3.1}$$

we need to decompose Ω into cells, define a finite element on each, and then specify how the local function spaces are to be stitched together.

Assume for now that Ω is polytopic (polygonal or polyhedral) so that it can be decomposed into simplicial elements exactly. We now define a mesh.

Definition 8.3.1 (mesh)**.** *A* mesh *$\mathcal{M}$ is a geometric decomposition of a domain Ω into a finite set of* cells *$\mathcal{M} = \{K_i\}$ such that*

1. $\cup_i K_i = \overline{\Omega}$.

2. *If $K_i \cap K_j$ for $i \neq j$ is exactly one point, it is a common vertex of K_i and K_j.*

3. *If $K_i \cap K_j$ for $i \neq j$ is not exactly one point, it is a common facet of K_i and K_j (edge in two dimensions, face in three dimensions).*

Meshing is a huge subject in its own right, and often where commercial FEM solvers spend most of their time and effort. For an excellent introduction to the field, see Frey & George[4].

We equip each cell $K \in \mathcal{M}$ with a finite element, so we have a set of finite elements $\{(K, \mathcal{V}_K, \mathcal{L}_K) : K \in \mathcal{M}\}$. Typically we equip every element with the same type of element, but not always[5].

We specify how the elements fit together with the *local-to-global mapping*. For each cell $K \in \mathcal{M}$, the analyst must specify a local-to-global map

$$\iota_K : \{1, \ldots, d(K)\} \to \{1, \ldots, N\} \tag{8.3.2}$$

which specifies how the *local* degrees of freedom $\ell_i^K(v)$ relate to the *global* degrees of freedom. Each local degree of freedom corresponds to a global degree of freedom, under the action of the local-to-global map:

$$\ell_{\iota_K(i)}(v) = \ell_i^K(v|_K), \quad i = 1, \ldots, d(K). \tag{8.3.3}$$

The properties of the local-to-global mapping determine the continuity of the global approximation space. Consider the mesh consisting of two triangular cells depicted in figure 8.3. Both cells are equipped with second order Lagrange elements, so the degrees of freedom consist of pointwise evaluation at the vertices and the midpoints of the edges. The two cells share a common edge: each triangle has three degrees of freedom on that common edge, that match up in a natural way. *If the local-to-global mapping maps the matching local degrees of freedom to the same global degree of freedom, the global approximation will be continuous*; conversely, if the local-to-global mapping maps the matching local degrees of freedom to different global degrees of freedom, the global approximation will permit discontinuous functions. In the mapping depicted in figure 8.3, the continuity of the global approximation across the interface is enforced, and so $V_h \subset H^1(\Omega)$. If continuity is not enforced, the approximation has no weak derivative and so merely $V_h \subset L^2(\Omega)$, as shown in figure 8.4.

Definition 8.3.2 (conforming approximation)**.** *Suppose the continuous variational problem is posed over a Hilbert space V. If $V_h \subset V$, the approximation is* conforming*; if $V_h \not\subset V$, then the approximation is said to be nonconforming.*

[4] P.J. Frey and P.L. George. *Mesh Generation. Application to finite elements*. Wiley, 2nd edition, 2008

[5] For example, in the hp-adaptive finite element method, the order of some elements is increased to improve the approximation locally (so-called p-refinement). This is typically done where the solution is smooth, as there higher-order elements yield the most benefit. Where the solution is less smooth (e.g. near a singularity), the method refines the mesh instead (so-called h-refinement).

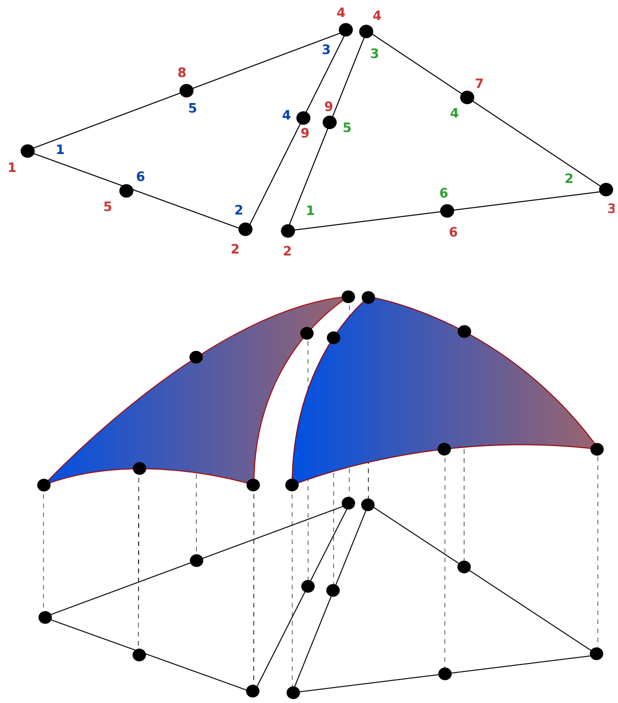

Figure 8.3: The local-to-global mapping for a simple mesh consisting of two triangles, both equipped with secord order Lagrange elements. By mapping matching local degrees of freedom at the common edge to the same global degree of freedom, the local-to-global map ensures the C^0 continuity of the approximation.

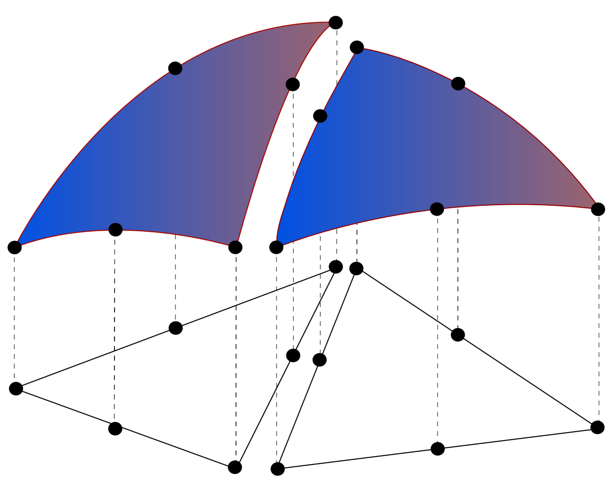

Figure 8.4: By *not* mapping matching local degrees of freedom at the common edge to the same global degree of freedom, a discontinuous approximation results.

In this course we will only consider conforming discretisations, although nonconforming ones are sometimes advantageous for various reasons[6].

The global function space V_h can be described by a mesh $\mathcal{M}$, a set of finite elements $\{(K, \mathcal{V}_K, \mathcal{L}_K) : K \in \mathcal{M}\}$, and a set of local-to-global mappings $\{\iota_K : K \in \mathcal{M}\}$. Once a local interpolation operator on each cell is defined, it is straightforward to define a global interpolation operator mapping to V_h: one merely demands that (8.2.1) is satisfied on each cell.

[6] It also happens sometimes that we wish to approximate a function $u \in L^2(\Omega)$, and thus *discontinuous Lagrange* discretisations are a natural, conforming choice. This arises in certain formulations of Stokes flow (where the pressure $p \in L^2(\Omega)$) and in PDE-constrained optimisation (where we might want to compute the source term $f \in L^2(\Omega)$ such that the solution of the Laplace equation is close to some target).

Definition 8.3.3 (global interpolation operator)**.** *Let V_h be a finite element function space constructed by equipping a mesh $\mathcal{M}$ with finite elements. Then the interpolation operator $\mathcal{I}_h : H \to V_h$ is defined by*

$$\mathcal{I}_h u|_K = \mathcal{I}_K u, \tag{8.3.4}$$

and that $\mathcal{I}_h u$ satisfies any necessary continuity requirements.

9 Local and global assembly

In this lecture, we will discuss the central algorithm executed by a finite element code, the *assembly* algorithm for computing the matrix and vector[1]:

$$A_{ji} = a(\phi_i, \phi_j), \quad b_j = F(\phi_j). \tag{9.0.1}$$

[1] Sometimes called the *stiffness* matrix and *load* vector, terminology that arose in the original context of structural mechanics.

9.1 The assembly algorithm

Given a mesh, and a basis for the trial space

$$V_h = \text{span}\{\phi_i, \quad i = 1, \ldots, N\}, \tag{9.1.1}$$

we seek to compute $A_{ji} = a(\phi_i, \phi_j)$ and $b_j = F(\phi_j)$. For simplicity of discussion, we will discuss the assembly of the stiffness matrix; the assembly of the load vector will be analogous. We also focus on the case where all cells in the domain are equipped with the same finite element basis functions.

The naïve algorithm for doing so is the following. This has two

Algorithm 9.1.1: The naïve algorithm for assembly.

1: **for** $i = 1, \ldots, N$ **do**
2:     **for** $j = 1, \ldots, N$ **do**
3:         Compute $A_{ji} = a(\phi_i, \phi_j)$.

major flaws. First, our finite element basis functions have local support, so most pairs of basis functions (ϕ_i, ϕ_j) do not overlap. In this case, the corresponding matrix entry is zero; we only need to compute A_{ji} for neighbouring basis functions, and algorithm 9.1.1 is very wasteful. Second, each evaluation of $a(\phi_i, \phi_j)$ involves an integration over the domain Ω, which in turn is broken up into integrations over each cell $K \in \mathcal{M}$. Each cell will be visited multiple times in the course of executing 9.1.1. This is wasteful, because it means that the relevant transformations will be recomputed several times.

A much better idea is to express the assembly operation as an iteration over the cells of the mesh, visiting each cell exactly once, and assembling only those contributions that we know in advance to be nonzero. For each cell $K \in \mathcal{M}$, let $\{\phi_i^K\}_{i=1}^d$ denote the restrictions of the relevant basis functions to that cell, where d is the dimension of the local function space $\mathcal{V}_K$; all other global basis functions are zero there. Furthermore, recall that ι_K is the local to global map that sends the local numbering $\{1, \ldots, d\}$ to the global numbering, a subset of $\{1, \ldots, N\}$. Over each cell, we will compute a local $d \times d$ matrix A_K that will contain *all* contributions for all basis functions relevant to this cell. We will then insert that into the global data structure, using the local-to-global map ι_K to discover the relevant indices. By A_{ι_K, ι_K}, we mean the $d \times d$ submatrix of A formed by

Algorithm 9.1.2: The efficient algorithm for assembly.

for $K \in \mathcal{M}$ **do**
    Fetch the local-to-global map ι_K.

    Compute the *local* tensor A_K:
    for $i = 1, \ldots, d$ **do**
        for $j = 1, \ldots, d$ **do**
            Compute $(A_K)_{ji} = a(\phi_i^K, \phi_j^K)$ (only on the cell K).

    Add the local tensor to the global tensor:
    $A_{\iota_K, \iota_K} \stackrel{+}{=} A_K$

taking the entries of the rows and columns corresponding to the global indices of the degrees of freedom associated with K. By $\stackrel{+}{=}$, we mean the mathematical operation of adding on the right-hand side to the existing value of the left-hand side. This is much clearer when explained with a figure; see figure 9.1, reproduced from Logg et al[2].

[2] A. Logg, K. A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2011

The process for the load vector b is analogous, except that instead of adding a local matrix to a submatrix of the global matrix, we will add a local vector (of length d) to a subvector (of length d) of the global vector (of length N).

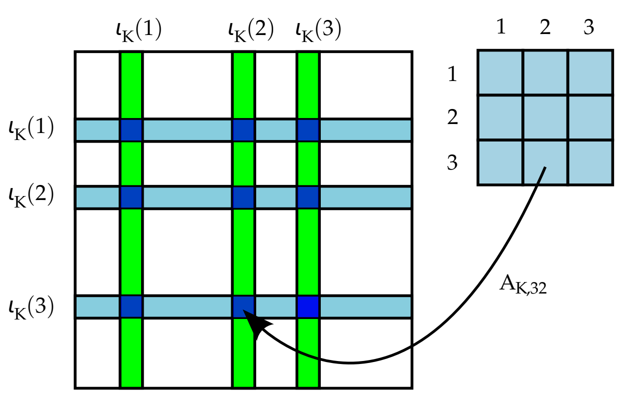

Figure 9.1: Finite element assembly. We loop over each cell K of the mesh and assemble the local stiffness matrix A_K (top right). We add this matrix to the submatrix of the global stiffness matrix A formed by taking the rows and columns associated with the local-to-global map ι_K.

We have seen that if we can compute a local tensor on each cell, we can combine them in a clever way to efficiently assemble the global tensor. How do we calculate the local tensor for each cell, though?

9.2 *Mapping to the reference element*

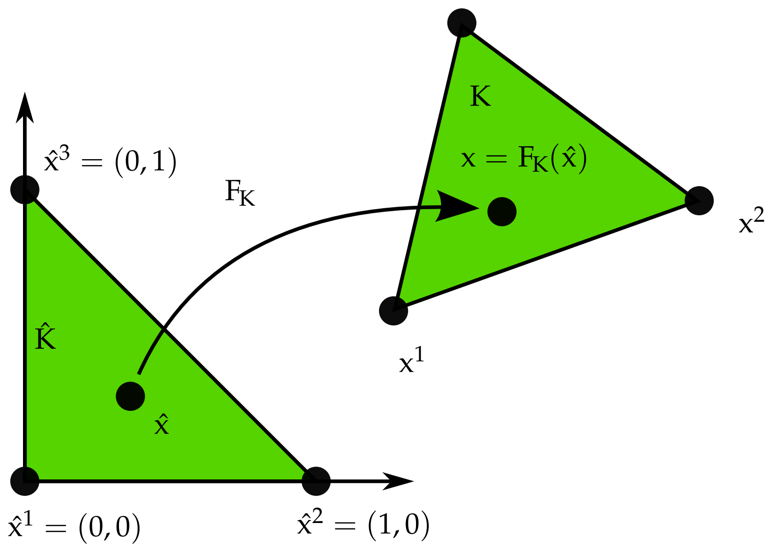

Figure 9.2: The finite element for a specific cell of the mesh K can be constructed via a map F_K applied to an abstract reference element $\hat{K}$.

One of the main implementational advantages of the finite element method is that this description can be significantly simplified by introducing a reference finite element $(\hat{K}, \hat{\mathcal{V}}, \hat{\mathcal{L}})$, and a set of diffeomorphisms $\{F_K : K \in \mathcal{M}\}$ such that

$$K = F_K(\hat{K}) \text{ for all } K \in \mathcal{M}. \tag{9.2.1}$$

This is illustrated in figure 9.2.

For each $K \in \mathcal{M}$, the map F_K generates a function space on K via

$$\mathcal{V}(K) = \{v = \hat{v} \circ F_K^{-1} : \hat{v} \in \hat{\mathcal{V}}\}, \tag{9.2.2}$$

and a set of degrees of freedom on K via

$$\mathcal{L}(K) = \{\ell(v) = \hat{\ell}(v \circ F_K) : \hat{\ell} \in \hat{\mathcal{L}}\}. \tag{9.2.3}$$

By construction, we also obtain the nodal basis $\{\phi_i^K\}_{i=1}^d$ on K from the nodal basis functions on $\hat{K}$. Suppose $\{\hat{\phi}_i\}_{i=1}^d$ is the set of nodal basis functions on $\hat{K}$ satisfying

$$\hat{\ell}_i(\hat{\phi}_j) = \delta_{ij}. \tag{9.2.4}$$

Define $\phi_i^K = \hat{\phi}_i \circ F_K^{-1}$. Computing, we find

$$\ell_j^K(\phi_i^K) = \hat{\ell}_j(\phi_i^K \circ F_K) = \hat{\ell}_j(\hat{\phi}_i \circ F_K^{-1} \circ F_K) = \hat{\ell}_j(\hat{\phi}_i) = \delta_{ij}. \tag{9.2.5}$$

Here we are hoping that the nodal basis functions constructed via the transformation agree with the nodal basis functions computed directly from the finite element on the physical cell. This is the case for most finite elements; such elements are called *affine-equivalent*. This discussion pertains only to affine-equivalent elements. For a more general theory that applies to other elements, see the recent work of Kirby[3].

[3] R. C. Kirby. A general approach to transforming finite elements. *SMAI Journal of Computational Mathematics*, 4:197–224, 2018

Given an expression involving the basis functions to be integrated on a physical cell, we will use the coordinate transformation to convert this to an expression on the reference cell. Let us see this by examples. Consider first the calculation of

$$\int_K \phi_i(x)\phi_j(x)\ \mathrm{d}x, \tag{9.2.6}$$

which is the local contribution to the matrix representing the $L^2(\Omega)$ inner product[4]. From Part A vector calculus, we know that we can transform coordinates in an integral, at the cost of the determinant of the Jacobian of the coordinate transformation, i.e.

[4] This matrix is called the *mass* matrix, another term arising from structural engineering. It is the matrix that must be solved to calculate the $L^2(\Omega)$ Riesz map. Mathematicians sometimes refer to this matrix as the *Gramian* of the basis $\{\phi_1, \dots, \phi_N\}$.

$$\int_K \phi_i(x)\phi_j(x)\ \mathrm{d}x = \int_{\hat{K}} \hat{\phi}_i(\hat{x})\hat{\phi}_j(\hat{x})|J_K(\hat{x})|\ \mathrm{d}\hat{x}. \tag{9.2.7}$$

This latter integral can then be approximated by *quadrature*.

Definition 9.2.1 (quadrature rule of degree m). *A quadrature rule over a cell $\hat{K}$ is a choice of q quadrature points $\hat{x}_i \in \hat{K}$ and weights w_i such that*

$$\int_{\hat{K}} f(\hat{x})\ \mathrm{d}\hat{x} \approx \sum_{i=1}^{q} w_i f(\hat{x}_i). \tag{9.2.8}$$

It has degree of precision *(or degree) m if the approximation is exact for polynomials of degree m or less.*

For example, in one dimension, Gaussian quadrature tells us the optimal choice of weights and quadrature points to maximise the degree of the rule. The fundamental result of Gaussian quadrature is that the optimal choice of q points in an interval gives degree $(2q-1)$; for an introduction to the subject, see Süli & Mayers[5]. In higher dimensions, quadrature is not as straightforward. Fortunately, specialist researchers have dedicated their academic careers to advancing and collating the best known quadrature rules for various domains in an encyclopaedia[6]. We assume that a quadrature scheme is chosen so that the integral is computed exactly[7].

[5] E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, 2003

[6] R. Cools. An encyclopaedia of cubature formulas. *Journal of Complexity*, 19(3):445–453, 2003

[7] This is not always possible. For example, if the right-hand side $f \in L^2(\Omega)$ is not polynomial, then no quadrature scheme will integrate it exactly, and the resulting perturbation to the discrete system must be analysed for a comprehensive error analysis.

Notice that we do not ever need to explicitly calculate the basis functions on each element to calculate the matrix corresponding to $a(\phi_i, \phi_j) = (\phi_i, \phi_j)_{L^2(\Omega)}$; all we need is the tabulation of the reference basis functions at the quadrature points on the reference element, and the Jacobian of the transformation. The quadrature formulae and basis functions can be tabulated in advance; the transformation must be calculated element-by-element.

Now consider the calculation of

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \, \mathrm{d}x, \tag{9.2.9}$$

which is the local contribution to the stiffness matrix[8] for the Laplacian. We have used the symbol ∇_x to emphasise that the derivatives are taken with respect to the physical coordinates. As before, we can transform the integrand, at the cost of a factor involving the determinant of the Jacobian:

[8] Incidentally, this is the matrix that must be solved to calculate the $H_0^1(\Omega)$ Riesz map.

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \, \mathrm{d}x = \int_{\hat{K}} \nabla_x \hat{\phi}_i(\hat{x}) \cdot \nabla_x \hat{\phi}_j(\hat{x}) |J_K(\hat{x})| \, \mathrm{d}\hat{x}, \tag{9.2.10}$$

but this expression is still not computable as it requires the derivatives with respect to the physical coordinate. To eliminate these, we apply the chain rule:

$$\frac{\partial \phi}{\partial x_k} = \sum_l \frac{\partial \hat{x}_l}{\partial x_k} \frac{\partial \phi}{\partial \hat{x}_l}, \tag{9.2.11}$$

i.e. the derivatives transform according to the matrix $\partial \hat{x}_l / \partial x_k$. After some calculation, it can be shown that this is the transpose inverse of the matrix $\partial x_k / \partial \hat{x}_l$, i.e.

$$\nabla_x \hat{\phi}(\hat{x}) = J_K^{-T}(\hat{x}) \nabla_{\hat{x}} \hat{\phi}(\hat{x}). \tag{9.2.12}$$

Thus, the integral can be written as

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \, \mathrm{d}x = \int_{\hat{K}} \left(J_K^{-T} \nabla_{\hat{x}} \hat{\phi}_i(\hat{x}) \right) \cdot \left(J_K^{-T} \nabla_{\hat{x}} \hat{\phi}_j(\hat{x}) \right) |J_K(\hat{x})| \, \mathrm{d}\hat{x}. \tag{9.2.13}$$

This formula is now in a form suitable for the application of a quadrature rule. Here, we need a tabulation of the various *derivatives* of the basis functions with respect to each direction at each quadrature point, which again can be computed offline.

To summarise, we don't need to compute basis functions, their derivatives, or quadrature rules on arbitrary cells: we can tabulate all of the necessary data offline on a reference cell, and then compute the integrals using only the coordinate transformation.

We now wish to study the maps F_K. Before we do so, we need a prelude.

9.3 Prelude: vector elements

Consider our standard Laplacian example (P). This is a scalar problem, as we are solving for a *scalar field*, a smooth function $u : \Omega \to \mathbb{R}$

that associates a scalar value to each point in the domain. We can approximate u_h as an expansion in the span of $\Phi = \{\phi_1, \phi_2, \ldots, \phi_N\}$, our global finite element basis functions construced from a Lagrange finite element of degree q.

Next consider the linear elasticity example discussed in section 7.6, for concreteness posed in two dimensions. This is a vector problem, as we are solving for a *vector field*, a smooth function $u : \Omega \to \mathbb{R}^2$ that associates a vector value to each point in the domain. How should we approximate its solution?

Since we are approximating a vector field, we will want to use vector-valued basis functions; we always want our degrees of freedom to be scalars. We can approximate u_h as an expansion in the span of

$$\Psi = \{(\phi_i, 0) : \phi_i \in \Phi\} \cup \{(0, \phi_i) : \phi_i \in \Phi\}, \tag{9.3.1}$$

a finite dimensional space of dimension $2N$. This can be interpreted as arising from a vector Lagrange element as follows.

Definition 9.3.1 (vector Lagrange element of degree q)**.** *The vector Lagrange element of degree q and dimension k is a finite element $(K, \mathcal{V}, \mathcal{L})$ where*

- *K is a simplex (interval, triangle, tetrahedron),*
- $\mathcal{V} = \underbrace{\mathcal{P}_q(K) \times \mathcal{P}_q(K) \cdots \times \mathcal{P}_q(K)}_{k \text{ times}},$
- $\ell_{ij} : v \mapsto (v(x_i))_j, \;\; i = 1, \ldots, n(q), \;\; j = 1, \ldots, k.$

We denote this element by $\mathrm{CG}_q^k = \mathrm{CG}_q \times \cdots \times \mathrm{CG}_q$.

In two dimensions, the space $\mathcal{V}$ is the space of pairs of q-degree polynomials, and $\mathcal{L}$ is the set of pointwise evaluations of both components at the usual Lagrange points[9]. Given such an expansion, we can write a vector field as

$$u_h = \sum_{j=1}^{2N} c_j \psi_j \tag{9.3.2}$$

for $\psi_j \in \Psi$.

[9] Notice that the geometric dimension of the cell K and the number of components of the vector field do not *always* have to match, although they usually do. For example, if we are solving a problem in atmospheric physics for the flow of air around the Earth, we might model this is as solving for a three-dimensional vector field on a two-dimensional manifold. Our mesh would consist of triangles embedded in $\mathbb{R}^3$; each cell K would be triangular, but we would have $k = 3$.

9.4 More details on the element map

In order to understand the element map, we have to understand that *the coordinate field is just another smooth vector field over the domain*. Just as the solution of linear elasticity smoothly associates a vector with every point in the domain Ω, so does the coordinate field. We can *therefore represent the coordinate field with finite element basis functions.*

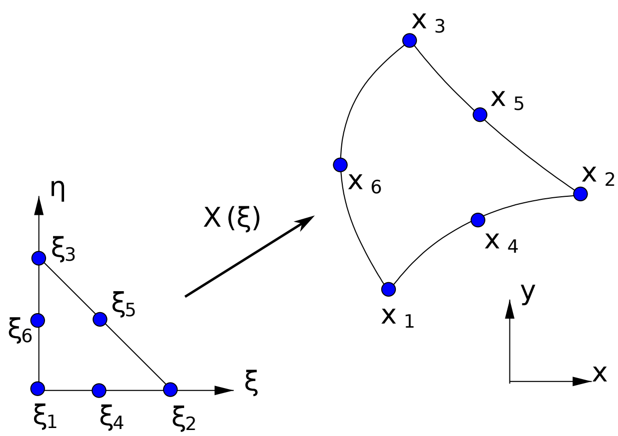

Figure 9.3: It is possible to represent the coordinate field with Lagrange elements of higher order, allowing us to bend the mesh. This is useful if Ω is not a polygon or polyhedron.

If the domain Ω is polyhedral or polygonal, then a linear Lagrange representation of the coordinate is sufficient; the facets of our mesh will be planar, which is enough to represent the domain exactly. If the domain is curved, then it can be advantageous to represent the coordinate field with higher-order basis functions, such as that shown in figure 9.3, adapted from Johnen et al.[10].

[10] A. Johnen, J.-F. Remacle, and C. Geuzaine. Geometrical validity of curvilinear finite elements. *Journal of Computational Physics*, 233:359–372, 2013

This means that for each element we can write

$$x = \sum_i x_i \hat{\psi}_i(\hat{x}) \tag{9.4.1}$$

for (scalar-valued) coefficients x_i and (vector-valued) basis functions $\hat{\psi}_i$. This is an explicit construction for the map $x = F_K(\hat{x})$. Since we have the map explicitly, we can also compute its Jacobian explicitly.

An important special case is when the elements are simplicial[11] and the basis functions used for the coordinate are linear. Since the derivative of a linear function is constant, the Jacobian $J_K(\hat{x})$ is constant over the reference element. This means that the element map is *affine*, i.e. a linear transformation composed with a translation. This means it can be written as

[11] An interval, triangle or tetrahedron.

$$F_K(\hat{x}) = J_K\hat{x} + b_K \tag{9.4.2}$$

where $J_K \in \mathbb{R}^{n\times n}$ and $b_K \in \mathbb{R}^n$. This greatly simplifies the expressions given in (9.2) above[12].

[12] This is *not* true on the first-order quadrilateral element; the local function space $\mathcal{V} = \text{span}\{1, \hat{x}_1, \hat{x}_2, \hat{x}_1\hat{x}_2\}$, and the nonlinear cross term ensures that the Jacobian varies over the element.

9.5 *Solving the assembled system*

Once we have assembled a linear system

$$Ax = b \tag{9.5.1}$$

we must solve it for the coefficients of the expansion of u in terms of our basis. The typical attitude among practitioners is to completely separate the assembly and solution algorithms: the matrix and vector are passed to a black-box solver that knows nothing about the infinite dimensional problem underpinning their construction. This approach works well for small problems[13] but is disastrous for larger problems where matrix factorisations are unaffordable. *A knowledge of functional analysis is essential for the efficient solution of the discrete problem; it is not just a theoretical exercise.*

To see why, let us consider the algorithms employed to solve large sparse[14] linear systems. All such algorithms fundamentally rely on the matrix action, using the matrix to multiply vectors in a clever way to converge to the solution. The simplest such algorithm is Richardson iteration, a fixed point iteration of the form

$$x_{m+1} = x_m - \alpha(Ax_m - b), \tag{9.5.2}$$

where α is a real parameter that must be chosen properly for the iteration to converge. This algorithm converges quite slowly (if at all). A better class of algorithms for this problem are Krylov methods, which rely on the construction of the *Krylov subspace* of order m, given by

$$K_m(A, b) = \text{span}\{b, Ab, \dots, A^{m-1}b\}. \tag{9.5.3}$$

The most celebrated Krylov method is the conjugate gradient method, which is designed for the case where A is symmetric and positive definite. In this case the unique solution x to the equation $Ax = b$ can be characterised as the minimum of the energy

$$x = \operatorname*{argmin}_{y \in \mathbb{R}^N} \frac{1}{2} y^T A y - b^T y \tag{9.5.4}$$

and the conjugate gradient method at iteration m approximates x by

$$x_m = \operatorname*{argmin}_{y \in K_m(A,b)} \frac{1}{2} y^T A y - b^T y. \tag{9.5.5}$$

In this regard, it is another Ritz-Galerkin method, just like the finite element method that assembled the A and b in the first place[15].

Suppose A arises from the discretisation of our running Laplacian example, and so is symmetric and positive-definite, and hence the method of conjugate gradients (CG) is applicable. Each iteration of CG requires one matrix-vector product, which has a cost of $O(N)$, proportional to the number of degrees of freedom, as A is sparse. Thus, *if we can solve the problem in a number of Krylov iterations independent of the mesh*, we will have an $O(N)$ solver[16], and we will be able to solve very fine problems. However, if the number of Krylov iterations increases as we refine the mesh, then our method will be $O(N^2)$, and we will be limited in the size of the problems we can consider.

[13] By a small problem, I mean a problem where sparse LU factorisation works in a reasonable time. Typical numbers on modern machines are a few million degrees of freedom in two dimensions, and perhaps a million degrees of freedom in three dimensions. The three-dimensional case is harder because the bandwidth of the matrix is larger.

[14] A sparse $N \times N$ matrix is one where most of the entries are zero. This means that its matrix-vector product $A : x \to Ax$ can be computed in $O(N)$ time, not the usual $O(N^2)$ time.

[15] Ritz refers to Walther Ritz, who suggested approximating the solution of minimisation problems in smaller subspaces.

[16] This is optimal: we have to compute the values of N degrees of freedom, so the best we can possibly hope for is $O(N)$ runtime.

Unfortunately, if CG is applied naïvely to the linear system

$$Ax = b \tag{9.5.6}$$

the number of iterations required will increase as the mesh is refined. The reasons for this have only recently been fully understood. The fundamental explanation of this phenomenon is that A *is a map from (coefficients of) the primal space* V_h *to (coefficients of) the dual space* V_h^*. That is, $Ax \in V_h^* \neq V_h$, and so *the mathematical expression* A^2x *is not valid*. It simply does not make any sense. Similarly, it does not make sense to look for our solution (a primal vector) in the span of $\{b, Ab, \ldots\}$, as these are dual vectors.

In order for our Krylov method to be defined, we need to compose our matrix A with an operator $P : V_h^* \to V_h$, so that the composition $PA : V_h \to V_h$ and we can look for our solution in the span of $\{Pb, PAb, (PA)^2b, \ldots\}$. The natural choice for P is the Riesz map. In fact, with this choice of P, one can prove mesh-independent convergence of conjugate gradients to a certain tolerance for symmetric positive-definite problems; the number of Krylov iterations required depends in a simple way on the coercivity constant α and the continuity constant C[17].

Just as in the case of optimisation problems in infinite dimensions as discussed in section 6.5, a working knowledge of functional analysis is essential for the efficient *solution* of our discrete problem, not just the analysis of well-posedness or convergence.

[17] K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011

10 Finite elements beyond Lagrange

Not all problems can or should be solved with Lagrange elements. Indeed, one of the beautiful properties of the finite element method is its adaptibility to the particular details of different PDEs.

Before we introduce any other finite elements, we will first discuss a useful coordinate system on triangles and tetrahedra.

10.1 Prelude: barycentric coordinates on a triangle

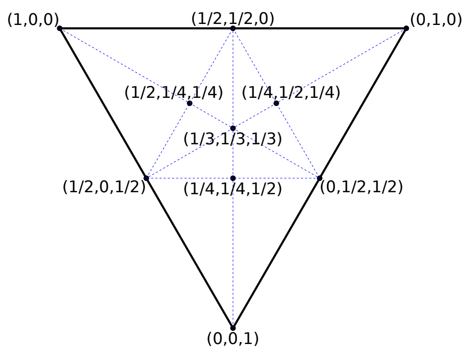

Figure 10.1: Barycentric coordinates on a triangle. Taken from the wikipedia, by Rubybrian, CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=4842309`.

In the next section we will introduce some elements other than the Lagrange family. To prove their unisolvence, it will be convenient to describe positions by barycentric coordinates. These coordinates were first considered by Möbius in 1827.

Let K be a triangle. Any point $p \in K$ can be written uniquely as a convex combination of the vertices. That is, there are three numbers $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$ such that $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and

$$p = \lambda_1 r_1 + \lambda_2 r_2 + \lambda_3 r_3. \tag{10.1.1}$$

In barycentric coordinates, we represent $p = (\lambda_1, \lambda_2, \lambda_3)$; while there are three numbers, there are only two degrees of freedom, due to the summation constraint. For a chart of barycentric coordinates on equilateral and right-angled triangles, see figure 10.1. One advantage of this coordinate system is that it allows us to compactly define polynomials with certain geometric features. For example, the polynomial

$p(x) = \lambda_1$ is the linear polynomial which is one at r_1 and zero on the edge opposite r_1.

Analogously, points in tetrahedra can be described by four numbers $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ which sum to one.

10.2 *The biharmonic equation: $H^2(\Omega)$-conforming elements*

Not all PDE problems can be posed over $H^1(\Omega)$, and thus not all problems can be solved in a conforming manner with continuous Lagrange finite elements. We now give an example of this arising in a physically important application.

The biharmonic equation is a fourth-order PDE arising in elasticity theory; it describes the equilibrium solutions of clamped plates under transverse loading, the stresses in an elastic body, the stream function in creeping flow of a viscous incompressible fluid, and other things besides. The equation is given by

$$\begin{aligned} \nabla^4 u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma, \\ \nabla u \cdot n &= 0 \text{ on } \Gamma. \end{aligned} \tag{10.2.1}$$

The operator ∇^4 can also be written as the square of the Laplacian Δ^2. In two dimensions, it means that

$$\frac{\partial^4 u}{\partial x^4} + \frac{\partial^4 u}{\partial y^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} = f. \tag{10.2.2}$$

Let us take this into variational form formally, i.e. without stating upfront what space we will take our test function v from. We will subsequently inspect the problem and enforce the conditions on v that make the variational formulation sensible. Multiplying by $v \in V$ for some V, we find

$$\int_\Omega \nabla^4 u v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x. \tag{10.2.3}$$

As we wish to invoke Lax–Milgram, we want the regularity requirements on u to be the same as v, and so we wish to equidistribute the derivatives somehow; we need to develop an integration by parts result analogous to (IBP).

Recall again the divergence theorem: if w is a sufficiently regular vector field defined on a sufficiently regular domain Ω, then

$$\int_\Omega \nabla \cdot w \,\mathrm{d}x = \int_\Gamma w \cdot n \,\mathrm{d}s, \tag{10.2.4}$$

where n is the unit outward facing normal to Ω on Γ. Defining

$$w = v \nabla(\nabla^2 u), \tag{10.2.5}$$

we have

$$\nabla \cdot w = \nabla v \cdot \nabla(\nabla^2 u) + v\nabla^4 u \tag{10.2.6}$$

and thus

$$\int_\Omega \nabla v \cdot \nabla(\nabla^2 u)\ \mathrm{d}x + \int_\Omega v\nabla^4 u\ \mathrm{d}x = \int_\Gamma v\nabla(\nabla^2 u)\cdot n\ \mathrm{d}s. \tag{10.2.7}$$

We can simplify the first term further, by applying our second-order integration by parts formula (IBP) to the first term:

$$\int_\Omega \nabla v \cdot \nabla(\nabla^2 u)\ \mathrm{d}x = -\int_\Omega \nabla^2 v\nabla^2 u + \int_\Gamma \nabla^2 u\nabla v\cdot n\ \mathrm{d}s. \tag{10.2.8}$$

Thus, (10.2.7) simplifies to

$$\int_\Omega v\nabla^4 u\ \mathrm{d}x = \int_\Omega \nabla^2 v\nabla^2 u\ \mathrm{d}x + \int_\Gamma v\nabla(\nabla^2 u)\cdot n\ \mathrm{d}s - \int_\Gamma \nabla^2 u\nabla v\cdot n\ \mathrm{d}s. \tag{BIBP}$$

Applying our integration by parts result to (10.2.3), we find

$$\int_\Omega \nabla^2 v\nabla^2 u + \int_\Gamma v\nabla(\nabla^2 u)\cdot n\ \mathrm{d}s - \int_\Gamma \nabla^2 u\nabla v\cdot n\ \mathrm{d}s = \int_\Omega fv\ \mathrm{d}x. \tag{10.2.9}$$

Recall that $u = 0 = \nabla u \cdot n$ on Γ. There is nowhere natural to enforce weakly this in the variational formulation, and so we will have to enforce this strongly in the choice of V. Furthermore, inspecting the variational form, it is clear we need square integrability of the second derivatives of u and v. Thus, the proper choice for the space V is

$$V = H_0^2(\Omega) = \{v \in H^2(\Omega) : v = 0 = \nabla v \cdot n \text{ on } \Gamma\}. \tag{10.2.10}$$

With this choice of function space, the surface integral terms disappear, and thus the variational form is

$$\text{find } u \in H_0^2(\Omega) \text{ such that } \int_\Omega \nabla^2 u\nabla^2 v\ \mathrm{d}x = \int_\Omega fv\ \mathrm{d}x \text{ for all } v \in H_0^2(\Omega). \tag{10.2.11}$$

This bilinear form is in fact an inner product on $H_0^2(\Omega)$, and so continuity and coercivity follow naturally with $C = \alpha = 1$. Thus the variational problem is well-posed.

If we discretise (10.2.11) with continuous Lagrange elements, then $V_h \subset H^1(\Omega)$ but $V_h \not\subset H^2(\Omega)$, and the discretisation is not conforming. As we know, a piecewise smooth function belongs to $H^1(\Omega)$ if and only if it is continuous. Since a function belongs to $H^2(\Omega)$ only if it and all its first derivatives belong to $H^1(\Omega)$, a piecewise smooth function belongs to $H^2(\Omega)$ if and only if it is $C^1(\Omega)$. This means that a finite element Galerkin method for the biharmonic problem requires C^1 finite elements, something that a continuous Lagrange element cannot satisfy.

In one dimension, it is easy to achieve this with the *Hermite* element: $K = [0,1]$, $\mathcal{V} = \mathcal{P}_3(K)$ the space of third degree polynomials,

Figure 10.2: The Hermite element in one dimension is $C^1(\Omega)$-conforming, and thus $H^2(\Omega)$-conforming. The degrees of freedom are pointwise evaluation at the endpoints, and pointwise evaluation of the derivative at the endpoints.

and $\mathcal{L}$ pointwise evaluation at the endpoints and derivative evaluation at the endpoints, figure 10.2.

Let us therefore consider the analogous element in two dimensions, depicted in figure 10.3. Here, the dimension of the space of cubic polynomials is 10. Pointwise evaluation at the vertices supplies three degrees of freedom; evaluation of the two components of the derivative at the endpoints supplies six more, leaving one left over, which we thus take to be pointwise evaluation at the barycentre.

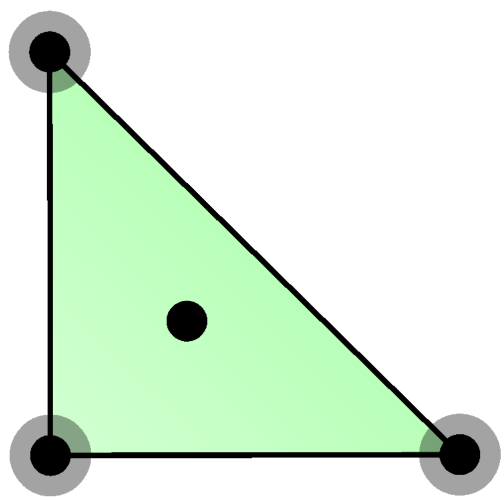

Figure 10.3: The Hermite element in two dimensions is not $C^1(\Omega)$-conforming, and thus is not $H^2(\Omega)$-conforming. The degrees of freedom are pointwise evaluation at the vertices, pointwise evaluation of both components of the derivative at the vertices, and pointwise evaluation at the barycentre.

Lemma 10.2.1 (Unisolvence of the triangular Hermite element)**.** *The Hermite element in two dimensions is unisolvent.*

Proof. Suppose u is a cubic polynomial on K and that all of the degrees of freedom evaluate to zero. Along an edge of the triangle, u belongs to the space of cubic polynomials along an interval, and both its pointwise values and its derivatives at the endpoints vanish, and thus $u = 0$ along each edge. This holds for each edge. Thus, u must be a multiple of the cubic bubble function $b(x) = \lambda_1\lambda_2\lambda_3$, $u(x) = cb(x)$ for some c. However, since the value at the barycentre is also zero, and $b(x) \neq 0$ at the barycentre, we must have $c = 0$. $\square$

Thus the degrees of freedom $\mathcal{L}$ determine the space $\mathcal{V}$. Unfortunately, however, this element is not $C^1(\Omega)$-conforming in two dimensions! In our argument we saw that the degrees of freedom on an edge of K determine $u|_K$, and so the element is $C^0(\Omega)$-conforming; for the global approximation to be $C^1(\Omega)$ we must also have that the degrees of freedom on an edge determine $\nabla u \cdot n$. The degrees of freedom only determine the value of $\nabla u \cdot n$ at the two endpoints of the edge; but $\nabla u \cdot n$ is a polynomial of degree two, which requires three degrees of freedom to be uniquely determined. Thus, the approximation is *not* $C^1(\Omega)$-conforming in two (or higher) dimensions.

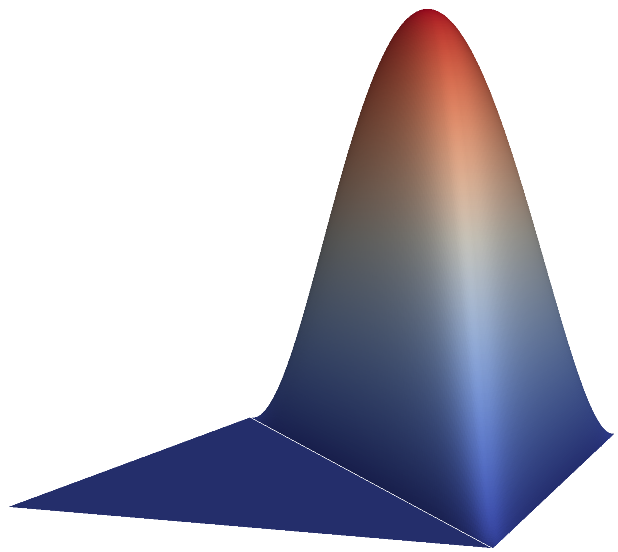

Figure 10.4: The Hermite element does not guarantee a $C^1(\Omega)$ discretisation in two dimensions. The function (10.2.12) is in the Hermite space but its normal derivative is not continuous at the shared edge.

For a specific counterexample, consider two adjacent elements K_1 and K_2, where

$$p(x) = \begin{cases} \lambda_1\lambda_2\lambda_3 & x \in K_1, \\ 0 & x \in K_2, \end{cases} \tag{10.2.12}$$

rendered in figure 10.4. This is an element of the function space induced by the Hermite element. All degrees of freedom evaluate to zero at the shared interface (i.e. $\nabla p \cdot n = 0$ at the common vertices), but $\nabla p \cdot n \neq 0$ over the whole edge.

Thus, if we wish to have a conforming discretisation, we must develop another finite element. The $C^1(\Omega)$-conforming element of minimal degree is called the *Argyris* element. In one dimension, it consists of $K = [0,1]$, $\mathcal{V} = \mathcal{P}_5(K)$, and $\mathcal{L}$ consists of the evaluation of the function, its first and its second derivatives at the endpoints[1]. In two dimensions, the degrees of freedom consist of all zeroth, first and second derivatives at the vertices, and the value of the normal derivative in the centre of the edges, figure 10.5.

[1] Thus, in one dimension, the Argyris element is $C^2(\Omega)$-conforming.

Lemma 10.2.2 (Unisolvence of the triangular Argyris element)**.** *The Argyris element in two dimensions is unisolvent.*

Proof. Suppose u is such that all degrees of freedom vanish. By the

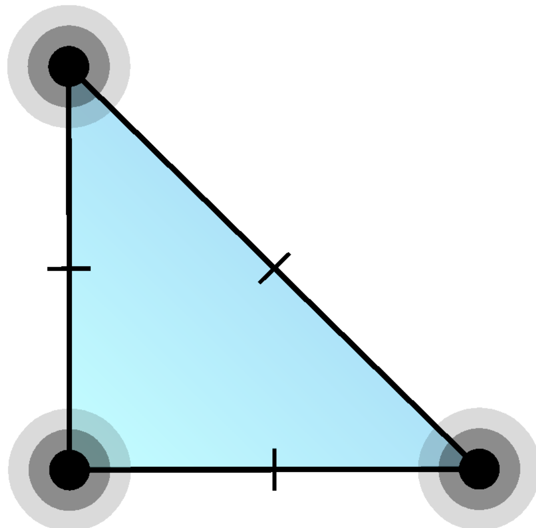

Figure 10.5: The Argyris element in two dimensions is $C^1(\Omega)$-conforming, and thus is $H^2(\Omega)$-conforming. The degrees of freedom are evaluation of the zeroth, first and second derivatives at the vertices, plus evaluation of the normal derivative at the midpoint of the edges.

unisolvence of the Hermite quintic in one dimension, u vanishes on each edge. On each edge, $\nabla u \cdot n$ is a quartic polynomial which vanishes at the endpoints and midpoint, and whose derivatives at the endpoints are zero, so $\nabla u \cdot n = 0$ on the edge too. Since a polynomial and its normal derivative vanish on the line $\lambda_i = 0$ if and only if it is divisible by λ_i^2, u must be a multiple of $b^2(x) = \lambda_1^2 \lambda_2^2 \lambda_3^2$, which is a polynomial of degree six. Since u can only be a polynomial of degree at most five, it must vanish. $\square$

The Argyris element is very useful for certain applications. The Lagrange elements are the most commonly used, but they are not universal.

11 *Interpolation error estimates*

In lecture 7, we saw that for coercive problems the Galerkin approximation enjoys a quasi-optimality property:

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_V,$$

where C is the continuity constant and α is the coercivity constant. In words, the Galerkin approximation in the space V_h is almost as good as it could be, up to problem-specific constants. While this is reassuring, it is not very concrete: ideally, we would like to know more about how the error behaves as we refine the parameters of our discretisation that are actually under our control, the mesh size h and the approximation order of our finite element p. This will be achieved by *interpolation error estimates*. We will bound the error between a function u and its interpolant $\mathcal{I}_h u$, and use the fact that

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_V \leq \frac{C}{\alpha} \|u - \mathcal{I}_h u\|_V. \tag{11.0.1}$$

In this lecture, we will see how this latter quantity can be bounded for different finite elements and choices of norm V.

11.1 *Prelude: Sobolev seminorms*

In lecture 3, we saw the Sobolev space $W_p^k(\Omega)$ for $p < \infty$ is equipped with the norm

$$\|u\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p}. \tag{11.1.1}$$

It will be convenient in what follows to use the Sobolev seminorm[1]:

$$|u|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| = k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p}. \tag{11.1.2}$$

This just means to take the $L^p(\Omega)$ norms of the k^{th} derivatives of u, instead of all derivatives up to and including order k. This is a

[1] A seminorm $|\cdot| : V \to \mathbb{R}^+$ satisfies the nonnegativity, scaling and the triangle inequality properties of a norm, but differs from a norm in that $|u| = 0 \not\Rightarrow u = 0$.

seminorm because there can exist functions $u \in W_p^k(\Omega)$ for which $u \neq 0$ but $|u| = 0$.

As a concrete example, observe that

$$\|u\|_{H^1(\Omega)}^2 = \|u\|_{L^2(\Omega)}^2 + |u|_{H^1(\Omega)}^2 \tag{11.1.3}$$

and that $|u|_{H^1(\Omega)}$ is a norm on $H_0^1(\Omega)$, the norm induced by the bilinear form for the negative Laplacian.

11.2 Prelude: measuring the mesh size

The interpolation error estimates will depend on the polynomial order of approximation p and some notion of the mesh size h. For the case of a one-dimensional mesh where every cell has the same length h, it is obvious that we should describe the mesh resolution with that quantity h; for non-uniform meshes in higher dimensions, it is less clear how we should characterise our meshes.

The quantity we will use on each cell is its *diameter*.

Definition 11.2.1 (diameter of a cell)**.**

$$h_K = \operatorname{diam}(K) = \sup\{\|x - y\| : x, y \in K\}. \tag{11.2.1}$$

For a triangle or tetrahedron, this resolves to the length of its longest edge.

To describe the resolution of the mesh as a whole, we will take a pessimistic viewpoint and consider the resolution of the worst element in that mesh.

Definition 11.2.2 (mesh size)**.** *Given a mesh $\mathcal{M}$, its mesh size h is given by*

$$h = \max_{K \in \mathcal{M}} \operatorname{diam}(K). \tag{11.2.2}$$

We will consider a sequence of meshes $(\mathcal{M}_h)_h$ indexed by the mesh size h. For the following results to hold, we will need a technical condition on the sequence of meshes.

Definition 11.2.3 (incircle diameter of a cell)**.** *The incircle diameter ρ_K of a cell K is the diameter of the largest hyperdisc (i.e. disc in two dimensions, ball in three dimensions) that is completely contained within K.*

Definition 11.2.4 (shape regularity of mesh sequence $(\mathcal{M}_h)_h$)**.** *A sequence of meshes $(\mathcal{M}_h)_h$ is* shape regular *if there exists a constant σ such that*

$$\sup_h \max_{K \in \mathcal{M}_h} \frac{h_K}{\rho_K} \leq \sigma. \tag{11.2.3}$$

Informally, this means that the meshes don't get arbitrarily stretched as the meshes are refined: the quantity h_K/ρ_K measures how anisotropic (needle-like) each element is, and we ensure that this remains bounded as we refine the mesh size h.

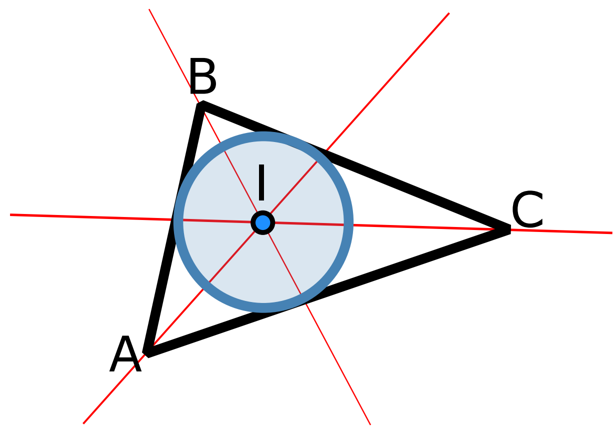

Figure 11.1: A triangle K and its incircle.

11.3 Interpolation error for Lagrange elements

First, consider problems posed in $V = H^1(\Omega)$ approximated with Lagrange elements. For this finite element, we have the following interpolation error bound[2].

[2] Verfürth, R. A note on polynomial approximation in Sobolev spaces. *Mathematical Modelling and Numerical Analysis*, 33(4):715–719, 1999; and D. N. Arnold, G. Awanou, and R. Winther. Nonconforming tetrahedral mixed finite elements for elasticity. *Mathematical Models and Methods in Applied Sciences*, 24(04):783–796, 2014

Theorem 11.3.1 (Interpolation error in the $H^1(\Omega)$-norm for continuous Lagrange elements of order p)**.** *Let $(V_h)_h$ be the function spaces constructed with continuous Lagrange elements of order p on a shape-regular sequence of meshes $(\mathcal{M}_h)_h$ indexed by mesh size h. Let $u \in H^{p+1}(\Omega)$, and let $\mathcal{I}_h : H^{p+1}(\Omega) \to V_h$ be the interpolation operator associated with each V_h. Then there exists a constant $D < \infty$ independent of u such that*

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq D h^p |u|_{H^{p+1}(\Omega)}. \tag{11.3.1}$$

Let's think about what this means. In words, it says that *the interpolation error with degree p depends on the size of the next highest derivatives of order $p+1$*, i.e. those not captured in the interpolation. So if we were interpolating a function in one dimension with linear elements, the error will be high wherever the second derivative (the curvature) of the function is large. This aligns exactly with our expectations from Part A Numerical Analysis. On the other hand, if the second derivatives were in fact zero, we would approximate a linear function with a linear function and get zero interpolation error. In higher dimensions, the result is the same, but we need to look at all the different $(p+1)^{\text{th}}$ derivatives.

Importantly, this bound gives us an idea of how the error in the $H^1(\Omega)$ norm will scale like $\mathcal{O}(h^p)$ as a function of our mesh spacing h and our polynomial degree p. If our solutions are smooth enough, i.e. they live in $H^{q+1}(\Omega)$ for $q > p$, then increasing p is our best path to fast convergence: we would much rather change the exponent in h^p than the base! This is called p-refinement, or p-adaptivity if it is driven by an error estimate of the solution. But if our solution isn't so regular (e.g. $u \in H^2(\Omega)$ but $u \notin H^3(\Omega)$), there's not much advantage to increasing p, and we should refine the mesh size h instead. This is called h-refinement, or h-adaptivity if it is driven by an error estimate of the solution. When combined, these two refinement strategies give rise to sophisticated hp-adaptive finite element methods[3].

Assuming that $u \in H^{p+1}(\Omega)$, and that Lagrange elements of degree p are used to approximate a variational problem posed in $H^1(\Omega)$, it follows from (11.0.1) and (11.3.1) that

$$\|u - u_h\|_{H^1(\Omega)} \leq \frac{DC}{\alpha} h^p \, |u|_{H^{p+1}(\Omega)}. \tag{11.3.2}$$

This is a very useful *a priori*[4] $H^1(\Omega)$-error estimate for the Poisson problem (P), for the advection-diffusion equation (5.3.1), for the Helmholtz problem (6.4.4), for the linear elasticity problem (7.6.3), and for problems like them.

[3] C. Schwab. *$p-$ and $hp-$ Finite Element Methods: Theory and Applications to Solid and Fluid Mechanics*. Numerical Mathematics and Scientific Computation. Clarendon Press, 1999

[4] An *a priori* error estimate is one that can be performed without actually computing u_h. There is an alternative (and extremely useful) *a posteriori* error analysis whose evaluation depends on actually computing u_h. These *a priori* estimates are useful for estimating the rate of convergence of the discretisation, whereas *a posteriori* error estimates are useful for driving adaptive refinement (in h or p) to *control* the error in the discretisation.

11.4 Elliptic regularity results

Notice that these estimates require $u \in H^2(\Omega)$ or greater. In general, this requires an elliptic regularity result, an auxiliary argument to prove that the unique solution to the variational problem at hand (that is posed in $H^1(\Omega)$) actually happens to live in $H^2(\Omega)$ or a stronger subspace. In fact, these elliptic regularity results typically bound some higher Sobolev seminorm of the solution in terms of the norms of the data. For example, a typical elliptic regularity result might look like the following.

Theorem 11.4.1 (Example elliptic regularity result)**.** *Let Ω be C^∞-smooth, i.e. possesses a local parametrisation by C^∞ functions. Then the solution $u \in H_0^1(\Omega)$ to the Poisson equation* (Q) *is an element of $H^2(\Omega)$ and satisfies*

$$|u|_{H^2(\Omega)} \leq c\|f\|_{L^2(\Omega)} \tag{11.4.1}$$

for some constant c.

The requirement that Ω has some kind of smoothness is indispensable. For example, if Ω has a re-entrant corner (an L-shape, or a cube with an octant cut out) then the result does not hold and the solution to the Poisson equation genuinely lives in $H^1(\Omega) \setminus H^2(\Omega)$.

11.5 Changing norms: the Aubin–Nitsche duality argument

The interpolation error bound also depends on the norm used to measure the error. The $L^2(\Omega)$ norm (also called the $H^0(\Omega)$ norm) is *weaker* than the $H^1(\Omega)$ norm used in (11.3.1): it only measures how good your approximation of the function values is, while the $H^1(\Omega)$ norm also takes in to account how good your approximation of its derivative is. When you measure the interpolation error in this weaker norm, the convergence rate improves by one[5]:

[5] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag New York, third edition edition, 2008

Theorem 11.5.1 (Interpolation error in the $L^2(\Omega)$-norm for continuous Lagrange elements of order p)**.** *Let $(V_h)_h$ be the function spaces constructed with continuous Lagrange elements of order p on a shape-regular sequence of meshes $(\mathcal{M}_h)_h$ indexed by mesh size h. Let $u \in H^{p+1}(\Omega)$, and let $\mathcal{I}_h : H^{p+1}(\Omega) \to V_h$ be the interpolation operator associated with each V_h. Then there exists a constant[6] $D < \infty$ independent of u such that*

[6] This constant is different to the one mentioned in theorem 11.3.1.

$$\|u - \mathcal{I}_h u\|_{L^2(\Omega)} \leq D h^{p+1} |u|_{H^{p+1}(\Omega)}. \tag{11.5.1}$$

However, this result just refers to interpolation errors between u and $\mathcal{I}_h u$, whereas we are primarily interested in the error between the true solution u and its Galerkin approximation u_h.

This gap may be addressed using the Aubin–Nitsche duality argument[7]. We demonstrate its use for the Poisson equation with linear Lagrange elements.

[7] J.-P. Aubin. *Analyse fonctionnelle appliqué*. Presses Universitaires de France, 1987

Consider the variational problem

$$\text{find } u \in H^1_0(\Omega) \text{ such that } a(u,v) = F(v) \text{ for all } v \in H^1_0(\Omega), \tag{11.5.2}$$

where Ω is C^∞-smooth and

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x. \tag{11.5.3}$$

We know that this has a unique solution $u \in H^1_0(\Omega)$ by the Riesz representation theorem. By the elliptic regularity result 11.4.1, we know that $u \in H^2(\Omega)$; its Galerkin approximation using linear Lagrange finite elements therefore satisfies

$$\|u - u_h\|_{H^1(\Omega)} \leq CD\alpha^{-1} h |u|_{H^2(\Omega)}. \tag{11.5.4}$$

Consider the error $e = u - u_h \in H^1_0(\Omega)$. Given any element of a Hilbert space, we can construct its associated dual element as

$$e^*(v) = (u - u_h, v)_{L^2(\Omega)}. \tag{11.5.5}$$

This $e^* \in H^{-1}(\Omega) = H^1_0(\Omega)^*$ and thus makes sense as the data for an auxiliary ("adjoint" or "dual") problem:

$$\text{find } w \in H^1_0(\Omega) \text{ such that } a(w,v) = e^*(v) \text{ for all } v \in H^1_0(\Omega). \tag{11.5.6}$$

Essentially, this seeks the $H^1_0(\Omega)-$Riesz representative of the error functional $e^*(v)$. We can apply the Riesz representation theorem again to show that this has a unique solution w and can apply the elliptic regularity theorem 11.4.1 to show that there exists a constant c such that

$$|w|_{H^2(\Omega)} \leq c\|e\|_{L^2(\Omega)}. \tag{11.5.7}$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$\begin{aligned}
\|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\
&= a(w, u - u_h) \\
&= a(u - u_h, w - \mathcal{I}_h w) && \text{(symmetry and Galerkin orthogonality)} \\
&\leq C\|u - u_h\|_{H^1(\Omega)}\|w - \mathcal{I}_h w\|_{H^1(\Omega)} && \text{(by continuity of } a\text{)} \\
&\leq CDh\|u - u_h\|_{H^1(\Omega)}|w|_{H^2(\Omega)} && \text{(by interpolation result (11.3.1))} \\
&\leq C^2D^2\alpha^{-1}h^2|u|_{H^2(\Omega)}|w|_{H^2(\Omega)} && \text{(by \textit{a priori} error estimate (11.5.4))} \\
&\leq C^2D^2\alpha^{-1}ch^2|u|_{H^2(\Omega)}\|u - u_h\|_{L^2(\Omega)} && \text{(by elliptic regularity result (11.5.7))}
\end{aligned}$$

and hence there exists a constant C' such that

$$\|u - u_h\|_{L^2(\Omega)} \leq C'h^2|u|_{H^2(\Omega)} \tag{11.5.8}$$

as required. That is, just as we expected from the improved interpolation error estimates in the $L^2(\Omega)$ norm, the Galerkin approximation error is one order better in the $L^2(\Omega)$-norm than in the $H^1(\Omega)$-norm.

This holds more generally: if $u \in H^{p+1}(\Omega)$, then a Galerkin approximation using degree p Lagrange finite elements converges at $\mathcal{O}(h^p)$ in the $H^1(\Omega)$-norm and at $\mathcal{O}(h^{p+1})$ in the $L^2(\Omega)$-norm.

11.6 *Interpolation error for the Argyris element*

Of course, the interpolation error estimates depend on the specific kind of finite element employed. For problems posed in $H^2(\Omega)$ such as the biharmonic equation, other finite elements and hence other interpolation error estimates must be employed. In the case of Argyris elements, the following interpolation error estimates hold[8].

[8] D. Braess. *Finite Elements: theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, third edition, 2007

Theorem 11.6.1 (Interpolation error for the lowest-order Argyris element)**.** *Let $(V_h)_h$ be the function spaces constructed with the fifth-order Argyris element on a shape-regular sequence of meshes $(\mathcal{M}_h)_h$ indexed by mesh size h. Let $u \in H^6(\Omega)$, and let $\mathcal{I}_h : H^6(\Omega) \to V_h$ be the interpolation operator associated with each V_h. Then there exists constants labelled D (all different) such that*

$$\|u - \mathcal{I}_h u\|_{H^2(\Omega)} \leq Dh^4|u|_{H^6(\Omega)}, \tag{11.6.1}$$

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq Dh^5|u|_{H^6(\Omega)}, \tag{11.6.2}$$

$$\|u - \mathcal{I}_h u\|_{H^0(\Omega)} \leq Dh^6|u|_{H^6(\Omega)}. \tag{11.6.3}$$

(11.6.4)

Notice again the general pattern: the weaker the norm, the better the convergence.

12 Nonlinear problems

We have now analysed in some detail linear coercive equations. However, not all physical problems are linear! In this lecture we consider how nonlinear problems may be treated in a variational framework, and how they may be discretised and solved using finite elements.

As a model problem, consider the following Bratu–Gelfand equation:

$$u''(x) + \lambda e^u = 0, \quad u(0) = 0 = u(1). \tag{12.0.1}$$

This problem "appears in a large variety of application areas such as the fuel ignition model of thermal combustion, radiative heat transfer, thermal reaction, the Chandrasekhar model of the expansion of the universe, chemical reactor theory and nanotechnology"[1]. Here u is the temperature, and $\lambda \in \mathbb{R}$ is known as the Frank–Kamenetskii parameter. The equation balances cooling via diffusion (due to the zero boundary conditions) and heating (due to the nonlinear reaction term). Intuitively, one might expect that for large λ, the reaction term will dominate and the temperature will explode, and this is indeed the case: the equation only has solutions for $\lambda \leq \lambda^*$, with

[1] A. Mohsen. A simple solution of the Bratu problem. *Computers & Mathematics with Applications*, 67(1):26–33, 2014

$$\lambda^* = 8 \left(\min_{x>0} \frac{x}{\cosh x} \right)^2 \approx 3.5138307 \tag{12.0.2}$$

in one dimension. The proof of well-posedness for $\lambda \leq \lambda^*$ is far beyond the scope of the course and relies on Schauder's fixed point theorem, discussed in C4.6, Fixed Point Methods for Nonlinear PDEs[2]. In fact, for $\lambda = \lambda^*$, the equation has one solution, and for $0 < \lambda < \lambda^*$ it has two solutions.

[2] M. Rupflin. Lecture notes on Fixed Point Methods for Nonlinear PDEs, 2017. https://courses.maths.ox.ac.uk/node/view_material/2037

How should we formulate this problem variationally, how should it be discretised, and how should it be solved?

12.1 Variational formulation of nonlinear problems

As usual, we multiply our equation by a test function v from some (not yet specified) test space V, and integrate by parts. Our weak form is

$$-\int_0^1 u'(x)v'(x)\,\mathrm{d}x + \int_0^1 \lambda e^u v\,\mathrm{d}x = 0, \tag{12.1.1}$$

and by inspection we take $V = H_0^1(0,1)$, since we only need the existence of one weak derivative for u and v[3].

For linear problems, our abstraction was: find $u \in V$ such that $a(u,v) = F(v)$ for all $v \in V$. For nonlinear problems, we will use a more general abstraction: find $u \in V$ such that

$$G(u;v) = 0 \tag{12.1.2}$$

for all $v \in V$. Here $G : V \times V \to \mathbb{R}$ is nonlinear in u but is linear in v; we remind ourselves of this property by putting the arguments in which G is linear to the right of the semicolon.

Just as in the linear case, it is very useful to reformulate this variational statement as a standard equality. We do this by introducing the operator $H : V \to V^*$, defined by

$$(H(u))\,(v) = \langle H(u), v\rangle = G(u;v). \tag{12.1.3}$$

Solutions of (12.1.2) are exactly roots of H, i.e. solutions $u \in V$ such that $H(u) = 0$, with the equality between elements of the dual space V^*.

For the model problem at hand, H has no roots for $\lambda > \lambda^*$, exactly one root at $\lambda = \lambda^*$, and two roots for $0 < \lambda < \lambda^*$.

To compute a solution, there are two steps required. We must discretise the equation, so that the problem becomes finite and thus amenable to finite computers, and we must devise a scheme to solve the resulting nonlinear problem. We can take these steps in either order. We will first discuss the choice where we discretise first.

[3] We are skipping over a major technical complication here, whether the second term is bounded if $u \in L^2(0,1)$. An operator that is the composition of a scalar function $f : \mathbb{R} \to \mathbb{R}$ with $u(x)$ is called a Nemytskii or superposition operator, and their analysis is rather delicate. We ignore this issue here.

12.2 *Discretisation first*

As before, we discretise by introducing a finite-dimensional closed subspace $V_h \subset V$, and posing the problem over this subspace:

$$\text{find } u_h \in V_h \text{ such that } G(u_h;v_h) = 0 \text{ for all } v_h \in V_h. \tag{12.2.1}$$

Given a basis $V_h = \operatorname{span}\{\phi_1, \dots, \phi_N\}$, this reduces to finding the roots of a nonlinear residual

$$\begin{aligned} H_h : \mathbb{R}^N &\to \mathbb{R}^N \\ H_h(x) &= 0, \end{aligned} \tag{12.2.2}$$

where

$$(H_h)_j\,(x) = G(x_1\phi_1 \cdots x_N\phi_N; \phi_j). \tag{12.2.3}$$

The discretised residual H_h is analogous to H defined in (12.1.3), and maps from (coefficients of) an element of V_h to (coefficients of) an element of V_h^*.

We are now faced with the very difficult task of ensuring that the solutions u_h of our Galerkin approximation approximate in a suitable norm. For if we want to control the norm $\|u - u_h\|$, the question immediately arises: which u, and which u_h? After all, the continuous and discretised nonlinear problems will in general support multiple solutions, and it is not clear how to pair solutions of the two problems together in the right way. Even if one could guarantee that each solution u was unique within a neighbourhood, and that asymptotically there was exactly one u_h inside each neighbourhood that converged $u_h \to u$, that would still be insufficient; for it might be the case that the discretised problem supported *spurious* solutions, i.e. solutions that did not correspond to solutions of the continuous problem. In our worst nightmares, a sequence of such spurious solutions might appear to converge — to something that is not a continuous solution. Unfortunately, this can actually happen, as shown in example 8.6 of Deuflhard[4].

[4] P. Deuflhard. *Newton Methods for Nonlinear Problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer-Verlag, 2011

We are treading deep waters; in many cases of interest the resolution of these questions is simply not known. Thus, in this course, we will sidestep these fundamental and intricate questions of existence and convergence[5]. We will do this elegantly by considering the alternative approach where we devise a scheme to solve the nonlinear problem *first*, and postpone discretisation until the very last possible moment.

[5] If you want to face them head on, a good place to start is the review article of Caloz and Rappaz:

G. Caloz and J. Rappaz. Numerical analysis for nonlinear and bifurcation problems. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume 5, pages 487–637. Elsevier, 1997

The scheme we will use to (abstractly) solve our infinite-dimensional nonlinear PDE is Newton's method, or the Newton–Kantorovich algorithm as it is known in Banach spaces. Before we discuss the general theory, let us remind ourselves of Newton's method in the more familiar Euclidean setting.

12.3 *Prelude: Newton's method in $\mathbb{R}$*

Numerical algorithms treat problems by breaking them down into easier subproblems. This pattern is ubiquitous. A timestepping algorithm for an initial value problem solves for the solution of an ODE by solving a sequence of nonlinear problems. An optimisation algorithm breaks down a minimisation problem into the computation of descent directions. In this section, we will see that Newton's method *solves a nonlinear problem by successively computing the roots of linearised approximations*.

Let $f \in C^1(\mathbb{R}; \mathbb{R})$, and consider figure 12.1. Our initial guess for a root of f is x_0, but $|f(x_0)| \gg 0$, and we would like to improve it.

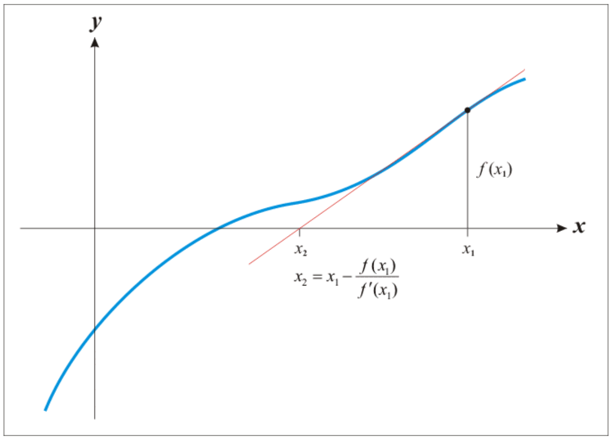

Figure 12.1: Each iteration of Newton's method consists of linearising the nonlinear problem and solving for the root of the linearised problem. Credit: Ralf Pfeifer, Wikipedia.

While we may not know how to find the roots of a nonlinear equation, we can certainly find the roots of a linear equation, and so we form the best linear approximation to f around x_0 by linearising the function at x_0, to yield the red function. We then solve for the (unique) root of the linearised function, which yields our next iterate x_1. We then linearise around x_1 and carry on.

It is a straightforward exercise of geometry to derive the following equation for the update $\delta x = x_{n+1} - x_n$:

$$f'(x_n)\delta x = -f(x_n). \tag{12.3.1}$$

Remark 12.3.1. *Note that if we had a solution, i.e. x such that $f(x) = 0$, then the right-hand side of our Newton update (12.3.1) would be zero, and thus $\delta x = 0$. That is, roots of f are fixed points of the iteration.*

Remark 12.3.2. *We require that $f'(x)$ is invertible (in one dimension, $f'(x) \neq 0$) for every iterate. If $f'(x)$ is not invertible, then the Newton step is not defined.*

Remark 12.3.3 (poor global convergence)**.** *Newton's method only converges* locally*, i.e. with sufficiently accurate initial guess. With poor initial guesses, Newton's method may diverge to infinity, or may get stuck in a cycle. For example, consider the application of Newton's method to*

$$f(x) = x^3 - 2x + 2 \tag{12.3.2}$$

from $x_0 = 0$. Computing (12.3.1), we find $x_1 = 1$. However, the Newton step from $x_1 = 1$ takes us back to $x_2 = 0$, and the 2-cycle repeats indefinitely.

Remark 12.3.4 (excellent local convergence)**.** *Newton's method can exhibit extremely fast* local *convergence under the right conditions. If the function f is C^2 and the root is isolated (i.e. is unique within a ball and of multiplicity one), then there exists a ball within which Newton converges* *quadratically. Let x^* be the root to which we converge. Quadratic convergence states that there exists a $\mu > 0$ such that*

$$\lim_{n\to\infty} \frac{|x^* - x_{n+1}|}{|x^* - x_n|^2} = \mu. \tag{12.3.3}$$

This is extremely fast; roughly speaking, the number of correct digits will double at every iteration. Once Newton gets the scent of the solution it zeroes in to machine accuracy in a handful of iterations.

Remark 12.3.5 (degenerate local convergence)**.** *If the root is degenerate (i.e. $f'(x^*) = 0$), then Newton's method will only converge linearly. This can happen if the solution is not isolated, or if it has multiplicity higher than one.*

Remark 12.3.6 (undecidability)**.** *It is a deep fact (proven by Fields medallist S. Smale and coauthors[6]) that Newton's method is* undecidable, *i.e. it is impossible in general to know in advance whether Newton's method will converge from an arbitrary initial guess without computing it. Since we always have to put an upper bound on the number of Newton iterations we are willing to do in practice, this has the unfortunate consequence that we never know if we would have converged had we been a little more patient.*

[6] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer-Verlag, 1998

12.4 Prelude: Newton's method in $\mathbb{R}^N$

The geometric reasoning of the previous section is hard to generalise to higher dimensions. Let us consider an alternative derivation of the iteration (12.3.1). Consider the Taylor expansion of f around x_n:

$$f(x_n + \delta x) \approx f(x_n) + f'(x_n)\delta x + \cdots . \tag{12.4.1}$$

We linearise the model by ignoring the higher-order terms,

$$f(x_n + \delta x) \approx f(x_n) + f'(x_n)\delta x, \tag{12.4.2}$$

and we *solve for the δx that brings our linearised model to zero*:

$$0 = f(x_n) + f'(x_n)\delta x \tag{12.4.3}$$

$$\implies f'(x_n)\delta x = -f(x_n), \tag{12.4.4}$$

exactly iteration (12.3.1).

If we apply this same argument to a function $F \in C^2(\mathbb{R}^N;\mathbb{R}^N)$, we find that the linearised Taylor expansion is

$$F(x_n + \delta x) \approx F(x_n) + J(x_n)\delta x, \tag{12.4.5}$$

and so the Newton update satisfies the linearised system

$$J(x_n)\delta x = -F(x_n), \tag{12.4.6}$$

where $J(x_n)$ is the Jacobian matrix evaluated at the guess x_n.

Thus, to apply Newton's method to a vector equation (such as our discretised nonlinear problem, (12.2.2)), we must solve linear systems involving the Jacobian matrix. This will require a matrix factorisation or a Krylov method, as discussed in section 9.5.

All of the remarks of the previous section still apply, with one important addition.

Remark 12.4.1 (Affine covariance)**.** *Newton's method is* affine-covariant*: given any nonsingular linear operator $A \in \mathbb{R}^{N\times N}$, Newton's method applied to $AF : \mathbb{R}^N \to \mathbb{R}^N$ yields* exactly the same sequence *of iterates (x_n).*

This property has deep consequences for the analysis and understanding of Newton's method, explored in detail in the book of Deuflhard[7]. A key corollary of this fact is that *measuring the residual norm $\|F(x_n)\|$ is not a good way of measuring convergence*, since by choosing a clever A we can get whatever residual we like, and yet the error in our approximation is unchanged.

[7] P. Deuflhard. *Newton Methods for Nonlinear Problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer-Verlag, 2011

The convergence of Newton's method is excellent when the iteration is within striking distance of an isolated root, but its global convergence is erratic. This can be visualised in the idea of a *Newton fractal*.

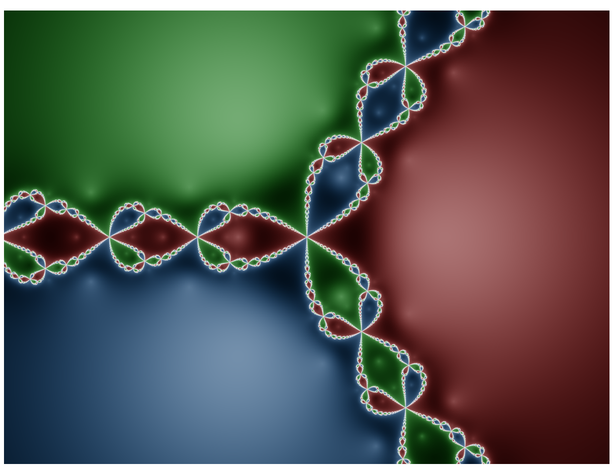

Figure 12.2: The Newton fractal for the function $f(z) = z^3 - 1$. Each pixel is coloured according to its basin of attraction, and is shaded according to how many iterations it took to converge there. White points did not converge within 256 iterations. Credit: Wikipedia.

Consider the nonlinear residual

$$\begin{aligned} f &: \mathbb{C} \to \mathbb{C} \\ f(z) &= z^3 - 1. \end{aligned} \tag{12.4.7}$$

Solutions of $f(z) = 0$ are exactly the cube roots of unity. Fix a colour for each root (e.g. $1 \mapsto$ red, $-\frac{1}{2} + \frac{\sqrt{3}}{2}i \mapsto$ green, $-\frac{1}{2} - \frac{\sqrt{3}}{2}i \mapsto$ blue). This problem can be considered equivalently as a residual mapping

$f : \mathbb{R}^2 \to \mathbb{R}^2$. Consider a rectangular section $R = [a, b] \times [a, b]$ of the plane, and for each $x \in R$ apply Newton's method to f. Colour the point according to the root found; colour the point white if Newton's method failed to converge after a fixed number of iterations.

The resulting image, displayed in figure 12.2, hints at the breathtaking complexity of the structures underpinning the convergence of Newton's method. Large regions of the plane are coloured red, green or blue, but between these boring regions there are beautiful fractal shapes, where the smallest perturbation causes you to converge to an entirely different solution, and where the set exhibits self-similarity: zooming in on figure 12.2 yields another figure very much like it[8].

[8] For much more details on Newton fractals, see `http://www.chiark.greenend.org.uk/~sgtatham/newton/`.

12.5 *The Newton–Kantorovich algorithm in Banach spaces*

The proceeding sections dealt with Newton's method in finite-dimensional spaces. We now consider the infinite-dimensional problem, i.e. *before* discretisation[9]. The extension of Newton's method to this context is referred to as the Newton–Kantorovich algorithm, after Kantorovich, who proved (a variant of) the theorem below in the 1940s[10,11].

[9] We have seen in sections 6.5 and 9.5 that disaster can ensue if we discretise and then blindly solve the resulting discrete system. Always postpone discretisation to the last possible moment!

[10] L. Kantorovich. On Newton's method for functional equations. *Doklady Akademii Nauk SSSR*, 59:1237–1249, 1948

[11] Kantorovich was a very interesting man. He independently invented linear programming, a fundamental branch of mathematical optimisation, some time before Dantzig. He did this in response to a problem brought to him by his comrades at the Leningrad Plywood Trust. Unfortunately, his theory of dual variables as shadow prices was insufficiently Marxist and Kantorovich narrowly avoided a trip to the Gulag.
During world war two, Kantorovich was a professor at the military college in Leningrad. During the war, Leningrad was besieged by Axis forces for nearly three years, with over one million deaths. Many more millions would have died had it not been for Kantorovich, whose precise calculations underpinned the only route for supplying the city with food via a road across a frozen lake.
After the war, he solved an optimisation problem to increase the efficiency of the Egorov railroad car-plant, minimising the amount of scrap metal wastage. The unfortunate consequence of this was that it disrupted the supply of scrap iron to the steel mills, and Kantorovich only escaped punishment by the regional Party because of his work on atomic reactors.
He also made fundamental contributions to many other areas of mathematics, computation and economics.

Kantorovich's theorem is a triumph: it is both the fundament of the numerical solution of nonlinear equations, but it is also a fundamental theorem of nonlinear functional analysis. This is because it *does not assume the existence of a solution*; it *proves* the existence of a solution, given certain conditions on H and its Fréchet derivative[12] verified around the initial guess x_0. Indeed, it even goes further: it proves *local uniqueness*, i.e. uniqueness within a neighbourhood of the root of H. Thus, with finite computations (and excellent insight in choosing the right x_0), it is possible in principle to prove the existence of solutions to very general equations.

[12] Defined in section 6.1.

For a self-contained, elegant and clear exposition of the proof of the Newton–Kantorovich theorem, see Ciarlet and Mardare[13].

[13] P. G. Ciarlet and C. Mardare. On the Newton–Kantorovich theorem. *Analysis and Applications*, 10(3):249–269, 2012

Theorem 12.5.1 (Newton–Kantorovich)**.** *Let X and Y be two Banach spaces. Let Ω be an open subset of X, the set where the residual is defined. Let $H \in C^1(\Omega, Y)$ be the residual of our nonlinear problem, and let $x_0 \in \Omega$ be an initial guess such that the Fréchet derivative $H'(x_0)$ is invertible (hence $H'(x_0) \in L(X; Y)$ and $H'(x_0)^{-1} \in L(Y; X)$). Let $B(x_0, r)$ denote the open ball of radius r centred at x_0. Assume that there exists a constant $r > 0$ such that*

(1) $\overline{B(x_0, r)} \subset \Omega$,

(2) $\|H'(x_0)^{-1} H(x_0)\|_X \leq \frac{r}{2}$,

(3) For all $\tilde{x}, x \in B(x_0, r)$,

$$\|H'(x_0)^{-1}\left(H'(\tilde{x}) - H'(x)\right)\|_{L(X;X)} \leq \frac{1}{r}\|\tilde{x} - x\|_X. \tag{12.5.1}$$

Then

(1) $H'(x) \in L(X;Y)$ *is invertible at each* $x \in B(x_0, r)$.

(2) The Newton sequence $(x_n)_{n=0}^{\infty}$ *defined by*

$$x_{n+1} = x_n - H'(x_n)^{-1}H(x_n) \tag{12.5.2}$$

is such that $x_n \in B(x_0, r)$ *for all* $n \geq 0$ *and converges to a root* $x^* \in \overline{B(x_0, r)}$ *of* H.

(3) For each $n \geq 0$,

$$\|x^* - x_n\|_X \leq \frac{r}{2^n}. \tag{12.5.3}$$

(4) The root x^* *is the locally unique, i.e.* x^* *is the only root of* H *in the ball* $\overline{B(x_0, r)}$.

Remark 12.5.2. *The proof is reasonably straightforward, but beautiful; it is outside the scope of this course, but I recommend you read it if you are keen.*

Remark 12.5.3. *This is a slightly simplified variant, corresponding to theorem 5 of Ciarlet. There are more complicated versions that expand the single constant* r *into distinct constants and so achieve a finer analysis. In particular, this expansion is required to prove locally superlinear convergence; the theorem above only claims* linear *convergence in conclusion (3).*

Remark 12.5.4. *Notice that the conditions are all affine-covariant, i.e. if we replace* $H \mapsto AH$ *for a nonsingular* $A \in L(Y, Y)$ *the bounds are unchanged.*

12.6 Example: the Bratu–Gelfand equation

Let us see how this works in practice. Consider again our model problem

$$u''(x) + \lambda e^u = 0, \quad u(0) = 0 = u(1). \tag{12.6.1}$$

and its variational formulation: find $u \in H_0^1(0, 1)$ such that

$$G(u; v) = -\int_0^1 u'(x)v'(x)\,\mathrm{d}x + \int_0^1 \lambda e^u v\,\mathrm{d}x = 0 \tag{12.6.2}$$

for all $v \in H_0^1(0, 1)$. Taking the Fréchet derivative of G with respect to u, we find that the linearisation at a fixed u in the direction $w \in H_0^1(0, 1)$ is

$$G_u(u; v, w) = -\int_0^1 w'(x)v'(x)\,\mathrm{d}x + \int_0^1 \lambda e^u wv\,\mathrm{d}x. \tag{12.6.3}$$

Thus, the Newton update solves: find $\delta u \in H_0^1(0,1)$ such that

$$-\int_0^1 \delta u'(x) v'(x) \, \mathrm{d}x + \int_0^1 \lambda e^u \delta u v \, \mathrm{d}x = \int_0^1 u'(x) v'(x) \, \mathrm{d}x - \int_0^1 \lambda e^u v \, \mathrm{d}x \tag{12.6.4}$$

for all $v \in H_0^1(0,1)$. Once δu is computed, we update $u \leftarrow u + \delta u$ and repeat. Assuming $\delta u \in H^2(0,1) \cap H_0^1(0,1)$, the equation for the Newton update can be written in strong form as

$$\delta u'' + \lambda e^u \delta u = -u'' - \lambda e^u, \quad \delta u(0) = 0 = \delta u(1). \tag{12.6.5}$$

We have thus resolved the problem of solving the infinite-dimensional nonlinear problem into a sequence of infinite-dimensional linear problems. These linear problems may then be discretised using our standard finite element techniques.

The following code uses FEniCS[14] to approximately solve the Bratu problem on a mesh of 500 elements for $\lambda = 2$, starting from two initial guesses: $u_0(x) = 0$, and $u_0(x) = 3$[15].

[14] A. Logg, K. A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2011

[15] A sharp reader will notice that this initial guess does not satisfy the boundary conditions. A robust code deals with this by ensuring that the first Newton update has boundary conditions to correct the values of the solution on the boundary.

```
from dolfin import *

mesh = UnitIntervalMesh(500)

element = FiniteElement("Lagrange", interval, 1)
V = FunctionSpace(mesh, element)

u = Function(V)
v = TestFunction(V)

lmbda = Constant(2.0)

F = -inner(grad(u), grad(v))*dx + lmbda*exp(u)*v*dx
bc = DirichletBC(V, 0, DomainBoundary())

for u_init in [0, 3]:
    u.interpolate(Constant(u_init))
    solve(F == 0, u, bc) # apply Newton-Kantorovich
    plot(u, interactive=True)
```

The two Newton iterations converge to the two distinct solutions of the problem, shown in figure 12.3.

12.7 Further questions

Many important questions remain.

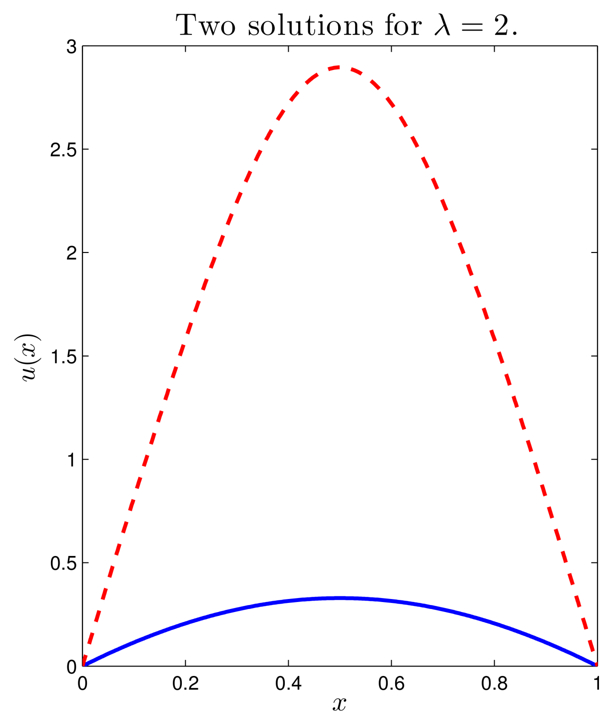

Figure 12.3: Two solutions of the Bratu–Gelfand equation for $\lambda = 2$.

First, are the linearised problems well-posed? While we may accept that the underlying nonlinear problem supports distinct solutions, we must demand that the linearised problem have a unique solution at every step if the algorithm is to make sense. In this case, it is not at all clear that the variational equation (12.6.4) satisfies the conditions of the Lax–Milgram Theorem, since the bilinear form $a(\delta u, v)$ consists of the addition of a coercive form (as $e^u(x) > 0$) to an anticoercive one (i.e. one whose negation is coercive). In fact, the Lax–Milgram conditions are merely sufficient to prove well-posedness, and not necessary; in a subsequent lecture we will develop a more general theory of necessary and sufficient conditions for the well-posedness of linear variational problems, which will concisely resolve this question.

Second, we will never be able to compute the exact solution of the Newton step (12.6.4); at best we will compute an approximation δu_h with some control over the error $\delta u - \delta u_h$ in a suitable norm. This naturally leads to the concept of *inexact* Newton methods, where the Newton update is not solved for exactly, but only up to some residual. By cleverly adapting the residual tolerance of the inner solver, one can guarantee the retention of locally superlinear or quadratic convergence of the Newton method[16]. In fact, this development is also necessary even when the nonlinear PDE is discretised first, as

[16] S. Eisenstat and H. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*, 17(1):16–32, 1996

the discrete Newton step (12.4.6) will only be solved to within the accuracy of the machine and the tolerance of any Krylov solver used.

Third, once we have a scheme for deciding how accurate the solution of each Newton step must be, we must adapt the discretisation of (12.6.4) to achieve it. This naturally leads to the subject of *a posteriori* error estimation: given a computed approximation δu_h, how can we compute an upper bound for the specific error $\|\delta u - \delta u_h\|$, and how can we refine our mesh to achieve it? There is a very large body of literature on this topic; for an introduction to the field, see Ainsworth & Oden[17].

[17] M. T. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, New York, 2000

Fourth, Newton's major downfall is its lack of global robustness, i.e. far from a solution it may fail to converge. Here we offer two suggestions. First, just as in nonlinear optimisation, the algorithm may be globalised by the addition of a *line search*: after the Newton update δu has been computed, instead of executing

$$u_{n+1} = u_n + \delta u, \tag{12.7.1}$$

the iteration proceeds by

$$u_{n+1} = u_n + \mu\delta u, \tag{12.7.2}$$

where μ is to be specified. Different globalisation schemes choose μ differently, but the overall theme is to take small steps ($\mu \ll 1$) far from a solution to improve robustness, while to take full steps ($\mu = 1$) close to a solution to exploit the natural quadratic convergence of the method. The concept of affine covariance introduced in remark 12.4.1 is crucial for the design of sensible globalisation schemes, and is discussed in detail in Deuflhard[18].

[18] P. Deuflhard. *Newton Methods for Nonlinear Problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer-Verlag, 2011

A second, and often more robust, approach to globalising Newton's method is to solve a *sequence* of nonlinear problems, starting from a very easy one to which we know the answers and culminating in the problem whose solutions we seek. This can be interpreted as the computation of the *bifurcation diagram* of

$$F : V \times \mathbb{R}, \tag{12.7.3}$$

i.e. understanding the solutions $u(\lambda)$ of $F(u, \lambda) = 0$ as a scalar real parameter λ is varied. This is the subject of bifurcation analysis and is an exciting field of functional and numerical analysis in its own right.

13 *Noncoercive variational problems*

In lecture 4, we proved the Lax–Milgram theorem about the well-posedness (existence, uniqueness and stability) of *coercive* problems. Recall that in the finite-dimensional case, coercivity is equivalent to positive-definiteness of the matrix. Since positive-definiteness is merely a sufficient condition for invertibility of a matrix (not a necessary one), we might analogously expect coercivity to be merely sufficient for the well-posedness of a linear variational equality. This is indeed the case: such a problem need not be coercive for it to be well-posed. It turns out that a weaker set of conditions, the *Babuška–Brezzi* conditions, are both sufficient and *necessary*.

This subject is not well explained in most books on finite elements (and functional analysis). In most expositions, the conditions parachute in from the sky, and are used to prove well-posedness of the problem without much in the way of motivation for their origins. In this lecture we will take a slower pace, building intuition from the finite-dimensional case, and explaining how the Babuška–Brezzi conditions arise naturally from considerations of the stability of the solution of a finite-dimensional system to changes in the right-hand side. The exposition is adapted from that of Brezzi and Bathe[1].

A note on the history and naming of these conditions is in order. Babuška stated the conditions in generality in 1971[2], but they had been developed before in the context of a specific application by Ladyzhenskaya in 1969[3], and for this reason are sometimes referred to as the Ladyzhenskaya–Babuška or Ladyzhenskaya–Babuška–Brezzi conditions. In fact, they had already been given in general form by Nečas in 1962[4], but the result did not draw widespread attention. Some authors such as Ern and Guermond[5] refer to them as the Banach–Nečas–Babuška conditions, adding Banach because the theorem follows from combining two theorems of Banach's; other authors refer to them as the generalised Lax–Milgram conditions. For reasons that will become obvious by the end of the lecture, the conditions are also commonly referred to as the inf-sup conditions.

Brezzi's contributions to the subject were twofold[6]. First, he specialised the conditions to the important case of *saddle-point* systems,

[1] F. Brezzi and K.-J. Bathe. A discourse on the stability conditions for mixed finite element formulations. *Computer Methods in Applied Mechanics and Engineering*, 82(1):27–57, 1990

[2] I. Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16(4):322–333, 1971

[3] O. A. Ladyzhenskaya. *The Mathematical Theory of Viscous Flows*. Gordon and Breach, 1969

[4] J. Nečas. Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, 16(4):305–326, 1962

[5] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer, 2004

[6] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 8(R2):129–151, 1974

where we are minimising an energy functional subject to a constraint. For example, in the Stokes equations of fluid mechanics, we solve for the velocity that minimises the Dirichlet energy subject to the incompressibility constraint that $\nabla \cdot u = 0$ (as described more thoroughly in lecture 14). Second, he was the first to prove the *necessity* of the conditions.

The weaker nature of the Babuška–Brezzi conditions have an important consequence for finite element discretisations. For coercive problems, any Galerkin method automatically inherits the well-posedness of the infinite-dimensional problem. For noncoercive problems, this is no longer true. The well-posedness of a given choice of discretisation depends on its satisfaction of the analogous inf-sup condition for the particular choices of function spaces used, and must be proven on a case-by-case basis.

13.1 Prelude: the dual norm

We recall the definition of the dual space and its norm (given in definition 2.3.5.) The *dual* V^* of a Hilbert space V is the space of all bounded linear functionals on V. This has a natural norm induced by the norm on the underlying space:

$$\|j\|_{V^*} = \sup_{\|u\|_V=1} |j(u)| = \sup_{\substack{u\in V\\ u\neq 0}} \frac{|j(u)|}{\|u\|_V}. \tag{13.1.1}$$

13.2 The stability of finite-dimensional linear systems

Consider the N-dimensional linear system

$$\text{find } x \in \mathbb{R}^N \text{ such that } Mx = b, \tag{13.2.1}$$

arising from a Galerkin discretisation of

$$\text{find } u \in V \text{ such that } a(u, v) = F(v) \text{ for all } v \in V. \tag{13.2.2}$$

We know from linear algebra that the linear system $Mx = b$ has a unique solution for all $b \in \mathbb{R}^N$ if and only if the associated homogeneous problem $Mx = 0$ has only one solution, $x = 0$. Let us suppose for now that this is the case.

Now suppose that we wish to understand how much a perturbation to the right-hand side δb can affect the solution. Denote by δx the corresponding change in the solution. For the system to be stable, we have the notion that a small change in b will induce only a small change in x. Therefore, we must introduce norms for the right-hand side and solution spaces. From the perspective of finite-dimensional linear algebra, it would be unusual to consider different norms for b and x, but we know better; the solution will live in some Hilbert space V and the right-hand side will live in its dual V^*. We therefore equip x and b with the norms $\|\cdot\|_V$ and $\|\cdot\|_{V^*}$ respectively.

Definition 13.2.1 (Stability constant)**.** *The stability constant of M with respect to the norms $\|\cdot\|_V$, $\|\cdot\|_{V^*}$ is the smallest possible constant S such that*

$$\frac{\|\delta x\|_V}{\|x\|_V} \leq S \frac{\|\delta b\|_{V^*}}{\|b\|_{V^*}} \tag{13.2.3}$$

for all vectors x and δx in $\mathbb{R}^N$ such that $Mx = b$ and $M\delta x = \delta b$.

Such a stability constant always exists if the matrix is invertible. However, if we consider a *sequence* of linear systems with increasing dimension N (corresponding to a finer and finer discretisation) it might be the case that the associated constants (S) depend on N and become infinitely large as $N \to \infty$. We thus say that a sequence of linear systems is stable with respect to the norms $\|\cdot\|_V$, $\|\cdot\|_{V^*}$ if the sequence of stability constants is uniformly bounded.

We can use matrix norms to clarify the nature of the stability constant. Define

$$\|M\| = \sup_{\substack{y \in V \\ y \neq 0}} \frac{\|My\|_{V^*}}{\|y\|_V}, \tag{13.2.4}$$

where we denote the input space by V. Choosing $y = x$, $My = b$, we have

$$\|M\| \geq \frac{\|b\|_{V^*}}{\|x\|_V}. \tag{13.2.5}$$

This implies that

$$\|M\| \frac{\|x\|_V}{\|b\|_{V^*}} \geq 1. \tag{13.2.6}$$

Now let us consider the inverse norm (assuming, again, that the inverse exists). We have

$$\|M^{-1}\| = \sup_{\substack{z \in V^* \\ z \neq 0}} \frac{\|M^{-1}z\|_V}{\|z\|_{V^*}}, \tag{13.2.7}$$

where we denote the output space by V^*. By choosing $z = \delta b$, $M^{-1}z = \delta x$, we have

$$\|M^{-1}\| \geq \frac{\|\delta x\|_V}{\|\delta b\|_{V^*}}. \tag{13.2.8}$$

Multiplying the upper bound by a quantity greater than one will not change the inequality, so

$$\frac{\|\delta x\|_V}{\|\delta b\|_{V^*}} \leq \|M^{-1}\| \|M\| \frac{\|x\|_V}{\|b\|_{V^*}} \tag{13.2.9}$$

which implies

$$\frac{\|\delta x\|_V}{\|x\|_V} \leq \|M\| \|M^{-1}\| \frac{\|\delta b\|_{V^*}}{\|b\|_{V^*}}. \tag{13.2.10}$$

Since x and δx are arbitrary, $\|M\| \|M^{-1}\|$ is the smallest number for which this holds, and so

$$S = \|M\| \|M^{-1}\|. \tag{13.2.11}$$

Remark 13.2.2 (Condition number)**.** *This is exactly the condition number of the matrix. Typically this is specialised to the Euclidean case* $\| \cdot \|_V = \| \cdot \|_{V^*} = \| \cdot \|_{\ell_2}$*, where the condition number resolves to the ratio of the largest and smallest singular values.*

Thus, for the stability of our problem to be uniformly bounded, we will require both $\|M\|$ and $\|M^{-1}\|$ to be uniformly bounded from above. We will now consider them in turn.

13.3 *The forward operator norm*

Let us assume that the bilinear form a is bounded: that is, there exists a constant C such that

$$|a(u, v)| \leq C \|u\| \|v\|, \tag{13.3.1}$$

and since our matrix encodes the action of the bilinear form, we have

$$|y^T M x| \leq C \|y\|_V \|x\|_V. \tag{13.3.2}$$

The forward operator norm $\|M\|$ is exactly the boundness constant C in disguise. To see this, expand the definitions:

$$\|M\| = \sup_{\substack{x \in V \\ x \neq 0}} \frac{\|Mx\|_{V^*}}{\|x\|_V} \tag{13.3.3}$$

$$= \sup_{\substack{x \in V \\ x \neq 0}} \left\{ \frac{1}{\|x\|_V} \sup_{\substack{y \in V \\ y \neq 0}} \frac{|y^T M x|}{\|y\|_V} \right\} \tag{13.3.4}$$

$$= \sup_{\substack{x, y \in V \\ x, y \neq 0}} \frac{|y^T M x|}{\|x\|_V \|y\|_V} = C. \tag{13.3.5}$$

Thus, if M arises from a conforming discretisation of a continuous bilinear form, then $\|M\|$ is uniformly bounded above by the continuity constant of the form. Thus, the sequence of problems will be stable if and only if the inverse operator norm $\|M^{-1}\|$ is uniformly bounded above.

13.4 The inverse operator norm

Lemma 13.4.1 (Characterising the inverse operator norm). *Let $M \in \mathbb{R}^{N\times N}$ be nonsingular, and let $\|\cdot\|_V$ be the norm for its input space, and let the associated dual norm be used for its output space. Then*

$$\|M^{-1}\|^{-1} = \inf_{\substack{x\in V\\ x\neq 0}} \sup_{\substack{y\in V\\ y\neq 0}} \frac{|y^T M x|}{\|y\|_V \|x\|_V}. \tag{13.4.1}$$

Proof.

$$\begin{aligned}
\|M^{-1}\|^{-1} &= \left(\sup_{\substack{z\in V^*\\ z\neq 0}} \frac{\|M^{-1}z\|_V}{\|z\|_{V^*}} \right)^{-1} && (13.4.2)\\
&= \inf_{\substack{z\in V^*\\ z\neq 0}} \frac{\|z\|_{V^*}}{\|M^{-1}z\|_V} && (13.4.3)\\
&= \inf_{\substack{x\in V\\ x\neq 0}} \frac{\|Mx\|_{V^*}}{\|x\|_V} \qquad \text{(set } z = Mx\text{)} && (13.4.4)\\
&= \inf_{\substack{x\in V\\ x\neq 0}} \left\{ \frac{1}{\|x\|_V} \sup_{\substack{y\in V\\ y\neq 0}} \frac{|y^T M x|}{\|y\|_V} \right\} \text{(defn of dual norm)} && (13.4.5)\\
&= \inf_{\substack{x\in V\\ x\neq 0}} \sup_{\substack{y\in V\\ y\neq 0}} \frac{|y^T M x|}{\|y\|_V \|x\|_V}. && (13.4.6)
\end{aligned}$$

□

Remark 13.4.2. *Since*

$$\sup_{\substack{y\in V\\ y\neq 0}} \frac{|y^T M x|}{\|y\|_V} = \sup_{\substack{y\in V\\ y\neq 0}} \frac{y^T M x}{\|y\|_V}, \tag{13.4.7}$$

we can if we wish ignore the absolute value sign on the numerator. (If the supremum were achieved at a negative argument to the absolute value function, we could just consider $y \mapsto -y$ as V is a vector space, and swap the sign of the numerator without changing the denominator.) It is a matter of taste whether we drop the minus sign, but many expositions do, so we do that here.

Thus, for the sequence of problems to be stable, we need that $\|M^{-1}\|^{-1}$ to be uniformly bounded below, and so we require a constant $\gamma \in \mathbb{R}$ such that

$$\inf_{\substack{x\in V\\ x\neq 0}} \sup_{\substack{y\in V\\ y\neq 0}} \frac{y^T M x}{\|y\|_V \|x\|_V} \geq \gamma > 0. \tag{13.4.8}$$

Since the matrix M encodes the bilinear form, the analogous condition for the infinite-dimensional problem is: there exists $\gamma \in \mathbb{R}$ such that

$$\inf_{\substack{u \in V \\ u \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u,v)}{\|u\|_V \|v\|_V} \geq \gamma > 0. \tag{13.4.9}$$

This is the famous "inf-sup" condition of Babuška and Brezzi.

13.5 *The inf-sup condition and the kernel*

We started our analysis by assuming that M was nonsingular. Of course, we now need to ensure that the inf-sup condition (derived based on stability arguments) does indeed imply that the kernel of the operator is trivial in the first place[7]. The argument, in finite dimensions, is really just unwinding that of lemma 13.4.1.

[7] For square finite dimensional systems, this is the same thing as invertibility.

The inf-sup condition (13.4.8) can be restated as: there exists $\gamma > 0$ such that for all $0 \neq x \in V$

$$\sup_{\substack{y \in V \\ y \neq 0}} \frac{y^T M x}{\|y\|_V} \geq \gamma \|x\|_V. \tag{13.5.1}$$

We recognise the quantity on the left-hand side as the dual norm of Mx, so this in turn is equivalent to: there exists $\gamma > 0$ such that for all $0 \neq x \in V$

$$\|Mx\|_{V^*} \geq \gamma \|x\|_V. \tag{13.5.2}$$

Suppose the kernel of M was nontrivial, i.e. there exists $x \neq 0$ such that $Mx = 0$. Since $x \neq 0$, our inf-sup inequality holds, and

$$Mx = 0 \implies \|Mx\|_{V^*} = 0 \implies \|x\|_V = 0 \implies x = 0, \tag{13.5.3}$$

a contradiction. Thus the inf-sup condition exactly implies the condition on the kernel of M that we sought.

If we apply the discrete inf-sup condition (13.4.8) with *Euclidean* norms chosen for V and V^*, the inf-sup condition has a very intuitive interpretation. A straightforward calculation shows that the inf-sup constant is given by

$$\gamma = \lambda_{\min}(M^T M)^{1/2}, \tag{13.5.4}$$

i.e. the inf-sup constant is nothing other than the *smallest singular value* of the matrix M. No wonder it must be bounded away from zero!

13.6 *The inf-sup condition and coercivity*

The inf-sup condition (13.4.9) is a generalisation of the notion of coercivity. Recall that a bilinear form $a : V \times V \to \mathbb{R}$ is coercive if there exists a constant $\alpha > 0$ such that

$$\alpha \|u\|_V^2 \leq a(u,u) \tag{13.6.1}$$

for all $u \in V$. Dividing both sides by $\|u\|_V \neq 0$, we find

$$\alpha \|u\|_V \leq \frac{a(u,u)}{\|u\|_V} \leq \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u,v)}{\|v\|_V} \text{ for all } u \in V, u \neq 0, \tag{13.6.2}$$

and infimising over u we conclude that coercivity of a implies the existence of α such that

$$0 < \alpha \leq \inf_{\substack{u \in V \\ u \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u,v)}{\|u\|_V \|v\|_V}, \tag{13.6.3}$$

that is, the coercivity constant provides an inf-sup constant. Coercivity implies the inf-sup condition, but not the other way around.

13.7 *The inf-sup condition and necessity*

The inf-sup condition is the key ingredient of the most general set of conditions that guarantee well-posedness. This is because it is *necessary*, i.e. it is implied by the well-posedness of the problem. If

$$\text{find } u \in V \text{ such that } a(u,v) = F(v) \text{ for all } v \in V \tag{13.7.1}$$

has a unique solution for all $F \in V^*$, that implies the unique solvability of the associated operator equation

$$Au = F \tag{13.7.2}$$

where $A : V \to V^*$ is defined by $(Au)(v) = a(u,v)$. If the operator equation is uniquely solvable, then $A^{-1} : V^* \to V$ exists, and the stability of the problem implies its norm is bounded above, so the reciprocal of its norm is bounded below away from zero. Untangling the definitions of the operator norm (reversing the argument of lemma 13.4.1), we find

$$0 < \gamma = \|A^{-1}\|^{-1} = \inf_{\substack{u \in V \\ u \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u,v)}{\|u\|_V \|v\|_V}, \tag{13.7.3}$$

exactly the inf-sup condition[8].

[8] Coercivity is necessary in a special case. If the bilinear form a is symmetric and monotone (i.e. $a(v,v) \geq 0$ for all $v \in V$), then coercivity is necessary and sufficient for well-posedness.

13.8 *Rectangular linear systems: the transpose condition*

For square linear systems, invertibility of the matrix M is equivalent to its kernel being trivial, and to the invertibility of its transpose

M^T. It will be useful in what follows to consider nonsquare linear systems: this will motivate the third and final condition that will guarantee well-posedness.

Consider a rectangular matrix $M \in \mathbb{R}^{P \times N}$. We denote by V_1 the input space (in this case $\mathbb{R}^N$) with associated norm $\|\cdot\|_{V_1}$. Similarly, we denote by V_2 the output space (in this case $\mathbb{R}^P$) with associated norm $\|\cdot\|_{V_2}$. In this case, the inf-sup condition (13.4.8) reads: there exists $\gamma \in \mathbb{R}$ such that

$$0 < \gamma \leq \inf_{\substack{x \in V_1 \\ x \neq 0}} \sup_{\substack{y \in V_2 \\ y \neq 0}} \frac{y^T M x}{\|y\|_{V_2} \|x\|_{V_1}}. \tag{13.8.1}$$

First, consider the case where M is an underdetermined matrix (more columns than rows, more variables than equations, $N > P$). Then the rank-nullity theorem guarantees us that the matrix has a nontrivial nullspace, so there are nonzero vectors $x \in V_1$ such that $Mx = 0$. Thus the discrete inf-sup condition (13.4.8) fails, as it should; we are happy that our conditions fail when this problem is not well-posed.

Next consider the case where M is an overdetermined matrix (more rows than columns, more equations than variables, $P > N$). Such a problem cannot be well-posed, for in this case the range of the matrix cannot be the whole of $\mathbb{R}^P$, and so the equation $Mx = b$ does not have a solution for arbitrary data b (only for those vectors b in the range). However, such a matrix might still have a trivial nullspace, and so can satisfy the inf-sup condition. As a concrete example, consider

$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}. \tag{13.8.2}$$

Then if $x = [x_1, x_2]^T$, $Mx = [x_1, x_2, x_1 + x_2]^T$, and $Mx = 0 \implies x = 0$. M satisfies the inf-sup condition (which encodes the fact that the nullspace is trivial), but this is not enough to guarantee well-posedness.

We thus need a third condition: that the *nullspace of the transpose* is trivial. That is, we also require that $y^T M = 0 \implies y = 0$. Recall that the fundamental theorem of linear algebra tells us that

$$\text{range}(M) = \text{kernel}(M^T)^{\perp}, \tag{13.8.3}$$

that is, the range of M is the orthogonal complement of the nullspace of M^T. Thus, in order for the operator to be surjective (and have a complete range), we must therefore require the nullspace of M^T to be trivial. This is the condition that fails in this case; for example, choose $y^T = [1, 0, -1]$.

Expressed variationally, the statement that the nullspace of M^T is trivial is equivalent to the following[9]: there exists a $\gamma' \in \mathbb{R}$ such that

$$0 < \gamma' \leq \inf_{\substack{y \in V_2 \\ y \neq 0}} \sup_{\substack{x \in V_1 \\ x \neq 0}} \frac{x^T M^T y}{\|x\|_{V_1} \|y\|_{V_2}} \tag{13.8.4}$$

$$= \inf_{\substack{y \in V_2 \\ y \neq 0}} \sup_{\substack{x \in V_1 \\ x \neq 0}} \frac{y^T M x}{\|x\|_{V_1} \|y\|_{V_2}}. \tag{13.8.5}$$

That is, the variables over which the inf and sup are taken are reversed.

No nonsquare matrix can satisfy both this condition (13.8.4) and (13.4.8). That is, no nonsquare matrix is invertible, exactly as we expect.

[9] Go through the argument of section 13.5 with $M \mapsto M^T$.

13.9 Babuška's Theorem

We are now in a position to state Babuška's theorem for the well-posedness of an abstract variational equation.

Theorem 13.9.1 (Babuška's theorem: necessary and sufficient conditions)**.** *Let V_1 and V_2 be two Hilbert spaces with inner products $(\cdot,\cdot)_{V_1}$ and $(\cdot,\cdot)_{V_2}$ respectively. Let $a : V_1 \times V_2 \to \mathbb{R}$ be a bilinear form for which there exist constants $C < \infty$, $\gamma > 0$, $\gamma' > 0$ such that*

(1) $|a(u,v)| \leq C\|u\|_{V_1}\|v\|_{V_2}$ *for all $u \in V_1, v \in V_2$;*

(2) $\gamma \leq \inf_{\substack{u \in V_1 \\ u \neq 0}} \sup_{\substack{v \in V_2 \\ v \neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$;

(3) $\gamma' \leq \inf_{\substack{v \in V_2 \\ v \neq 0}} \sup_{\substack{u \in V_1 \\ u \neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$.

Then for all $F \in V_2^*$ there exists exactly one element $u \in V_1$ such that

$$a(u,v) = F(v) \text{ for all } v \in V_2. \tag{13.9.1}$$

Furthermore the problem is stable in that

$$\|u\|_{V_1} \leq \frac{\|F\|_{V_2^*}}{\gamma}. \tag{13.9.2}$$

Proof. See theorem 2.1 of Babuška[10]. □

[10] I. Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16(4):322–333, 1971

Remark 13.9.2. *The first assumption states that the bilinear form is continuous. The second assumption implies that the nullspace of the operator $A : V_1 \to V_2^*$ arising in the associated operator equation*

$$Au = F, \quad (Au)(v) := a(u,v) \tag{13.9.3}$$

is trivial, which encodes the injectivity of A. The third assumption states that the nullspace of the adjoint operator $A' : V_2 \to V_1^*$ is trivial, which is equivalent to the surjectivity of A by the Banach closed range theorem.

Remark 13.9.3. *The third condition is often replaced by the following:*

$$\forall u \in V_1, a(u,v) = 0 \implies v = 0, \tag{13.9.4}$$

or

$$\forall\ 0 \neq v \in V_2\ \exists\ u \in V_1 \text{ such that } a(u,v) \neq 0, \tag{13.9.5}$$

or

$$\sup_{u \in V_1} |a(u,v)| > 0 \text{ for all } 0 \neq v \in V_2. \tag{13.9.6}$$

I state the version above as I like the symmetry between conditions (2) and (3).

Remark 13.9.4. *The first condition is often replaced by the following:*

$$\sup_{\substack{u \in V \\ u \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u,v)}{\|u\|_V \|v\|_V} \leq C, \tag{13.9.7}$$

in other words a "sup–sup" condition. In this form we can drop the absolute value function from the numerator for the same reason as in Remark 13.4.2. I prefer this form for symmetry with the other two conditions, but kept the original statement of continuity for familiarity.

Remark 13.9.5. *In fact, the Babuška conditions are even more general than this. They apply if the trial space V_1 is a Banach space, and the test space V_2 is a reflexive Banach space. (A reflexive Banach space is one where $V^{**} = V$.)*

Remark 13.9.6. *The generalisation to different spaces V_1 and V_2 is crucial. Consider a PDE with an odd number of derivatives, such as a first-order equation: no matter which way you integrate by parts, the test and trial spaces will be different. It will also be necessary in Brezzi's formulation of the conditions for the well-posedness of a mixed variational problem. It is also occasionally necessary in second-order scalar problems: for example, one of the ways to adapt Galerkin discretisations to advection-dominated problems is to modify the test functions used according to the direction of the advecting velocity. Such a discretisation is referred to as a* Petrov–Galerkin *discretisation, where the test and trial spaces are distinct*[11]*. The proof of well-posedness of the formulation thus requires a more general theory.*

[11] As opposed to a *Bubnov–Galerkin* discretisation, where the test and trial spaces are the same.

13.10 Quasioptimality for noncoercive problems

For coercive problems, Céa's Lemma tells us that a Galerkin discretisation is *quasi-optimal*: optimal in the approximation space up to problem-specific constants. We will now see that *stable Galerkin discretisations of well-posed problems are quasi-optimal*, i.e. the quasi-optimality property carries over to noncoercive well-posed problems. The key difference with the coercive case is that *the Galerkin discretisation is not automatically stable*: in general we will have to prove a nontrivial result to ensure that the discretisation satisfies a discrete version of the inf-sup conditions.

Consider a general variational problem[12]:

$$\text{find } u \in V \text{ such that } a(u, v) = F(v) \text{ for all } v \in V, \tag{13.10.1}$$

where a is bilinear and $F \in V^*$. Assume that this problem is well-posed, i.e. it satisfies the Babuška conditions. Given a closed subspace $V_h \subset V$, the Galerkin approximation is:

$$\text{find } u_h \in V_h \text{ such that } a(u_h, v_h) = F(v_h) \text{ for all } v_h \in V_h. \tag{13.10.2}$$

Note that the *Galerkin orthogonality* property discussed in section 7.2 still holds for noncoercive problems: for all $v_h \in V_h$,

$$a(u - u_h, v_h) = 0. \tag{13.10.3}$$

Let us consider whether the Galerkin approximation (13.10.2) is well-posed. The continuity requirement (1) of the Babuška conditions clearly follows from the corresponding condition for the infinite-dimensional problem. It therefore remains to check requirement (2), the inf-sup condition[13]. That is, we need to ensure that there exists $\tilde{\gamma}$

$$\inf_{\substack{u_h \in V_h \\ u_h \neq 0}} \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{a(u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \geq \tilde{\gamma} > 0, \tag{13.10.4}$$

with $\tilde{\gamma}$ independent of the mesh size h. Unfortunately this does *not* follow from the inf-sup condition for the infinite-dimensional problem; even if it is true, we might have $\tilde{\gamma} \neq \gamma$. We therefore have to *assume* this for the argument below.

Theorem 13.10.1 (Quasi-optimality of stable Galerkin discretisations)**.** *Consider the variational problem* (13.10.1) *and its Galerkin discretisation* (13.10.2)*. Assume that they satisfy the necessary and sufficient conditions for well-posedness given in theorem 13.9.1 and* (13.10.4)*. Then*

$$\|u - u_h\|_V \leq \left(1 + \frac{C}{\tilde{\gamma}}\right) \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{13.10.5}$$

[12] For simplicity, we pose the problem using the same test and trial spaces, but the same argument extends to the more general case of distinct Banach spaces (with the test space reflexive).

[13] In this case, we do not need to check requirement (3), the transpose inf-sup condition, because for square matrices in finite dimensions triviality of the nullspace is necessary and sufficient for well-posedness: we only need to check triviality of the transpose nullspace for rectangular matrices or infinite-dimensional problems.

Proof. For every $v_h \in V_h$, we have

$$
\begin{aligned}
\tilde{\gamma}\|v_h - u_h\|_V &\leq \sup_{\substack{w_h \in V_h \\ w_h \neq 0}} \frac{a(v_h - u_h, w_h)}{\|w_h\|_V} && \text{(discrete inf-sup)} \\
&= \sup_{\substack{w_h \in V_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h) + a(u - u_h, w_h)}{\|w_h\|_V} && \text{(bilinearity of } a\text{)} \\
&= \sup_{\substack{w_h \in V_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h)}{\|w_h\|_V} && \text{(Galerkin orthogonality)} \\
&\leq \sup_{\substack{w_h \in V_h \\ w_h \neq 0}} \frac{C\|v_h - u\|_V \|w_h\|_V}{\|w_h\|_V} && \text{(boundedness of } a\text{)} \\
&= C\|v_h - u\|_V. && (13.10.6)
\end{aligned}
$$

Now apply the triangle inequality to $\|u - u_h\|_V$:

$$
\begin{aligned}
\|u - u_h\|_V &\leq \|u - v_h\|_V + \|v_h - u_h\|_V && (13.10.7) \\
&\leq \|u - v_h\|_V + \frac{C}{\tilde{\gamma}}\|u - v_h\|_V && (13.10.8) \\
&= \left(1 + \frac{C}{\tilde{\gamma}}\right)\|u - v_h\|_V. && (13.10.9)
\end{aligned}
$$

□

As before, we can combine this with an approximation result and a regularity result to derive error estimates for finite element discretisations.

Remark 13.10.2. *For problems posed in Hilbert spaces, the quasi-optimality result* (13.10.5) *can be sharpened further to*

$$\|u - u_h\|_V \leq \frac{C}{\tilde{\gamma}} \inf_{u_h \in V_h} \|u - v_h\|_V, \qquad (13.10.10)$$

exactly the same form as Céa's Lemma. See Xu and Zikatanov[14] *for details.*

[14] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numerische Mathematik*, 94(1):195–202, 2003

14 Mixed finite element methods

All of the problems we have considered up to now have considered the approximation of a single physical quantity. However, it is often useful to consider mathematical models which involve several physically distinct quantities, which must be approximated simultaneously. In this lecture we will consider the variational formulation and discretisation of such equations.

This lecture draws on many sources, including the notes of Süli[1], the finite-dimensional exposition of Brezzi and Bathe[2], and chapter 12 of Brenner and Scott[3].

[1] E. Süli. A brief excursion into the mathematical theory of mixed finite element methods, 2017. http://people.maths.ox.ac.uk/suli/mixed_FEM_lectures.pdf

[2] F. Brezzi and K.-J. Bathe. A discourse on the stability conditions for mixed finite element formulations. *Computer Methods in Applied Mechanics and Engineering*, 82(1):27–57, 1990

[3] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag New York, third edition edition, 2008

14.1 *Example: the Stokes equations*

The Stokes equations are an elementary model in fluid mechanics. They describe the motion of a steady, incompressible, viscous, Newtonian, isothermal, slow-moving fluid. In strong form the equations are

$$-\nabla^2 u + \nabla p = f \quad \text{in } \Omega, \tag{14.1.1}$$

$$\nabla \cdot u = 0 \quad \text{in } \Omega, \tag{14.1.2}$$

subject to some boundary conditions to be specified. The first equation, the momentum equation, relates the vector-valued velocity $u : \Omega \to \mathbb{R}^n$ and the scalar-valued pressure $p : \Omega \to \mathbb{R}$. The vector-valued variable $f : \Omega \to \mathbb{R}^n$ represents the body forces acting on the fluid (such as gravity). The second equation, the continuity equation, enforces the *incompressibility* of the velocity: in words, the amount of fluid flowing into an infinitesimal volume is the same as the amount of fluid flowing out of the infinitesimal volume.

For simplicity, suppose the boundary conditions are that $u = 0$ on the entire boundary $\Gamma = \partial\Omega$[4]. Let us multiply (14.1.1) by a vector-valued test function $v \in V$, and (14.1.2) by a scalar-valued test function $q \in Q$, with V and Q to be determined, and integrate the

[4] If the velocity is nonzero on the boundary, we can convert the problem into a homogeneous one by the linearity of the PDE.

viscosity term by parts:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x, \tag{14.1.3}$$

$$\int_\Omega q \nabla \cdot u \, \mathrm{d}x = 0. \tag{14.1.4}$$

We could leave it in this form, but as it is we require that both u and p are weakly differentiable. If we integrate the pressure gradient term in the momentum equation by parts once, we can relax the differentiability requirement on the pressure space:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_\Omega p \nabla \cdot v \, \mathrm{d}x + \int_\Gamma p v \cdot n \, \mathrm{d}s = \int_\Omega f \cdot v \, \mathrm{d}x, \tag{14.1.5}$$

and the surface integral term disappears if we choose $v = 0$ on Γ. Note also that we can negate both sides of the continuity equation to arrive at the symmetric formulation

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_\Omega p \nabla \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x, \tag{14.1.6}$$

$$-\int_\Omega q \nabla \cdot u \, \mathrm{d}x = 0. \tag{14.1.7}$$

By inspection, we clearly need one weak derivative on u and v, and to enforce the Dirichlet bondition, so $V = H_0^1(\Omega; \mathbb{R}^n)$ is the appropriate choice. An obvious choice for the pressure space is $L^2(\Omega)$, but this is not quite sufficient; it is clear from the strong formulation of the problem that if (u, p) is a solution, then so is $(u, p + c)$ where $c \in \mathbb{R}$[5]. That is, the pressure is only defined up to a constant, and so to fix a unique pressure we choose

[5] Fundamentally, this is because

$$\int_\Omega \nabla \cdot v \, \mathrm{d}x = \int_\Gamma v \cdot n \, \mathrm{d}s = 0$$

with this choice of boundary conditions.

$$Q = L_0^2(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q \, \mathrm{d}x = 0\}. \tag{14.1.8}$$

Thus, the abstract formulation of this symmetric problem is: find $(u, p) \in V \times Q$ such that

$$a(u, v) + b(v, p) = F(v), \tag{14.1.9}$$

$$b(u, q) = 0, \tag{14.1.10}$$

for all $(v, q) \in V \times Q$, where in this case

$$a(u, v) = \int_\Omega \nabla u : \nabla v \, \mathrm{d}x, \tag{14.1.11}$$

and

$$b(u, q) = -\int_\Omega q \nabla \cdot u \, \mathrm{d}x. \tag{14.1.12}$$

14.2 *Stokes as an energy minimisation problem*

The Stokes equations arise from the minimisation of an energy functional, just as symmetric coercive problems do. The difference in the Stokes equations is that *there is an additional constraint*: that the velocity u be divergence-free.

Consider the following minimisation problem:

$$u = \underset{v \in H_0^1(\Omega;\mathbb{R}^n)}{\operatorname{argmin}} \ \frac{1}{2}\int_\Omega \nabla v : \nabla v \,\mathrm{d}x - \int_\Omega f \cdot v \,\mathrm{d}x. \tag{14.2.1}$$

If we take the Gâteaux derivative and set it equal to zero, we find

$$\text{find } u \in H_0^1(\Omega;\mathbb{R}^n) \text{ such that } \int_\Omega \nabla u : \nabla v \,\mathrm{d}x = \int_\Omega f \cdot v \,\mathrm{d}x \text{ for all } v \in H_0^1(\Omega). \tag{14.2.2}$$

This is the weak form of

$$-\nabla^2 u = f \qquad \text{in } \Omega, \tag{14.2.3}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{14.2.4}$$

the vector Laplacian.

Now consider the following modification to this problem:

$$u = \underset{v \in H_0^1(\Omega;\mathbb{R}^n)}{\operatorname{argmin}} \ \frac{1}{2}\int_\Omega \nabla v : \nabla v \,\mathrm{d}x - \int_\Omega f \cdot v \,\mathrm{d}x, \tag{14.2.5}$$

$$\text{subject to} \qquad \nabla \cdot v = 0. \tag{14.2.6}$$

As in finite-dimensional optimisation, we can write the optimality conditions for this problem by introducing a Lagrange multiplier $p \in L_0^2(\Omega)$ and writing the Lagrangian $L : H_0^1(\Omega;\mathbb{R}^n) \times L_0^2(\Omega) \to \mathbb{R}$

$$L(u,p) = \frac{1}{2}\int_\Omega \nabla u : \nabla u \,\mathrm{d}x - \int_\Omega f \cdot u \,\mathrm{d}x - \int_\Omega p \nabla \cdot u \,\mathrm{d}x. \tag{14.2.7}$$

The optimality conditions for this problem are: find (u,p) such that

$$L_u(u,p;v) = 0, \tag{14.2.8}$$

$$L_p(u,p;q) = 0, \tag{14.2.9}$$

for all $(v,q) \in H_0^1(\Omega;\mathbb{R}^n) \times L_0^2(\Omega)$. On computing the Gâteaux derivatives, we find exactly the Stokes system (14.1.6). That is, the Stokes equations *are the saddle point system arising from the minimisation of the Dirichlet energy subject to the incompressibility constraint*; furthermore, the pressure *arises as the Lagrange multiplier that enforces this constraint*.

Remark 14.2.1 (Saddle point systems)**.** *It will turn out that this problem has a unique solution. This system is referred to as a* saddle point system *because its unique solution (u,p) satisfies*

$$L(u,q) \le L(u,p) \le L(v,p) \text{ for all } v \in H_0^1(\Omega;\mathbb{R}^n),\ q \in L_0^2(\Omega). \tag{14.2.10}$$

That is, (u,p) is a saddle point of the associated Lagrangian[6].

[6] This is only true if the bilinear form a is symmetric and coercive.

As with any discretisation, the first step in the analysis is to investigate the well-posedness of the underlying problem. One route to take is to analyse the Babuška conditions on the Hilbert space $V \times Q$, but it will turn out to be more convenient to verify Brezzi's reformulation of the conditions for mixed variational problems. To understand Brezzi's conditions, we turn to the finite dimensional case. Before we do so, we need to learn one more fact about Hilbert spaces.

14.3 Prelude: orthogonal decompositions in Hilbert spaces

A fundamental fact about Hilbert spaces is that they can be cleanly decomposed into any subspace and its orthogonal complement, just like Euclidean spaces.

Theorem 14.3.1 (Orthogonal decomposition of a Hilbert space)**.** *Let H be a Hilbert space, and suppose $K \subset H$ is a closed subspace of H. Then its orthogonal complement*

$$K^\perp := \{v \in H : v \perp k \text{ for all } k \in K\} \tag{14.3.1}$$

is also a closed subspace, and

$$H = K \oplus K^\perp, \tag{14.3.2}$$

which means that every $v \in V$ can be uniquely *written as*

$$v = v^K + v^\perp, \tag{14.3.3}$$

with $v^K \in K$ and $v^\perp \in K^\perp$.

Proof. See Proposition 2.3.5 of Brenner and Scott, or any textbook on functional analysis. □

14.4 Saddle point systems in finite dimensions: the homogeneous case

Consider the following $N \times N$ linear system:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{14.4.1}$$

where the submatrix A is a square $NA \times NA$ matrix, and B is a rectangular $NB \times NA$ matrix with $NA + NB = N$ and $NA > NB$. The unknowns are $u \in V = \mathbb{R}^{NA}$ and $p \in Q = \mathbb{R}^{NB}$. In the case of the Stokes equations, A represents the discretised vector Laplacian, B^T represents the gradient (mapping from scalar-valued Q to vector-valued V), and B represents the divergence (mapping from vector-valued V to scalar-valued Q). We refer to this as the homogeneous case because the right hand side of the equation $Bu = 0$ is homogeneous.

The second equation tells us that $Bu = 0$[7], and thus we know that u must be an element of the space

[7] We will consider the case where the second component of the right-hand side is nonzero in the next section.

$$K := \text{kernel}(B) = \{v \in V : Bv = 0\}. \tag{14.4.2}$$

This set is certainly nonempty, since $0 \in K$. If K is trivial, i.e. $K = \{0\}$, then $u = 0$ and the equations reduce to the solvability of $B^T p = f$, which we will consider below. We therefore assume for the moment that the kernel is nontrivial.

By multiplying the first equation on the left with v^T for $v \in K$, we have

$$v^T Au + v^T B^T p = v^T f, \tag{14.4.3}$$

and since $v^T B^T p = p^T (Bv)^T = 0$, we derive the problem

$$\text{find } u \in K \text{ such that } v^T Au = v^T f \text{ for all } v \in K. \tag{14.4.4}$$

This is now a variational problem posed on a *closed subspace of a Hilbert space*, and we understand how to ensure its well-posedness. In the easier case, we can suppose that A is coercive *on the kernel* K, i.e.

$$v^T Av > 0 \text{ for all } v \in K \setminus \{0\}, \tag{14.4.5}$$

in which case we can invoke the Lax–Milgram theorem 4.2.1. In general, we can suppose that A satisfies the Babuška conditions of theorem 13.9.1 on the kernel K.

Let us now suppose that we have solved the problem on the kernel for u, and let us understand what we have achieved. Write the orthogonal decomposition of K in V:

$$V = K \oplus K^{\perp}, \tag{14.4.6}$$

where $K^{\perp}$ is the space of vectors orthogonal to every element of K. Every vector $f \in V$ can be uniquely written as the sum of one element of K and one element of $K^{\perp}$,

$$f = f^K + f^{\perp}, \tag{14.4.7}$$

where $f^K \in K$ and $f^{\perp} \perp K$. The expression $v^T f$ in the variational problem (14.4.4) simplifies to

$$v^T f = v^T f^K + v^T f^{\perp} = v^T f^K, \tag{14.4.8}$$

and so $Au = f^K$, and $f - Au = f^{\perp}$.

Having solved for the variable u, we must complete the solution of the problem by computing the unique $p \in Q$ such that

$$B^T p = f - Au = f^{\perp}. \tag{14.4.9}$$

By the fundamental theorem of linear algebra,

$$\mathrm{range}(B^T) = \mathrm{kernel}(B)^\perp = K^\perp, \tag{14.4.10}$$

and since $f^\perp \in \mathrm{range}(B^T)$, there exists at least one $p \in Q$ such that $B^T p = f - Au$. However, we must ensure that there is only one such p; that is, we must ensure that B^T is injective.

To ensure injectivity, we wish to ensure that $B^T p = 0 \implies p = 0$. We saw in the previous lecture that one way to formulate this is via the inf-sup condition: there exists $\gamma \in \mathbb{R}$ such that

$$0 < \gamma \leq \inf_{\substack{q \in Q \\ q \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{v^T B^T q}{\|q\| \|v\|}. \tag{14.4.11}$$

If this holds, then the operator $B^T : Q \to K^\perp$ is a bijection, and we can solve for $p \in Q$ uniquely.

Remark 14.4.1. *Carefully distinguish between applying the inf-sup condition to B^T (a rectangular sub-matrix of M) and the entire square matrix M.*

14.5 Saddle point systems in finite dimensions: the inhomogeneous case

Now consider the modified problem

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{14.5.1}$$

the inhomogeneous case. Again, define K to be the kernel of B

$$K = \mathrm{kernel}(B) \tag{14.5.2}$$

and write

$$u = u^K + u^\perp \tag{14.5.3}$$

using the orthogonal decomposition of V[8]. Suppose that we change our basis so that we may write

[8] The main difference in the homogeneous case is that $u^\perp = 0$.

$$u = \begin{pmatrix} u^K \\ u^\perp \end{pmatrix}. \tag{14.5.4}$$

Such a change of basis is always possible. We can therefore write

$$f = \begin{pmatrix} f^K \\ f^\perp \end{pmatrix} \tag{14.5.5}$$

and

$$A = \begin{pmatrix} A^{KK} & A^{K\perp} \\ A^{\perp K} & A^{\perp\perp} \end{pmatrix}. \tag{14.5.6}$$

We can therefore rewrite our system of equations as

$$A^{KK}u^K + A^{K\perp}u^\perp = f^K, \tag{14.5.7}$$
$$A^{\perp K}u^K + A^{\perp\perp}u^\perp + B^T p = f^\perp, \tag{14.5.8}$$
$$Bu^\perp = g. \tag{14.5.9}$$

There is no $B^T p$ term in the first equation because its range is $K^\perp$ and so it can only contribute to the second equation after our change of basis.

If we assume the inf-sup condition (14.4.11) on B^T, this ensures that $B^T : Q \to K^\perp$ is a bijection. This also ensures that $B : K^\perp \to Q$ is a bijection. Thus, this condition will be enough to ensure that (14.5.9) can be solved uniquely for $u^\perp$. The unique solvability of (14.5.7) for u^K is exactly the same as the homogeneous case: either we have a coercivity condition of A on the nullspace of B, or more generally a set of Babuška conditions. Finally, the inf-sup condition (14.4.11) on B^T is necessary and sufficient for the unique solvability of (14.5.8) for p.

Thus, the inhomogeneous case requires no additional assumptions on A beyond that of the homogeneous case.

14.6 *Saddle point theory in infinite dimensions: Brezzi's theorem*

We now state the Brezzi conditions for the well-posedness of the abstract saddle point problem.

Theorem 14.6.1 (Well-posedness of saddle point problems)**.** *Let V and Q be Hilbert spaces. Given $F \in V^*$ and $G \in Q^*$, we consider the problem: find $(u, p) \in V \times Q$ such that*

$$a(u, v) + b(v, p) = F(v), \tag{14.6.1}$$
$$b(u, q) = G(q), \tag{14.6.2}$$

for all $(v, q) \in V \times Q$. Let

$$K = \{v \in V : b(v, q) = 0 \text{ for all } q \in Q\}. \tag{14.6.3}$$

Suppose that

(1) $a : V \times V$ and $b : V \times Q$ are bounded bilinear forms;

(2) The variational problem

$$\text{find } u \in K \text{ such that } a(u, v) = F(v) \text{ for all } v \in K \tag{14.6.4}$$

is well-posed;

(3) b satisfies the following inf-sup condition: there exists $\gamma \in \mathbb{R}$ such that

$$0 < \gamma \leq \inf_{\substack{q \in Q \\ q \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{b(v,q)}{\|v\|_V \|q\|_Q}. \tag{14.6.5}$$

Then there exists a unique pair $(u,p) \in V \times Q$ that solves the variational problem (14.6.1), and the solution is stable with respect to the data F and G.

Proof. See Brezzi[9]. □

[9] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 8(R2):129–151, 1974

Remark 14.6.2 (The Stokes equations)**.** *In the Stokes equations, the bilinear forms $a : H_0^1(\Omega;\mathbb{R}^n) \times H_0^1(\Omega;\mathbb{R}^n)$ and $b : H_0^1(\Omega;\mathbb{R}^n) \times L_0^2(\Omega)$ are bounded by Cauchy–Schwarz. The Poincaré–Friedrichs inequality guarantees that the $H^1(\Omega;\mathbb{R}^n)$ seminorm is a norm on $H_0^1(\Omega;\mathbb{R}^n)$, so a is coercive on the entirety of $H_0^1(\Omega;\mathbb{R}^n)$ with coercivity constant 1. Finally, the inf-sup condition: there exists $\gamma > 0$ such that*

$$\gamma \|q\|_{L^2(\Omega)} \leq \sup_{\substack{v \in H_0^1(\Omega;\mathbb{R}^n) \\ v \neq 0}} \frac{(q, \nabla \cdot v)_{L^2(\Omega)}}{|v|_{H^1(\Omega;\mathbb{R}^n)}} \tag{14.6.6}$$

was established by Ladyzhenskaya in 1969[10]. We therefore conclude that the Stokes equations are well-posed.

[10] O. A. Ladyzhenskaya. *The Mathematical Theory of Viscous Flows*. Gordon and Breach, 1969

Remark 14.6.3 (Equivalence of the Babuška and Brezzi conditions)**.** *One could also formulate the saddle-point system as a single bilinear variational form on*

$$\Lambda : (V \times Q) \times (V \times Q) \to \mathbb{R} \tag{14.6.7}$$

and consider the Babuška conditions for this problem. The Brezzi conditions follow from the Babuška conditions, and vice versa; this equivalence is proven in a technical report of Demkowicz[11].

[11] L. Demkowicz. Babuška $\iff$ Brezzi ?? Technical Report 06-08, University of Texas at Austin, 2006

14.7 *Finite element discretisations of mixed problems*

Let $V_h \subset V$ be a closed subspace of V, and let Q_h be a closed subspace of Q. Consider the Galerkin approximation of the Stokes problem (14.1.9): find $(u_h, p_h) \in V_h \times Q_h$ such that

$$a(u_h, v_h) + b(v_h, p_h) = F(v_h), \tag{14.7.1}$$
$$b(u_h, q_h) = 0, \tag{14.7.2}$$

for all $(v_h, q_h) \in V_h \times Q_h$.

In order for this to be well-posed, we will require that the variational problem involving a is well-posed on the discrete kernel

$$K_h = \{v_h \in V_h : b(v_h, q_h) = 0 \text{ for all } q_h \in Q_h\}. \tag{14.7.3}$$

Recall that the full kernel is

$$K = \{v \in V : b(v, q) = 0 \text{ for all } q \in Q\}. \tag{14.7.4}$$

Now, if $v_h \in K \cap V_h$, then $v_h \in K_h$ also: but the fact that $b(v_h, q_h) = 0$ for all $q_h \in Q_h$ might *not* imply that $b(v_h, q) = 0$ for all $q \in Q$. (It will for some choices of finite element pairs, and won't for others.) So $(K \cap V_h) \subseteq K_h$; there may be entries in the discrete kernel that are not in the full kernel. Thus, *well-posedness of a on the discrete kernel does not necessarily follow automatically from well-posedness of a on the full kernel*. In the particular situation of the Stokes equations, it does, because a is coercive on the entirety of V; but in more difficult problems this will not be the case.

Similarly, it does not follow from the fact that b satisfies the inf-sup condition over V and Q, it does *not* follow that b satisfies the inf-sup condition: there exists $\tilde{\gamma} \in \mathbb{R}$ such that

$$0 < \tilde{\gamma} \leq \inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{b(v_h, q_h)}{\|v_h\| \|q_h\|}. \tag{14.7.5}$$

We will see a surprising example of the failure of this condition with a familiar discretisation momentarily.

Before turning to this counterexample, we state the theorem of Brezzi that guarantees quasi-optimality of Galerkin discretisations under these additional assumptions.

Theorem 14.7.1 (Well-posedness and quasi-optimality of Galerkin discretisations of saddle-point problems)**.** *Consider the Galerkin approximation of* (14.6.1) *over $V_h \times Q_h$, a closed subspace of $V \times Q$:*

$$a(u_h, v_h) + b(v_h, p_h) = F(v_h), \tag{14.7.6}$$

$$b(u_h, q_h) = G(q_h). \tag{14.7.7}$$

Let

$$K_h = \{v_h \in V_h : b(v_h, q_h) = 0 \textit{ for all } q_h \in Q_h\}. \tag{14.7.8}$$

In addition to the assumptions of 14.6.1 that guarantee well-posedness of the continuous problem, suppose that

(1) The variational problem

$$\textit{find } u_h \in K_h \textit{ such that } a(u_h, v_h) = F(v_h) \textit{ for all } v_h \in K_h \tag{14.7.9}$$

is well-posed.

(2) b satisfies the following inf-sup condition over $V_h \times Q_h$: there exists $\tilde{\gamma} \in \mathbb{R}$ such that

$$0 < \tilde{\gamma} \leq \inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_Q}. \tag{14.7.10}$$

Then the Galerkin approximation (14.7.6) is well-posed. Furthermore, the approximate solutions are quasi-optimal: there exists $c < \infty$ such that

$$\|u - u_h\|_V + \|p - p_h\|_Q \le c \left(\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right). \tag{14.7.11}$$

Proof. See Theorem 2.1 of Brezzi[12]. □

[12] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 8(R2):129–151, 1974

Finer analyses are possible, identifying c in terms of continuity and inf-sup constants of the bilinear forms involved. It is also possible (through duality arguments) to analyse the approximation error of u_h and p_h separately.

14.8 Not all finite elements are stable

Consider the Stokes equations once more. Since we have had such success with the Lagrange elements for coercive problems, let us naïvely consider the simplest Lagrange finite element

$$[CG_1]^n \times [CG_1],$$

or in plain English let us use piecewise linear basis functions for each component of velocity and for the pressure. Let $V_h \times Q_h$ be the finite element function space that arises from equipping each cell K of a mesh $\mathcal{M}$ with this element. *The resulting finite element discretisation is unstable.*

We will show its instability by explicitly demonstrating that the discrete inf-sup condition (14.7.10) does not hold. We will do this by explicitly constructing a pressure $0 \neq p_h \in Q_h$ such that $b(v_h, p_h) = 0$ for all $v_h \in V_h$. We refer to such a p_h as a *spurious pressure mode*.

Let $\Omega = (0,1)^2$, and divide the edges into N intervals, with N divisible by three. Break up each square of the mesh into two triangles with a line of positive slope. Such a mesh with $N = 9$ is shown in figure 14.1. Each degree of freedom in the pressure space may be described by its logical coordinates (i, j), with $i, j = 0 \ldots N$.

Since the pressure field is determined completely by its degrees of freedom at the vertices, we specify its values there: let

$$p_h(i,j) = \begin{cases} 0 & \text{if } i + j \equiv 0 \pmod 3, \\ +1 & \text{if } i + j \equiv 1 \pmod 3, \\ -1 & \text{if } i + j \equiv 2 \pmod 3. \end{cases} \tag{14.8.1}$$

The resulting function is shown in figure 14.2.

This function is not equal to zero but has integral zero on each cell[13]. Now, since v_h is piecewise linear on each triangle K, $(\nabla \cdot v_h)|_K$

[13] To see this, use the simplest quadrature rule on a triangle, evaluation at the midpoint. This quadrature rule is exact for piecewise linear functions and evaluates to zero.

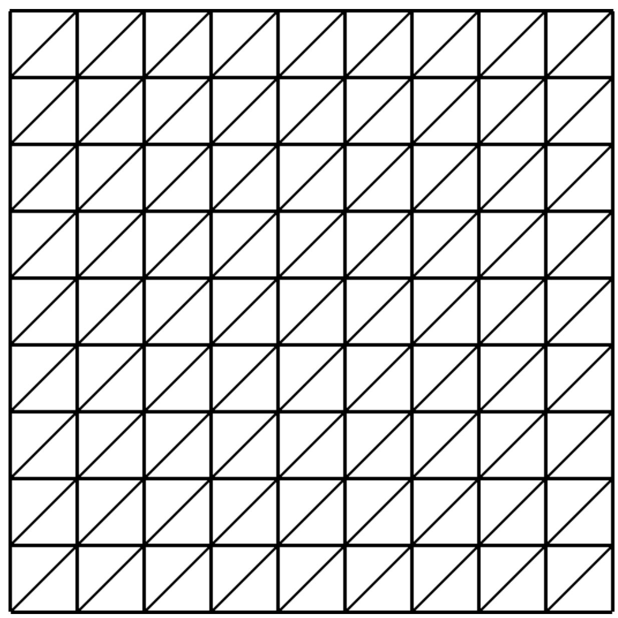

Figure 14.1: The mesh on which we will construct a spurious pressure mode.

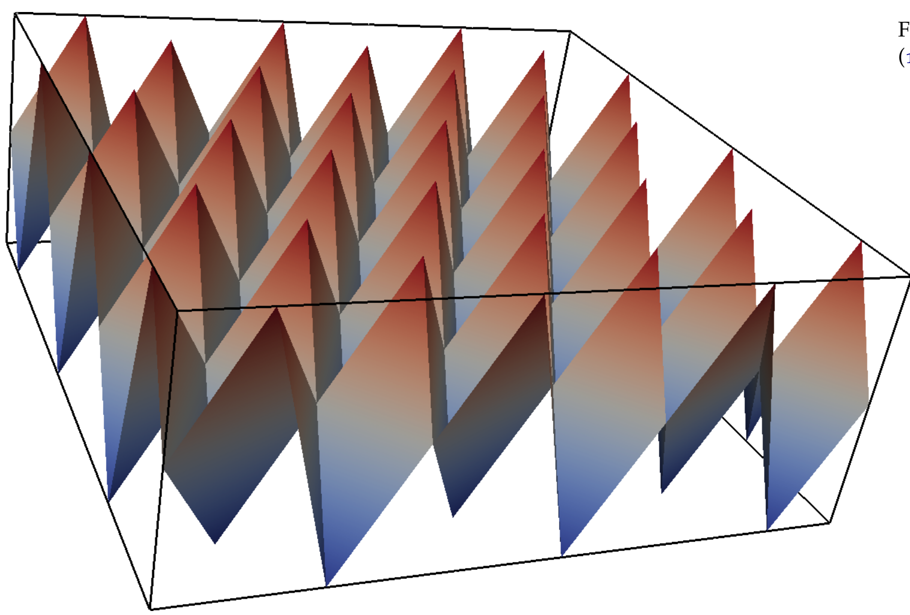

Figure 14.2: A spurious pressure mode (14.8.1).

is a constant. Therefore, for arbitrary $v_h \in V_h$,

$$\begin{aligned} b(v_h, p_h) &= -\int_\Omega p_h \nabla \cdot v_h \; \mathrm{d}x && (14.8.2) \\ &= \sum_{K \in \mathcal{M}} (\nabla \cdot v_h)|_K \int_K p_h \; \mathrm{d}x && (14.8.3) \\ &= 0. && (14.8.4) \end{aligned}$$

We thus have a $0 \neq p_h \in Q_h$ such that

$$\sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{b(v_h, p_h)}{\|v_h\|_V} = 0, \qquad (14.8.5)$$

and so the inf-sup condition cannot hold.

The Brezzi inf-sup condition (14.7.10) encodes a *compatibility condition* between the spaces V_h and Q_h: they cannot be chosen independently.

A final word: while this element is unstable (and generally any equal-order element is unstable), there are many choices of finite element discretisation which *are* stable. The simplest to implement is the Taylor–Hood finite element

$$[CG_2]^n \times [CG_1],$$

which does indeed satisfy the Brezzi inf-sup condition and yields a stable, convergent discretisation[14].

[14] C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Computers & Fluids*, 1(1):73–100, 1973

A Topics for further study

In this chapter we mention some of the topics that could be covered in a more advanced course on the finite element method.

A.1 Time-dependent PDEs

Consider the time-dependent PDE

$$\begin{aligned} \frac{\partial u}{\partial t} - \nabla^2 u &= f(x,t) && \text{in } \Omega \times (0,T), \\ u(x,t) &= 0 && \text{on } \partial\Omega \times (0,T), \\ u(x,0) &= u_0(x) && \text{in } \Omega. \end{aligned}$$

The analysis of such equations begins with *Bochner spaces*, a generalisation of Lebesgue spaces to functions whose values themselves lie in a Banach space. The solution at any instant is a function of space; we think of the solution of this problem as a function defined on $(0,T)$ whose value is itself a function of space:

$$u(x,t) = (u(t))\,(x). \tag{A.1.1}$$

That is, $u : (0,T) \to H^1_0(\Omega)$. Given a Banach space X, the Bochner space $L^p((0,T);X)$ for $p \in [1,\infty)$ is the set of all functions whose Bochner norm is finite:

$$\|u\|_{L^p([0,T];X)} = \int_0^T \|u(t)\|_X^p \, \mathrm{d}t < \infty. \tag{A.1.2}$$

For more details of these spaces, see Evans, section 5.9.2[1].

When discretising this equation, we can choose to either discretise in space first or in time first. These choices have different advantages and disadvantages, but in simple cases the operations commute. Let us suppose we discretise in space first; this is referred to as the *method of lines*. Multiplying by a space-dependent test function $v(x) \in H^1_0(\Omega)$ and integrating by parts, we find

$$\int_\Omega v\dot{u} \, \mathrm{d}x + \int_\Omega \nabla u \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x. \tag{A.1.3}$$

[1] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010

Now choose a closed subspace $V_h \subset H_0^1(\Omega)$ and expand $u(x,t)$ in terms of the finite element basis:

$$u(x,t) = \sum_{i=1}^{N} c_i(t)\phi_i(x), \tag{A.1.4}$$

where $V_h = \operatorname{span}(\phi_1, \dots, \phi_N)$. The finite element basis functions depend only on space and the coefficients depend only on time. The Galerkin approximation in space *reduces the PDE system to a large vector system of ODEs*:

$$M\dot{c} + Ac(t) = g(t), \tag{A.1.5}$$

$$c(0) = c_0, \tag{A.1.6}$$

where the mass matrix M is given by

$$M_{ji} = (\phi_i, \phi_j)_{L^2(\Omega)} = \int_\Omega \phi_i \phi_j \,\mathrm{d}x. \tag{A.1.7}$$

and the stiffness matrix A is given by

$$A_{ji} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j \,\mathrm{d}x. \tag{A.1.8}$$

This system, referred to as the *semidiscretised* system, may then be fully discretised using any standard algorithm for ODE initial value problems, such as the forward or backward Euler algorithms.

For a brief introduction to the Galerkin solution of parabolic equations, see Ern & Guermond[2], section 6.1; for a full exposition, see Thomée[3].

[2] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer, 2004

[3] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer, 2006

A.2 *Eigenvalue problems*

When one wishes to understand the stability of a system in time, the question often reduces to the solution of an eigenvalue problem. As a concrete example, let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain and consider: find eigenfunctions u and eigenvalues λ such that

$$\begin{aligned} -\nabla^2 u &= \lambda u \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \\ \int_\Omega u^2 \,\mathrm{d}x &= 1. \end{aligned} \tag{A.2.1}$$

This may be cast as a mixed variational problem in the following manner, with $V = H_0^1(\Omega)$: find $(u, \lambda) \in V \times \mathbb{R}$ such that

$$\int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega uv \,\mathrm{d}x, \tag{A.2.2}$$

$$\int_\Omega \mu u^2 \,\mathrm{d}x = \int_\Omega \mu \,\mathrm{d}x, \tag{A.2.3}$$

for all $(v, \mu) \in V \times \mathbb{R}$. The eigenvalues are real because the bilinear form is symmetric.

Given a closed subspace $V_h \subset V$, we construct the Galerkin approximation of this problem via: find $(u_h, \lambda) \in V \times \mathbb{R}$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h \, \mathrm{d}x = \int_\Omega u_h v_h \, \mathrm{d}x, \tag{A.2.4}$$

$$\int_\Omega \mu u_h^2 \, \mathrm{d}x = \int_\Omega \mu \, \mathrm{d}x, \tag{A.2.5}$$

for all $(v_h, \mu) \in V \times \mathbb{R}$.

Given a basis $V_h = \mathrm{span}(\phi_1, \ldots, \phi_N)$, this is equivalent to the eigenvalue problem

$$Ax = \lambda Mx, \tag{A.2.6}$$

where x is the vector of coefficients of u_h, the stiffness matrix A is given by

$$A_{ji} = a(\phi_i, \phi_j) = \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, \mathrm{d}x, \tag{A.2.7}$$

and the mass matrix

$$M_{ji} = (\phi_i, \phi_j)_{L^2(\Omega)} = \int_\Omega \phi_i \phi_j \, \mathrm{d}x. \tag{A.2.8}$$

Note that PDE eigenvalue problems become *generalised* eigenvalue problems from the perspective of linear algebra. The mass matrix on the right-hand side serves to enforce the correct inner product in the discrete space: the correct inner product to use in $\mathbb{R}^N$ is not the standard Euclidean inner product

$$(x, y) = x^T I y, \tag{A.2.9}$$

but the $L^2(\Omega)$ inner product

$$(x, y) = x^T M y. \tag{A.2.10}$$

As a concrete example, consider the L-shaped domain

$$\Omega = \Big((-1, -1) \times (1, 1) \Big) \setminus \Big((0, 1) \times (-1, 0) \Big). \tag{A.2.11}$$

We solve for the first five eigenfunctions of the Laplace Dirichlet eigenproblem: find $u \in H_0^1(\Omega)$ and $\lambda \in \mathbb{R}$ such that

$$-\nabla^2 u = \lambda u. \tag{A.2.12}$$

The eigenfunction with smallest eigenvalue $\lambda_1 \approx 9.644$ is shown in figure A.1.

For an excellent review of the finite element discretisation of eigenvalue problems, see Boffi[4].

[4] D. Boffi. Finite element approximation of eigenvalue problems. *Acta Numerica*, 19:1–120, 2010

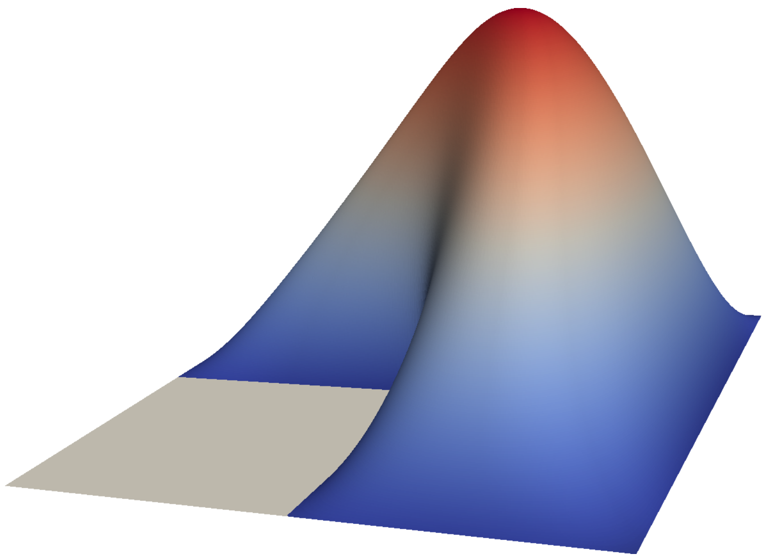

Figure A.1: The first eigenfunction of the Dirichlet Laplacian on the L-shaped domain. Any resemblance to the logo of a commercial software package is purely coincidental.

A.3 Variational inequalities

Let V be a Hilbert space, let $a : V \times V \to \mathbb{R}$ be a symmetric coercive bilinear form, and let $F \in V^*$. Recall the relationship between the optimisation problem

$$u = \underset{v \in V}{\operatorname{argmin}}\ J(v) = \frac{1}{2}a(v,v) - F(v) \tag{A.3.1}$$

and the variational equation

$$\text{find } u \in V \text{ such that } a(u,v) = F(v) \text{ for all } v \in V. \tag{A.3.2}$$

In section 6.2, we proved that their solutions were equivalent. We now consider optimisation problems with *inequality constraints*, with the consequence that *the solution set is not a vector space*.

Consider the following *obstacle* problem, which models the deformation of an elastic membrane stretched taut over an obstacle. The obstacle is represented by a function $\psi(x,y)$; we demand that the vertical deformation of the membrane $u(x,y)$ is such that[5]

$$u(x,y) \geq \psi(x,y) \text{ almost everywhere in } \Omega. \tag{A.3.3}$$

[5] Recall that "almost everywhere" (summarised as a.e.) means "except possibly on sets of measure zero", i.e. of area zero in two space dimensions.

The membrane deformation still seeks to minimise the Dirichlet energy, but we now have an additional pointwise-a.e. constraint. The mathematical formulation is

$$u = \underset{v \in K}{\operatorname{argmin}}\ J(v) = \frac{1}{2}a(v,v) - F(v), \tag{A.3.4}$$

where the feasible set K is given by

$$K := \{v \in H_0^1(\Omega) : v \geq \psi \text{ a.e.}\}, \tag{A.3.5}$$

and the bilinear form represents the Dirichlet energy

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x. \tag{A.3.6}$$

We demand that $\psi \leq 0$ on the boundary $\partial\Omega$, so that K is nonempty. With this constraint, this problem has a unique solution (Ciarlet, section 5.1)[6].

[6] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1978. Reprinted by SIAM in 2002

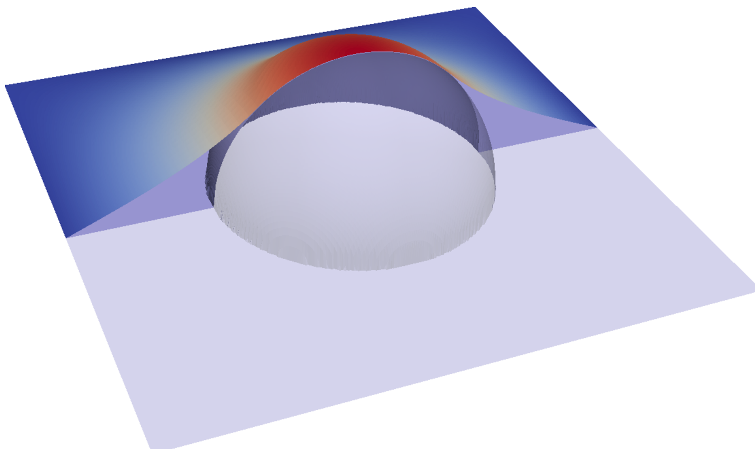

Figure A.2: A solution of the obstacle problem (A.3.4) with obstacle given by (A.3.7). The membrane is shown in colour (coloured by its vertical deformation) and the obstacle is shown in grey. Only half of the membrane is visualised; the computed solution extends over the whole domain.

Before deriving the optimality conditions, we present a concrete example to build intuition. Let $\Omega = (-2, 2) \times (-2, 2)$, and let ψ be a hemisphere of radius 1 centred at the origin, given by

$$\psi(x, y) = \begin{cases} \sqrt{1 - x^2 - y^2} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{A.3.7}$$

If we apply no external forces ($F = 0$), the solution of the problem is shown in figure A.2 (in colour), along with the obstacle (in grey). In one part of the domain Ω, the membrane is in contact with the obstacle ($u = \psi$); in the rest of the domain, the membrane is not in contact ($u > \psi$), and the membrane acts to minimise the Dirichlet energy. The free boundary, where the solution meets the obstacle, is not known in advance and must be determined as part of the solution.

Another example is given by a "staircase" obstacle, where $\Omega = (-1, 1) \times (-1, 1)$ and

$$\psi(x, y) = \begin{cases} -0.2 & -1 \leq x < -0.5, \\ -0.4 & 0.5 \leq x < 0, \\ -0.6 & 0 \leq x < 0.5, \\ -0.8 & 0.5 \leq x \leq 1. \end{cases} \tag{A.3.8}$$

In this example the membrane is forced by gravity,

$$F(v) = \int_\Omega g v \, \mathrm{d}x, \tag{A.3.9}$$

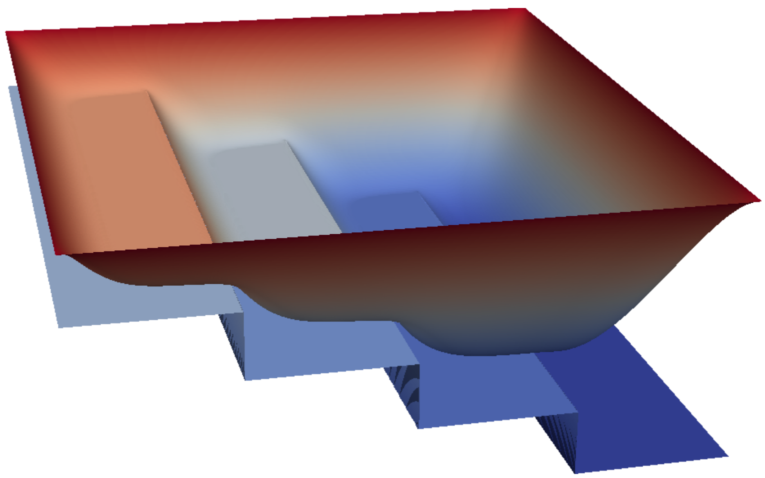

Figure A.3: A solution of the obstacle problem under gravity (A.3.4) with obstacle given by the staircase (A.3.8).

with $g = -10$. The solution is depicted in figure A.3[7].

The optimality conditions for this minimisation problem will capture the notion that a solution u is a minimiser if and only if the energy does not decrease along any *feasible* direction. Thus, we must understand what the feasible directions are. The feasible set K is convex; given $u, v \in K$, the convex combination of them is also feasible:

$$u + \varepsilon(v - u) \in K, \quad \varepsilon \in [0, 1]. \tag{A.3.10}$$

Thus, we may characterise every feasible perturbation of a candidate solution u by taking the set

$$D_\varepsilon(u) := \{u + \varepsilon(v - u) : v \in V\} \tag{A.3.11}$$

for $\varepsilon \in (0, 1]$. We now intuitively conjecture that the optimality condition will be that u is a solution to the constrained optimisation problem if

$$J'(u; v - u) \geq 0 \text{ for all } v \in K, \tag{A.3.12}$$

or in words that the energy can only increase along any feasible perturbation from u. This is a *variational inequality*: an inequality that must hold for all v varying over the set K.

Theorem A.3.1 (Optimality conditions for constrained optimisation problems)**.** *Let V be a Hilbert space, let $a : V \times V \to \mathbb{R}$ be a symmetric coercive bilinear form, and let $F \in V^*$. Given a nonempty closed convex subset $K \subset V$, any solution u of*

$$u = \underset{v \in K}{\operatorname{argmin}}\, J(v) = \frac{1}{2} a(v, v) - F(v) \tag{A.3.13}$$

must satisfy

$$J'(u; v - u) \geq 0 \ \text{ for all } v \in K. \tag{A.3.14}$$

[7] This example is taken from demo step-41 of the deal.ii finite element software package.

Proof. Let u be a solution of the minimisation problem, and let $v \in K$ be arbitrary. Since K is convex, $u + \varepsilon(v - u) \in K$ for $\varepsilon \in [0,1]$. As u is a minimiser, we must have that

$$J(u + \varepsilon(v - u) - J(u) \geq 0. \tag{A.3.15}$$

Therefore

$$\frac{J(u + \varepsilon(v - u) - J(u)}{\varepsilon} \geq 0, \tag{A.3.16}$$

and taking the limit as $\varepsilon \to 0^+$, we find

$$J'(u; v - u) \geq 0. \tag{A.3.17}$$

□

Expanding the definition of J' for the obstacle problem, it reads: find $u \in K$ such that

$$\int_\Omega \nabla u \cdot \nabla(v - u) \, \mathrm{d}x \geq \int_\Omega f(v - u) \, \mathrm{d}x \text{ for all } v \in K. \tag{A.3.18}$$

This is the weak form of

$$\begin{aligned} -\nabla^2 u \geq f & \quad \text{in } \Omega, \\ u = 0 & \quad \text{on } \partial\Omega, \\ u \geq \psi & \quad \text{a.e. in } \Omega, \\ (-\nabla^2 u - f)(u - \psi) = 0 & \quad \text{in } \Omega. \end{aligned} \tag{A.3.19}$$

The first equation states that the total force acting on the membrane is the applied force f plus something positive: the upward force that the obstacle exerts on the membrane where the membrane and obstacle are in contact. This additional force (a Lagrange multiplier) is unknown and must be solved for as part of the problem, but it must be so that the membrane does not penetrate the obstacle.

The final equation encodes a *complementarity* condition: where the membrane and the obstacle are not in contact ($u \neq \psi$), then the Laplace equation must hold, i.e. $-\nabla u = f$. In other words, the additional force preventing interpenetration can only act where there is contact. On the other hand, where $u = g$, then it might be the case that $-\nabla u - f \neq 0$ (although it could be the case: this corresponds to the membrane just touching the obstacle, not pressing against it).

We conclude with a few words on the numerical solution of such variational inequalities. Galerkin approximations are well-posed: they also have a unique solution, are stable, and possess a quasi-optimality property, just as in the case of linear variational equations[8]. We introduce an additional variable λ to represent the un-

[8] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1978. Reprinted by SIAM in 2002

known contact force; λ is the Lagrange multiplier that ensures non-penetration of the membrane and obstacle. The displacement and Lagrange multiplier satisfy

$$\text{(i) } u - \psi \geq 0 \quad \text{(ii) } \lambda \geq 0 \quad \text{(iii) } \lambda(u - \psi) = 0, \tag{A.3.20}$$

where all relations are understood pointwise-a.e. These inequalities are in turn transformed into a *rootfinding problem*: (A.3.20) is equivalent to

$$\phi(u - \psi, \lambda) = 0, \tag{A.3.21}$$

with

$$\phi(a, b) = \sqrt{a^2 + b^2} - a - b. \tag{A.3.22}$$

The catch is that the residual (A.3.21) is *not* Fréchet-differentiable, but it turns out that it has just enough smoothness to define a Newton-type method, called a *semismooth Newton method*. These algorithms are an extremely exciting recent development, with locally superlinear convergence and excellent performance in practical applications. For more information on the solution of variational inequalities with semismooth Newton methods, see the book of Ulbrich[9].

[9] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of *MOS-SIAM Series on Optimization*. SIAM, 2011

Bibliography

[1] M. T. Ainsworth and J. T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, New York, 2000.

[2] D. N. Arnold, G. Awanou, and R. Winther. Nonconforming tetrahedral mixed finite elements for elasticity. *Mathematical Models and Methods in Applied Sciences*, 24(04):783–796, 2014.

[3] J.-P. Aubin. *Analyse fonctionnelle appliqué*. Presses Universitaires de France, 1987.

[4] I. Babuška. Error-bounds for finite element method. *Numerische Mathematik*, 16(4):322–333, 1971.

[5] L. Blum, F. Cucker, M. Shub, and S. Smale. *Complexity and Real Computation*. Springer-Verlag, 1998.

[6] D. Boffi. Finite element approximation of eigenvalue problems. *Acta Numerica*, 19:1–120, 2010.

[7] D. Braess. *Finite Elements: theory, fast solvers, and applications in solid mechanics*. Cambridge University Press, third edition, 2007.

[8] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag New York, third edition edition, 2008.

[9] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 8(R2):129–151, 1974.

[10] F. Brezzi and K.-J. Bathe. A discourse on the stability conditions for mixed finite element formulations. *Computer Methods in Applied Mechanics and Engineering*, 82(1):27–57, 1990.

[11] G. Caloz and J. Rappaz. Numerical analysis for nonlinear and bifurcation problems. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume 5, pages 487–637. Elsevier, 1997.

[12] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North-Holland, 1978. Reprinted by SIAM in 2002.

[13] P. G. Ciarlet and C. Mardare. On the Newton–Kantorovich theorem. *Analysis and Applications*, 10(3):249–269, 2012.

[14] R. Cools. An encyclopaedia of cubature formulas. *Journal of Complexity*, 19(3):445–453, 2003.

[15] L. Demkowicz. Babuška $\iff$ Brezzi ?? Technical Report 06-08, University of Texas at Austin, 2006.

[16] P. Deuflhard. *Newton Methods for Nonlinear Problems*, volume 35 of *Springer Series in Computational Mathematics*. Springer-Verlag, 2011.

[17] S. Eisenstat and H. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM Journal on Scientific Computing*, 17(1):16–32, 1996.

[18] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite Elements and Fast Iterative Solvers: with applications in incompressible fluid dynamics*. Oxford University Press, 2014.

[19] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer, 2004.

[20] L. C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 2010.

[21] P.J. Frey and P.L. George. *Mesh Generation. Application to finite elements*. Wiley, 2nd edition, 2008.

[22] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, third edition, 2001.

[23] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, 2009.

[24] R. Hiptmair and C. Schwab. Numerical methods for elliptic and parabolic boundary value problems, 2008. http://www.sam.math.ethz.ch/~hiptmair/tmp/NAPDE_08.pdf.

[25] A. Johnen, J.-F. Remacle, and C. Geuzaine. Geometrical validity of curvilinear finite elements. *Journal of Computational Physics*, 233:359–372, 2013.

[26] L. Kantorovich. On Newton's method for functional equations. *Doklady Akademii Nauk SSSR*, 59:1237–1249, 1948.

[27] R. C. Kirby. From functional analysis to iterative methods. *SIAM Review*, 52(2):269–293, 2010.

[28] R. C. Kirby. A general approach to transforming finite elements. *SMAI Journal of Computational Mathematics*, 4:197–224, 2018.

[29] O. A. Ladyzhenskaya. *The Mathematical Theory of Viscous Flows*. Gordon and Breach, 1969.

[30] A. Logg, K. A. Mardal, G. N. Wells, et al. *Automated Solution of Differential Equations by the Finite Element Method*. Springer, 2011.

[31] K.-A. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. *Numerical Linear Algebra with Applications*, 18(1):1–40, 2011.

[32] V. Maz'ya. *Sobolev Spaces*, volume 342 of *A Series of Comprehensive Studies in Mathematics*. Springer, 2011.

[33] A. Mohsen. A simple solution of the Bratu problem. *Computers & Mathematics with Applications*, 67(1):26–33, 2014.

[34] J. Nečas. Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, 16(4):305–326, 1962.

[35] M. Rupflin. Lecture notes on Fixed Point Methods for Nonlinear PDEs, 2017. `https://courses.maths.ox.ac.uk/node/view_material/2037`.

[36] C. Schwab. *$p-$ and $hp-$ Finite Element Methods: Theory and Applications to Solid and Fluid Mechanics*. Numerical Mathematics and Scientific Computation. Clarendon Press, 1999.

[37] T. Schwedes, D. A. Ham, S. W. Funke, and M. D. Piggott. *Mesh Dependence in PDE-Constrained Optimisation*. Springer International Publishing, 2017.

[38] E. Süli. Lecture notes on Finite Element Methods for Partial Differential Equations, 2012. `http://people.maths.ox.ac.uk/suli/fem.pdf`.

[39] E. Süli. A brief excursion into the mathematical theory of mixed finite element methods, 2017. `http://people.maths.ox.ac.uk/suli/mixed_FEM_lectures.pdf`.

[40] E. Süli and D. F. Mayers. *An Introduction to Numerical Analysis*. Cambridge University Press, 2003.

[41] C. Taylor and P. Hood. A numerical solution of the Navier-Stokes equations using the finite element technique. *Computers & Fluids*, 1(1):73–100, 1973.

[42] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer, 2006.

[43] M. Ulbrich. *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*, volume 11 of *MOS-SIAM Series on Optimization*. SIAM, 2011.

[44] Verfürth, R. A note on polynomial approximation in Sobolev spaces. *Mathematical Modelling and Numerical Analysis*, 33(4):715–719, 1999.

[45] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numerische Mathematik*, 94(1):195–202, 2003.