C6.4 Finite Element Methods for PDEs

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

In this lecture I shall:

- explain what the finite element method is for;
- give a sketch of how it works;
- outline the questions we will address in the rest of the course.

What is the finite element method for?

The finite element method is a framework for computing numerical approximations to boundary and initial-boundary value problems.

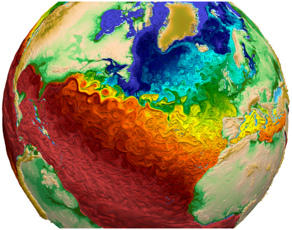
What is the finite element method for?

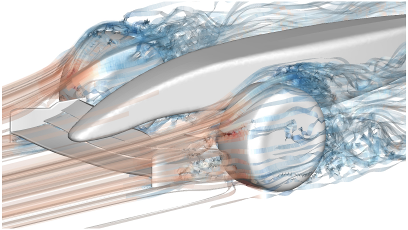
The finite element method is a framework for computing numerical approximations to boundary and initial-boundary value problems.

Example: the Navier–Stokes equations

Find velocity $u : \Omega \to \mathbb{R}^3$ and pressure $p : \Omega \to \mathbb{R}$ such that

$$-\nabla \cdot \left(\nu \left(\nabla u + (\nabla u)^{\mathrm{T}}\right)\right) + \nabla \cdot (u \otimes u) + \nabla p = f$$
$$\nabla \cdot u = 0$$





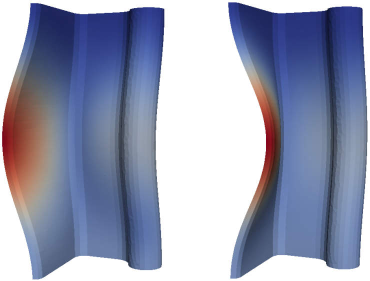
What is the finite element method for?

The finite element method is a framework for computing numerical approximations to boundary and initial-boundary value problems.

Example: the equations of elasticity

Find displacement $u : \Omega \to \mathbb{R}^3$ such that

$$-\mu\nabla^2 u - (\mu + \gamma)\nabla(\nabla \cdot u) = f$$



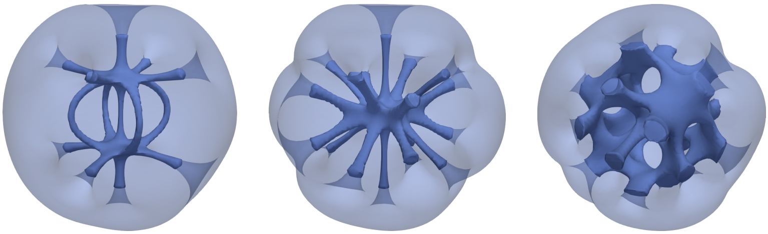
What is the finite element method for?

The finite element method is a framework for computing numerical approximations to boundary and initial-boundary value problems.

Example: the nonlinear Schrödinger equation

Find wave function $\psi : \Omega \times (0, T] \to \mathbb{C}$ such that

$$i\frac{\partial \psi}{\partial t} = -\frac{1}{2}\nabla^2\psi + |\psi|^2\psi + (x^2 + y^2 + z^2)\psi$$

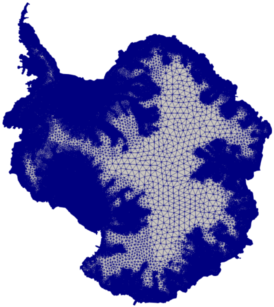


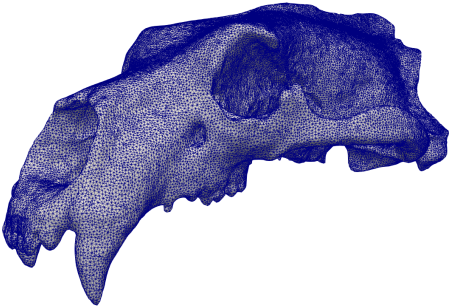
How does it work?

The finite element method converts PDE problems into algebraic ones.

$$Lu = f \rightsquigarrow Ax = b$$

The discrete system arises by breaking up the domain Ω into a mesh:





Comparing the finite element method

There are many different techniques used to compute numerical approximations of PDE. How does FEM compare?

✓ Able to handle complicated geometries (vs e.g. finite differences)

Comparing the finite element method

There are many different techniques used to compute numerical approximations of PDE. How does FEM compare?

- ✓ Able to handle complicated geometries (vs e.g. finite differences)
- ✓ Able to handle very general PDE (vs e.g. boundary element methods)

Comparing the finite element method

There are many different techniques used to compute numerical approximations of PDE. How does FEM compare?

- ✓ Able to handle complicated geometries (vs e.g. finite differences)
- ✓ Able to handle very general PDE (vs e.g. boundary element methods)
- ✓ Rooted in the modern variational theory of PDE

Comparing the finite element method

There are many different techniques used to compute numerical approximations of PDE. How does FEM compare?

- ✓ Able to handle complicated geometries (vs e.g. finite differences)
- ✓ Able to handle very general PDE (vs e.g. boundary element methods)
- ✓ Rooted in the modern variational theory of PDE
- ✓ Can design *structure-preserving* discretisations

Comparing the finite element method

There are many different techniques used to compute numerical approximations of PDE. How does FEM compare?

- ✓ Able to handle complicated geometries (vs e.g. finite differences)
- ✓ Able to handle very general PDE (vs e.g. boundary element methods)
- ✓ Rooted in the modern variational theory of PDE
- ✓ Can design *structure-preserving* discretisations
- ✗ Algebraic convergence (vs exponential with spectral methods), but this can be overcome: hp-version finite element method

Comparing the finite element method

There are many different techniques used to compute numerical approximations of PDE. How does FEM compare?

- ✓ Able to handle complicated geometries (vs e.g. finite differences)
- ✓ Able to handle very general PDE (vs e.g. boundary element methods)
- ✓ Rooted in the modern variational theory of PDE
- ✓ Can design *structure-preserving* discretisations
- ✗ Algebraic convergence (vs exponential with spectral methods), but this can be overcome: hp-version finite element method
- ✗ Can take more work per degree-of-freedom (vs e.g. finite differences)

Comparing the finite element method

There are many different techniques used to compute numerical approximations of PDE. How does FEM compare?

- ✓ Able to handle complicated geometries (vs e.g. finite differences)
- ✓ Able to handle very general PDE (vs e.g. boundary element methods)
- ✓ Rooted in the modern variational theory of PDE
- ✓ Can design *structure-preserving* discretisations
- ✗ Algebraic convergence (vs exponential with spectral methods), but this can be overcome: hp-version finite element method
- ✗ Can take more work per degree-of-freedom (vs e.g. finite differences)
- ✓ The most popular approach used in science and industry

Section 2

Basic steps of the finite element method

We illustrate the essential steps of the finite element method with the problem: for given $f : \Omega \to \mathbb{R}$, find $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla^2 u := -\nabla \cdot (\nabla u) = f & \quad \text{in } \Omega, \\ u = 0 & \quad \text{on } \partial\Omega. \end{aligned}$$

For $\Omega \subset \mathbb{R}^2$, this is

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) = f,$$

while for $\Omega \subset \mathbb{R}^3$, this is

$$-\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right) = f.$$

We illustrate the essential steps of the finite element method with the problem: for given $f : \Omega \to \mathbb{R}$, find $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\nabla^2 u := -\nabla \cdot (\nabla u) = f \quad & \text{in } \Omega, \\ u = 0 \quad & \text{on } \partial\Omega. \end{aligned}$$

The steps are:

- write as a variational problem over a suitable function space V;
- formulate over a finite-dimensional subspace $V_h \subset V$;
- construct V_h + basis with a mesh of Ω;
- assemble and solve the resulting linear system of equations.

We test the equation with a *test function* v and integrate:

$$-\int_\Omega (\nabla^2 u)\, v \,\mathrm{d}x = \int_\Omega f\, v \,\mathrm{d}x$$

and demand that this equality holds for "all" test functions (tbd).

We test the equation with a *test function* v and integrate:

$$-\int_\Omega (\nabla^2 u)\, v \; \mathrm{d}x = \int_\Omega f\, v \; \mathrm{d}x$$

and demand that this equality holds for "all" test functions (tbd).

We can now integrate by parts to shift derivatives onto v:

$$-\int_\Omega (\nabla^2 u)\, v \; \mathrm{d}x = \int_\Omega \nabla u \cdot \nabla v \; \mathrm{d}x - \int_{\partial\Omega} (\nabla u \cdot n)\, v \; \mathrm{d}s = \int_\Omega f\, v \; \mathrm{d}x.$$

We test the equation with a *test function* v and integrate:

$$-\int_\Omega (\nabla^2 u)\, v \, \mathrm{d}x = \int_\Omega f\, v \, \mathrm{d}x$$

and demand that this equality holds for "all" test functions (tbd).

We can now integrate by parts to shift derivatives onto v:

$$-\int_\Omega (\nabla^2 u)\, v \, \mathrm{d}x = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x - \int_{\partial\Omega} (\nabla u \cdot n)\, v \, \mathrm{d}s = \int_\Omega f\, v \, \mathrm{d}x.$$

Since we know u on the boundary, there is no need to test there. Let us therefore fix $v = 0$ on the boundary; hence,

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

for "all" test functions v. This is the *variational* or *weak* formulation of the Dirichlet boundary-value problem for Poisson's equation.

Our problem is to find the *trial function* $u \in V$ such that

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

for all *test functions* $v \in V$. What function space V should we look in?

Our problem is to find the *trial function* $u \in V$ such that

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

for all *test functions* $v \in V$. What function space V should we look in?

We need:

- u to have all first-order derivatives;
- the gradient of u to be square-integrable;
- u to satisfy $u = 0$ on the boundary.

The set of such functions is denoted by $H_0^1(\Omega) =: V$.

Our problem is to find the *trial function* $u \in V$ such that

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

for all *test functions* $v \in V$. What function space V should we look in?

We need:

- u to have all first-order derivatives;
- the gradient of u to be square-integrable;
- u to satisfy $u = 0$ on the boundary.

The set of such functions is denoted by $H_0^1(\Omega) =: V$.

Key advantage of variational formulation

Classical formulation would require $u \in C^2(\Omega) \cap C(\overline{\Omega})$; we now only require u to have square-integrable *first* derivatives.

V is infinite-dimensional; it is too big to search in! So to compute we look instead at *finite-dimensional subspaces* $V_h \subset V$.

The *Galerkin approximation* of our problem is: find $u_h \in V_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega f\, v_h \,\mathrm{d}x$$

for all $v_h \in V_h$. Here h represents the resolution of our discretisation, i.e. the maximal diameter of an element in the mesh.

V is infinite-dimensional; it is too big to search in! So to compute we look instead at *finite-dimensional subspaces* $V_h \subset V$.

The *Galerkin approximation* of our problem is: find $u_h \in V_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h \, \mathrm{d}x = \int_\Omega f \, v_h \, \mathrm{d}x$$

for all $v_h \in V_h$. Here h represents the resolution of our discretisation, i.e. the maximal diameter of an element in the mesh.

In this simple case the well-posedness of the discrete problem follows from the continuous one for any V_h. Some discretisations will be convergent ($u_h \to u$ as $h \to 0$ in a suitable norm) and some will not. We want discretisations that are well-posed and which converge quickly.

The finite element method is a way to construct a V_h ($+$ a basis) with good approximation properties and convenient computational properties.

The finite element method is a way to construct a V_h (+ a basis) with good approximation properties and convenient computational properties.

Key idea

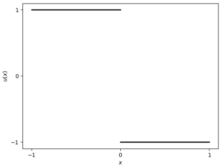
Represent a function with a polynomial on each cell of the mesh.

The finite element method is a way to construct a V_h (+ a basis) with good approximation properties and convenient computational properties.

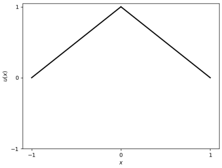
Key idea

Represent a function with a polynomial on each cell of the mesh.

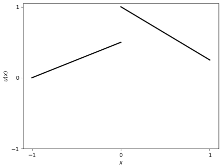
. . . with a specified degree of continuity.



discont., degree $= 0$



cont., degree $= 1$

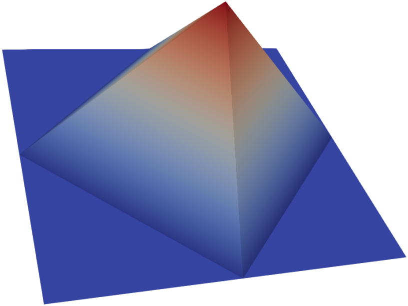


discont., degree $= 1$

The finite element method is a way to construct a V_h (+ a basis) with good approximation properties and convenient computational properties.

Key idea

Represent a function with a polynomial on each cell of the mesh.



Suppose we now have $V_h = \mathrm{span}\{\phi_1, \dots, \phi_N\}$. For brevity we write

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x, \quad F(v) := \int_\Omega f \, v \, \mathrm{d}x.$$

Suppose we now have $V_h = \mathrm{span}\{\phi_1, \ldots, \phi_N\}$. For brevity we write

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad F(v) := \int_\Omega f\, v \,\mathrm{d}x.$$

Let us expand u_h and v_h in terms of our basis. First write

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

Suppose we now have $V_h = \mathrm{span}\{\phi_1, \ldots, \phi_N\}$. For brevity we write

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x, \quad F(v) := \int_\Omega f\, v \,\mathrm{d}x.$$

Let us expand u_h and v_h in terms of our basis. First write

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

Then our problem becomes:

$$\begin{aligned} & a(u_h, v_h) = F(v_h) \\ \Longrightarrow\ & a\left(u_h, \sum_{i=1}^{N} V_i\phi_i\right) = F\left(\sum_{i=1}^{N} V_i\phi_i\right) \\ \Longrightarrow\ & \sum_{i=1}^{N} V_i a(u_h, \phi_i) = \sum_{i=1}^{N} V_i F(\phi_i). \end{aligned}$$

Suppose we now have $V_h = \mathrm{span}\{\phi_1, \ldots, \phi_N\}$. For brevity we write

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x, \quad F(v) := \int_\Omega f\, v \, \mathrm{d}x.$$

Let us expand u_h and v_h in terms of our basis. First write

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

Then our problem becomes:

$$\begin{aligned} & a(u_h, v_h) = F(v_h) \\ \implies \ & a\Big(u_h, \sum_{i=1}^{N} V_i \phi_i\Big) = F\Big(\sum_{i=1}^{N} V_i \phi_i\Big) \\ \implies \ & \sum_{i=1}^{N} V_i a(u_h, \phi_i) = \sum_{i=1}^{N} V_i F(\phi_i). \end{aligned}$$

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

$$a(u_h, \phi_i) = F(\phi_i) \quad \text{for } i = 1, \ldots, N.$$

Now expand u_h as

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

Now expand u_h as

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

Substituting, we find

$$a\left(\sum_{j=1}^{N} U_j \phi_j, \phi_i\right) = F(\phi_i), \quad i = 1, \ldots, N,$$

$$\implies \sum_{j=1}^{N} a(\phi_j, \phi_i) U_j = F(\phi_i), \quad i = 1, \ldots, N,$$

or in matrix notation

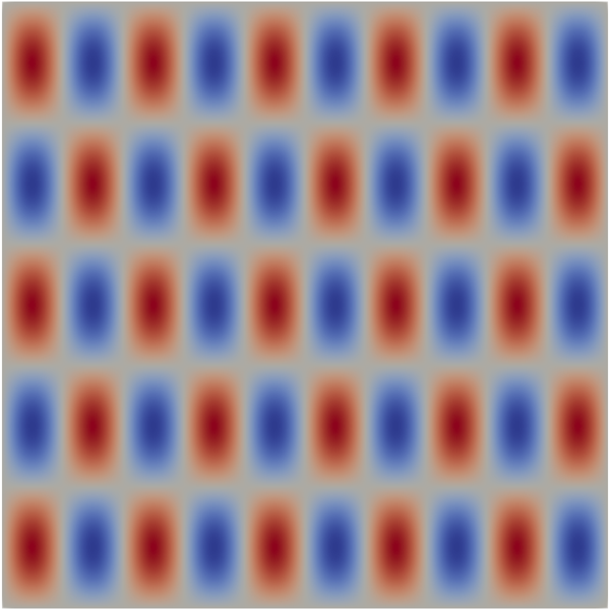
$$AU = b,$$

where $U = (U_1, \ldots, U_N)^{\mathrm{T}}$,

$$A_{ij} = a(\phi_j, \phi_i), \quad b_i = F(\phi_i).$$

Our numerical approximation is computed by solving this linear system.

```
from firedrake import *

mesh = UnitSquareMesh(128, 128, quadrilateral=True)
V = FunctionSpace(mesh, "CG", 1)
(x, y) = SpatialCoordinate(mesh)

f = sin(10*pi*x) * sin(5*pi*y)
bc = DirichletBC(V, 0, "on_boundary")

u = Function(V)
v = TestFunction(V)
G = inner(grad(u), grad(v))*dx - inner(f, v)*dx

solve(G == 0, u, bc)
File("output/poisson.pvd").write(u)
```



Section 3

Outlook

This sketch gives the core idea of how the finite element method works. However, each of these steps gets more complicated for real problems:

- There are different variational formulations for the same PDE, with different advantages and disadvantages.

This sketch gives the core idea of how the finite element method works. However, each of these steps gets more complicated for real problems:

- There are different variational formulations for the same PDE, with different advantages and disadvantages.
- For most problems the subspace V_h must be chosen carefully to achieve a convergent method; any old choice won't work.

This sketch gives the core idea of how the finite element method works. However, each of these steps gets more complicated for real problems:

- There are different variational formulations for the same PDE, with different advantages and disadvantages.
- For most problems the subspace V_h must be chosen carefully to achieve a convergent method; any old choice won't work.
- Fast solvers for the resulting linear systems must exploit the PDE structure.

The main questions we will address in the remainder of this course are:

- How do we formulate problems variationally?
  [Lec. 2, 3, 5, 6, 12]

The main questions we will address in the remainder of this course are:

- How do we formulate problems variationally?
  [Lec. 2, 3, 5, 6, 12]
- Are the continuous and discrete variational problems well-posed?
  [Lec. 4, 13, 14-16]

The main questions we will address in the remainder of this course are:

- How do we formulate problems variationally?
  [Lec. 2, 3, 5, 6, 12]
- Are the continuous and discrete variational problems well-posed?
  [Lec. 4, 13, 14-16]
- What error is incurred in the approximation?
  [Lec. 6, 7, 8, 11, 13]

The main questions we will address in the remainder of this course are:

- How do we formulate problems variationally?
  [Lec. 2, 3, 5, 6, 12]
- Are the continuous and discrete variational problems well-posed?
  [Lec. 4, 13, 14-16]
- What error is incurred in the approximation?
  [Lec. 6, 7, 8, 11, 13]
- How do we implement the finite element method?
  [Lec. 8, 9, 10]

C6.4 Finite Element Methods for PDEs
Lecture 2: Lebesgue spaces

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

Lecture 2: Lebesgue spaces

Our goal for the next three lectures is to prove the *Lax–Milgram* theorem about the well-posedness of the linear variational problem

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

where $a : V \times V \to \mathbb{R}$ is bilinear and $F : V \to \mathbb{R}$ is linear.

Lecture 2: Lebesgue spaces

Our goal for the next three lectures is to prove the *Lax–Milgram* theorem about the well-posedness of the linear variational problem

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

where $a : V \times V \to \mathbb{R}$ is bilinear and $F : V \to \mathbb{R}$ is linear.

When it applies, the Lax–Milgram theorem will prove the well-posedness of both the continuous problem and its Galerkin approximation.

Lecture 2: Lebesgue spaces

Our goal for the next three lectures is to prove the *Lax–Milgram* theorem about the well-posedness of the linear variational problem

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

where $a : V \times V \to \mathbb{R}$ is bilinear and $F : V \to \mathbb{R}$ is linear.

When it applies, the Lax–Milgram theorem will prove the well-posedness of both the continuous problem and its Galerkin approximation.

Moreover, we can bound the error in the Galerkin approximation in terms of the constants arising in the statement of the Lax–Milgram theorem.

Lecture 2: Lebesgue spaces

Our goal for the next three lectures is to prove the *Lax–Milgram* theorem about the well-posedness of the linear variational problem

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

where $a : V \times V \to \mathbb{R}$ is bilinear and $F : V \to \mathbb{R}$ is linear.

When it applies, the Lax–Milgram theorem will prove the well-posedness of both the continuous problem and its Galerkin approximation.

Moreover, we can bound the error in the Galerkin approximation in terms of the constants arising in the statement of the Lax–Milgram theorem.

First however we need to understand the *Sobolev* spaces V in which we look for solutions, and to that end we need to discuss *Lebesgue* spaces.

Definition (normed linear space)

A *normed linear space* X is a linear space equipped with a norm $\|\cdot\| : X \to \mathbb{R}$ that satisfies the following properties:

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

Definition (normed linear space)

A *normed linear space* X is a linear space equipped with a norm $\|\cdot\| : X \to \mathbb{R}$ that satisfies the following properties:

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

Definition (Banach space)

A Banach space is a complete normed linear space.

Definition (normed linear space)

A *normed linear space* X is a linear space equipped with a norm $\|\cdot\| : X \to \mathbb{R}$ that satisfies the following properties:

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

Definition (Banach space)

A Banach space is a complete normed linear space.

Recall that *completeness* of a normed linear space X means that all Cauchy sequences in X converge in X. A Cauchy sequence (x_n) is one where $(\forall \varepsilon > 0)\,(\exists\, N > 0)\,(\forall\, m, n > N)\ \|x_n - x_m\| < \varepsilon$.

Example

Euclidean space $\mathbb{R}^n$ equipped with the 1-norm, the 2-norm, and the supremum norm

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}, \quad \|x\|_\infty := \max_{1 \le i \le n} |x_i|$$

are all Banach spaces.

Example

Euclidean space $\mathbb{R}^n$ equipped with the 1-norm, the 2-norm, and the supremum norm

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|, \quad \|x\|_2 := \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}, \quad \|x\|_\infty := \max_{1\le i\le n} |x_i|$$

are all Banach spaces.

Example

The space of continuous functions from a domain $\overline{\Omega}$ to $\mathbb{R}$ equipped with the supremum norm

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

is a Banach space.

Definition (inner product space)

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

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

Definition (inner product space)

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

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

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

Definition (inner product space)

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

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

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

Definition (Hilbert space)

A *Hilbert space* is a complete inner product space.

Example

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

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

Example

The space of square-integrable functions on a domain $L^2(\Omega)$ with inner product

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

Example

The space $H^1_0(\Omega)$ of square-integrable functions that are zero on the boundary and that have square-integrable derivatives is a Hilbert space with inner product

$$(u, v)_{H^1_0(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x.$$

If $(u, u)_{H^1_0(\Omega)} = 0$ then u must be constant; but the only constant function in $H^1_0(\Omega)$ is the zero function thanks to the boundary condition $u|_{\partial\Omega} = 0$.

Example

The space $H^1(\Omega) = H(\mathrm{grad}, \Omega)$ of square-integrable functions that have square-integrable gradient is a Hilbert space with inner product

$$(u, v)_{H^1(\Omega)} = \int_\Omega uv + \nabla u \cdot \nabla v \, \mathrm{d}x.$$

Example

The space $H^1(\Omega) = H(\mathrm{grad}, \Omega)$ of square-integrable functions that have square-integrable gradient is a Hilbert space with inner product

$$(u, v)_{H^1(\Omega)} = \int_\Omega uv + \nabla u \cdot \nabla v \, \mathrm{d}x.$$

Example

The space $H(\mathrm{div}, \Omega)$ of square-integrable vector-valued functions that have square-integrable divergence is a Hilbert space with inner product

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

Example

The space $H^1(\Omega) = H(\mathrm{grad}, \Omega)$ of square-integrable functions that have square-integrable gradient is a Hilbert space with inner product

$$(u, v)_{H^1(\Omega)} = \int_\Omega uv + \nabla u \cdot \nabla v \, \mathrm{d}x.$$

Example

The space $H(\mathrm{div}, \Omega)$ of square-integrable vector-valued functions that have square-integrable divergence is a Hilbert space with inner product

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

Example

For $\Omega \subset \mathbb{R}^3$, the space $H(\mathrm{curl}, \Omega)$ of square-integrable vector-valued functions that have square-integrable curl is a Hilbert space with inner product

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

Theorem (Cauchy–Schwarz inequality)

For a Hilbert space X and any $u, v \in X$,

$$|(u, v)_X| \leq \|u\|_X \|v\|_X.$$

Proof.

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

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

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,

$$|2(u, v)_X|^2 - 4\|u\|_X^2 \|v\|_X^2 \leq 0.$$

Section 2

Dual of a Hilbert space

Definition (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

$$j(\alpha u + \beta v) = \alpha j(u) + \beta j(v).$$

Definition (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

$$j(\alpha u + \beta v) = \alpha j(u) + \beta j(v).$$

Example

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

Definition (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

$$j(\alpha u + \beta v) = \alpha j(u) + \beta j(v).$$

Example

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

Example

Drag over a wing, compliance of a structure, average global temperature.

Definition (Bounded linear functional)

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

$$|j(v)| \leq L\|v\|_X \quad \forall\, v \in X.$$

Definition (Bounded linear functional)

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

$$|j(v)| \leq L\|v\|_X \quad \forall\, v \in X.$$

Lemma (Boundedness and continuity)

Boundedness is equivalent to continuity.

Proof.

See notes, Lemma 2.3.4. □

Example

Given any $g \in X$, we can consider

$$j(v) := (g, v)_X.$$

This is a bounded linear functional as by the Cauchy–Schwarz inequality

$$|j(v)| = |(g, v)_X| \leq \|g\|_X \|v\|_X.$$

Definition (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:

$$\|j\|_{X^*} := \sup_{\|v\|_X = 1} |j(v)| \, .$$

This gives the "tightest L" in the definition of boundedness.

Definition (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:

$$\|j\|_{X^*} := \sup_{\|v\|_X = 1} |j(v)| \,.$$

This gives the "tightest L" in the definition of boundedness.

Given a $j \in X^*$, denote the action of j on v by

$$\langle j, v \rangle := j(v).$$

This is called the *duality pairing*.

Theorem (Riesz Representation Theorem)

Any bounded linear functional $j \in X^*$ *can be uniquely represented by a* $g \in X$, *via*

$$\langle j, v\rangle = (g, v) \qquad \text{for all } v \in X.$$

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

Theorem (Riesz Representation Theorem)

Any bounded linear functional $j \in X^*$ can be uniquely represented by a $g \in X$, via

$$\langle j, v\rangle = (g, v) \quad \textit{for all } v \in X.$$

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.

Theorem (Riesz Representation Theorem)

Any bounded linear functional $j \in X^*$ *can be uniquely represented by a* $g \in X$*, via*

$$\langle j, v\rangle = (g, v) \qquad \textit{for all } v \in X.$$

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.

Example

Let $X = L^2(\Omega)$ and let

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

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

Section 3

Lebesgue spaces

Definition (Lebesgue p-norm, $p \in [1, \infty)$)

Let $p \in [1, \infty)$. The $L^p(\Omega)$ norm is defined by

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

Definition (Lebesgue p-norm, $p = \infty$)

The $L^\infty(\Omega)$ norm is defined by

$$\|u\|_{L^\infty(\Omega)} = \inf\{C \geq 0 : |u(x)| \leq C \text{ almost everywhere}\}.$$

This is the *essential supremum* of $|u|$.

Definition (Lebesgue space)

For $p \in [1, \infty]$, consider the definition

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

These are Banach spaces for all p, a Hilbert space for $p = 2$.

Definition (Lebesgue space)

For $p \in [1, \infty]$, consider the definition

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

These are Banach spaces for all p, a Hilbert space for $p = 2$.

Important remark

The Lebesgue integral ignores differences on a set of measure zero. Two functions f and g that differ only on a set of measure zero will have $\|f - g\|_{L^p} = 0$. In order to fix this, we actually take elements of L^p to be *equivalence classes* of functions that differ up to sets of measure zero.

Definition (Lebesgue space)

For $p \in [1, \infty]$, consider the definition

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

These are Banach spaces for all p, a Hilbert space for $p = 2$.

Important remark

The Lebesgue integral ignores differences on a set of measure zero. Two functions f and g that differ only on a set of measure zero will have $\|f - g\|_{L^p} = 0$. In order to fix this, we actually take elements of L^p to be *equivalence classes* of functions that differ up to sets of measure zero.

Consequence

For $f \in L^p(\Omega)$, to evaluate $f(x)$, we have to prove that there is a *continuous* function in the equivalence class and evaluate *that*.

Fix $\Omega \subset \mathbb{R}^n$ to have finite measure: $|\Omega| < \infty$.

Example

The function $f(x) = 1$ is in $L^p(\Omega)$ for all p:

$$\begin{aligned} \|1\|_{L^p(\Omega)} &= \left(\int_\Omega 1^p \, \mathrm{d}x \right)^{1/p} \\ &= |\Omega|^{1/p} < \infty. \end{aligned}$$

Fix $\Omega \subset \mathbb{R}^n$ to have finite measure: $|\Omega| < \infty$.

Example

The function $f(x) = 1$ is in $L^p(\Omega)$ for all p:

$$\begin{aligned} \|1\|_{L^p(\Omega)} &= \left(\int_\Omega 1^p \, \mathrm{d}x \right)^{1/p} \\ &= |\Omega|^{1/p} < \infty. \end{aligned}$$

Example

Let $\Omega = (0, 1)$ and let

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

Then $f_q \in L^p(\Omega) \iff q < 1/p$. That is, $\frac{1}{x} \notin L^1(\Omega)$, but $\frac{1}{x^{0.999}} \in L^1(\Omega)$, and $\frac{1}{\sqrt{x}} \notin L^2(\Omega)$, 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. For $L^\infty(\Omega)$ no blow-up whatsoever is allowed.

Theorem (Hölder's inequality)

Let $p, q \in [1, \infty]$ *such that*

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

The elements of such a pair are called Hölder conjugates. By convention here, 1 *and* ∞ *are conjugate.*

If $f \in L^p(\Omega)$ *and* $g \in L^q(\Omega)$*, then* $fg \in L^1(\Omega)$ *and*

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

Theorem (Hölder's inequality)

Let $p, q \in [1, \infty]$ *such that*

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

The elements of such a pair are called Hölder conjugates. By convention here, 1 *and* ∞ *are conjugate.*

If $f \in L^p(\Omega)$ *and* $g \in L^q(\Omega)$*, then* $fg \in L^1(\Omega)$ *and*

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

Theorem (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.

See notes, Theorem 2.5.10. $\square$

C6.4 Finite Element Methods for PDEs
Lecture 3: Sobolev spaces

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

In the previous lecture, we studied Lebesgue spaces, which capture the *integrability* properties of functions.

Lecture 3: Sobolev spaces

In the previous lecture, we studied Lebesgue spaces, which capture the *integrability* properties of functions.

In this lecture, we now study *Sobolev* spaces, which also capture the *differentiability* properties of functions. First, however, we must generalise our notion of differentiation.

Definition (Classical/strong differentiability in one dimension)

$$f'(x) := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

Definition (Classical/strong differentiability in one dimension)

$$f'(x) := \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}.$$

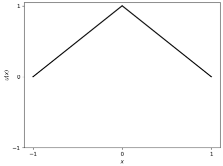
Remark

This is based on pointwise evaluation. We have seen however that pointwise evaluation is not a native concept to Lebesgue functions. We cannot always do it!

Recall that the variational formulation of the Poisson equation requires evaluating terms such as

$$\int_\Omega \nabla u_h \cdot \nabla v_h \, \mathrm{d}x.$$

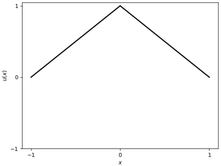
How do we do this if u_h is a C^0 piecewise polynomial?



Recall that the variational formulation of the Poisson equation requires evaluating terms such as

$$\int_\Omega \nabla u_h \cdot \nabla v_h \,\mathrm{d}x.$$

How do we do this if u_h is a C^0 piecewise polynomial?



Answer

We will develop a generalised notion of derivative, called the *weak derivative*, built on integration by parts.

To motivate the definition, first suppose $f \in C^1([a,b])$. Let $\phi \in C^1([a,b])$ and suppose that $\phi(a) = 0$ and $\phi(b) = 0$. Then, integration by parts tells us that

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

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 (Compact support in Ω)

A function $\phi \in C(\Omega)$ has *compact support* if

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

is compact (i.e. is bounded and closed) and is a subset of the *interior* of Ω.
[This means that ϕ vanishes on (and in a neighbourhood of) $\partial\Omega$.]

Definition (Compact support in Ω)

A function $\phi \in C(\Omega)$ has *compact support* if

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

is compact (i.e. is bounded and closed) and is a subset of the *interior* of Ω. [This means that ϕ vanishes on (and in a neighbourhood of) $\partial\Omega$.]

Definition (Bump functions)

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

Definition (Compact support in Ω)

A function $\phi \in C(\Omega)$ has *compact support* if

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

is compact (i.e. is bounded and closed) and is a subset of the *interior* of Ω. [This means that ϕ vanishes on (and in a neighbourhood of) $\partial\Omega$.]

Definition (Bump functions)

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

Example

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

is in $C_0^\infty((-2, 2))$.

What set of functions might have weak derivatives?

Definition (Locally integrable functions)

Given a domain Ω, the set of *locally integrable functions* is defined by

$$L^1_{\mathsf{loc}}(\Omega) = \{f : \Omega \to \mathbb{R},\ f|_K \in L^1(K) \text{ for all compact } K \subset \text{ interior } \Omega\}.$$

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

Definition (Weak first derivative)

A function $f \in L^1_{\mathrm{loc}}(\Omega)$ has a weak i^{th} partial derivative $\partial f / \partial x_i$ if there exists a function $g \in L^1_{\mathrm{loc}}(\Omega)$ such that

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

Definition (Weak first derivative)

A function $f \in L^1_{\mathrm{loc}}(\Omega)$ has a weak i^{th} partial derivative $\partial f / \partial x_i$ if there exists a function $g \in L^1_{\mathrm{loc}}(\Omega)$ such that

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

With this, we can define the weak gradient, curl, and divergence in the obvious way (collect the relevant weak partial derivatives).

Definition (Weak first derivative)

A function $f \in L^1_{\mathsf{loc}}(\Omega)$ has a weak i^{th} partial derivative $\partial f/\partial x_i$ if there exists a function $g \in L^1_{\mathsf{loc}}(\Omega)$ such that

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

With this, we can define the weak gradient, curl, and divergence in the obvious way (collect the relevant weak partial derivatives).

Theorem

Weak derivatives are unique, up to values on a set of measure zero.

Definition (Weak first derivative)

A function $f \in L^1_{\text{loc}}(\Omega)$ has a weak i^{th} partial derivative $\partial f/\partial x_i$ if there exists a function $g \in L^1_{\text{loc}}(\Omega)$ such that

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

With this, we can define the weak gradient, curl, and divergence in the obvious way (collect the relevant weak partial derivatives).

Theorem

Weak derivatives are unique, up to values on a set of measure zero.

From now on, whenever I take a derivative, I mean a weak derivative!

Example

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

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

Example

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

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

To verify this, let $\phi \in C_0^\infty(\Omega)$. Then,

$$\begin{aligned}
\int_{-1}^{1} f(x)\phi'(x)\ \mathrm{d}x &= \int_{-1}^{0} f(x)\phi'(x)\ \mathrm{d}x + \int_{0}^{1} f(x)\phi'(x)\ \mathrm{d}x \\
&= -\int_{-1}^{0} (-1)\phi(x)\ \mathrm{d}x + [f\phi]_{-1}^{0} - \int_{0}^{1} (+1)\phi(x)\ \mathrm{d}x + [f\phi]_{0}^{1} \\
&= -\int_{-1}^{1} f'(x)\phi(x)\ \mathrm{d}x + \big((f\phi)(0^-) - (f\phi)(0^+)\big) \\
&= -\int_{-1}^{1} f'(x)\phi(x)\ \mathrm{d}x.
\end{aligned}$$

Example

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

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

More generally ...

Any *continuous piecewise-differentiable* function is weakly differentiable, because the boundary terms arising in integration by parts will cancel.

This is important because this is what we will use to approximate the solutions of PDEs!

Example

A counterexample: take $\Omega = (-1, 1)$ and take $f(x) = \mathrm{sign}(x)$, i.e.

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

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.

To define higher-order weak derivatives for functions of several variables, we first need to introduce multi-index notation.

Definition (multi-index notation)

Let $n \geq 1$. A multi-index α is a tuple of n non-negative integers

$$\alpha = (\alpha_1, \ldots, \alpha_n), \quad \alpha_i \in \mathbb{N}^+.$$

Let $\Omega \subset \mathbb{R}^n$. Given a multi-index α and $\phi \in C^\infty(\Omega)$, define

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

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

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

Example

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.

Definition (Weak derivative)

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

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

Section 2

Sobolev spaces

Definition (Sobolev norm)

Let $f \in L^1_{\mathsf{loc}}(\Omega)$. Let k be a non-negative integer.

Definition (Sobolev norm)

Let $f \in L^1_{\text{loc}}(\Omega)$. Let k be a non-negative integer. Suppose that the weak derivatives $D^\alpha f$ exist for all $|\alpha| \leq k$ and belong to $L^p(\Omega)$. The set of all such functions is denoted by $W^k_p(\Omega)$.

Definition (Sobolev norm)

Let $f \in L^1_{\mathrm{loc}}(\Omega)$. Let k be a non-negative integer. Suppose that the weak derivatives $D^\alpha f$ exist for all $|\alpha| \leq k$ and belong to $L^p(\Omega)$. The set of all such functions is denoted by $W^k_p(\Omega)$. For $p \in [1, \infty)$, define the *Sobolev norm*

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

Definition (Sobolev norm)

Let $f \in L^1_{\text{loc}}(\Omega)$. Let k be a non-negative integer. Suppose that the weak derivatives $D^\alpha f$ exist for all $|\alpha| \le k$ and belong to $L^p(\Omega)$. The set of all such functions is denoted by $W^k_p(\Omega)$. For $p \in [1, \infty)$, define the *Sobolev norm*

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

In the case $p = \infty$

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

Definition (Sobolev norm)

Let $f \in L^1_{\mathrm{loc}}(\Omega)$. Let k be a non-negative integer. Suppose that the weak derivatives $D^\alpha f$ exist for all $|\alpha| \le k$ and belong to $L^p(\Omega)$. The set of all such functions is denoted by $W^k_p(\Omega)$. For $p \in [1, \infty)$, define the *Sobolev norm*

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

In the case $p = \infty$

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

Definition (Sobolev space)

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

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

Theorem

The Sobolev space $W_p^k(\Omega)$ is a Banach space.

Proof.

See theorem 1.3.2 of Brenner & Scott. □

Theorem

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

$$H^k(\Omega) := W_2^k(\Omega).$$

Example

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

Example

Suppose $l \geq k$. Then $W_p^l(\Omega) \subset W_p^k(\Omega)$; we are just asking for fewer derivatives.

Example

Suppose $1 \leq p \leq q \leq \infty$ and that Ω is bounded. Then $W_q^k(\Omega) \subset W_p^k(\Omega)$.

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

Definition (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.

This regularity condition is important: without it, the Sobolev inequality *need not be 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;
- p, the integrability of the function and derivatives.

Sobolev's inequality asserts that if a function has sufficiently many weak derivatives that have a high enough exponent of integrability, then the function is continuous and bounded.

Theorem (Sobolev's inequality)

Let $p \in [1, \infty)$. Suppose that

$$k \geq n \text{ when } p = 1,$$
$$k > n/p \text{ when } p > 1.$$

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

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

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

Example

For $n = 1$, the existence of *a single* weak derivative of any integrability suffices to ensure continuity; $W_1^1(\Omega) \subset C(\overline{\Omega})$.

Example

For $n = 2$, we have $W_1^1(\Omega) \not\subset C(\Omega)$, but $W_1^2(\Omega) \subset C(\overline{\Omega})$.

Example

For $n = 3$, we have $W_1^2(\Omega) \not\subset C(\Omega)$, but $W_1^3(\Omega) \subset C(\overline{\Omega})$.

Example

Let us consider the continuity properties of functions in $H^k(\Omega) = W_2^k(\Omega)$, i.e. $p = 2$. With $p = 2$, Sobolev's inequality tells us that we need

$$k > n/2.$$

In one dimension,

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

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

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

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

We are now in a position to see why the space

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

is the "right" one for the variational formulation of the Poisson equation: find $u \in H_0^1(\Omega)$ such that

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

for all $v \in H_0^1(\Omega)$.

- We want $v \in L^2(\Omega)$ and $f \in L^2(\Omega)$ to ensure that the RHS is a bounded linear functional of v.
- We want $u|_{\partial\Omega} = 0$ to satisfy the boundary condition.
- We need the first weak derivatives to exist to talk about ∇u and ∇v.
- We want u and v to have square-integrable weak derivatives, as this guarantees that $a(u, v) < \infty$ (by Cauchy-Schwarz).

C6.4 Finite Element Methods for PDEs
Lecture 4: The Lax–Milgram Theorem

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

In this course, we will see three main theorems regarding the well-posedness of the linear variational problem: for $F \in V^*$,

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

of increasing generality:

In this course, we will see three main theorems regarding the well-posedness of the linear variational problem: for $F \in V^*$,

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

of increasing generality:

- Riesz Representation Theorem: a *bounded*, *coercive*, symmetric;

In this course, we will see three main theorems regarding the well-posedness of the linear variational problem: for $F \in V^*$,

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

of increasing generality:

- Riesz Representation Theorem: a *bounded*, *coercive*, symmetric;
- Lax–Milgram Theorem: a bounded, coercive;

In this course, we will see three main theorems regarding the well-posedness of the linear variational problem: for $F \in V^*$,

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

of increasing generality:

- Riesz Representation Theorem: a *bounded*, *coercive*, symmetric;
- Lax–Milgram Theorem: a bounded, coercive;
- Babuška's Theorem: a bounded, satisfies an inf-sup condition.

In this course, we will see three main theorems regarding the well-posedness of the linear variational problem: for $F \in V^*$,

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

of increasing generality:

- Riesz Representation Theorem: a *bounded*, *coercive*, symmetric;
- Lax–Milgram Theorem: a bounded, coercive;
- Babuška's Theorem: a bounded, satisfies an inf-sup condition.

In this lecture we will study the first two.

Definition (Bounded bilinear form)

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

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

Definition (Bounded bilinear form)

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

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

As with linear functionals, this is equivalent to continuity.

Definition (Bounded bilinear form)

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

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

As with linear functionals, this is equivalent to continuity.

The best (i.e. least) constant C satisfying the definition is called the *continuity constant* of a:

$$C := \sup_{\substack{v \in H \\ v \neq 0}} \sup_{\substack{w \in H \\ w \neq 0}} \frac{|a(v, w)|}{\|v\|_H\|w\|_H}.$$

Definition (Coercive bilinear form)

A bilinear form $a : H \times H \to \mathbb{R}$ is said be to *coercive* on $V \subset H$ or V-coercive if there exists an $\alpha > 0$ such that

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

Definition (Coercive bilinear form)

A bilinear form $a : H \times H \to \mathbb{R}$ is said be to *coercive* on $V \subset H$ or V-coercive if there exists an $\alpha > 0$ such that

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

This is stronger than a being positive-definite ($a(u, u) > 0$ for $u \neq 0$).

Example

Consider the space $H = \ell_2(\mathbb{R})$, the space of all square-summable infinite sequences of real numbers: $x = (x_1, x_2, \ldots)$. The bilinear form

$$a(x, y) = \sum_{m=1}^{\infty} 2^{-m} x_m y_m, \quad \text{with } x, y \in H,$$

is positive-definite but not coercive.

Definition (Coercive bilinear form)

A bilinear form $a : H \times H \to \mathbb{R}$ is said be to *coercive* on $V \subset H$ or V-coercive if there exists an $\alpha > 0$ such that

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

The best (i.e. largest in this case) constant α satisfying the definition is called the coercivity constant of a:

$$\alpha := \inf_{\substack{u \in V \\ u \neq 0}} \frac{a(u, u)}{\|u\|_H^2}.$$

Note that we must have $\alpha \leq C$, as

$$\alpha \|u\|_H^2 \leq a(u, u) \leq C\|u\|_H^2.$$

Let us assume for now that a is also symmetric.

Theorem

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 $(V, \|\cdot\|_a)$ is a Hilbert space with the norm $\|\cdot\|_a$ defined by $\|v\|_a := \sqrt{a(v,v)}$.

Let us assume for now that a is also symmetric.

Theorem

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 $(V, \|\cdot\|_a)$ is a Hilbert space with the norm $\|\cdot\|_a$ defined by $\|v\|_a := \sqrt{a(v,v)}$.

We must prove that $a(\cdot,\cdot)$ is an inner product on V, and that V is complete with respect to the induced norm $\|\cdot\|_a$.

Clearly $a(v, v) \geq 0$ for all $v \in V$. If $0 = a(v, v) \geq \alpha \|v\|_H^2 \geq 0$, then $v = 0$. Symmetry and linearity are assumed, so $a(\cdot, \cdot)$ is an inner product on V.

Denote

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

It remains to show that $(V, \|\cdot\|_a)$ is complete.

Suppose that $\{v_n\}$ is a Cauchy sequence in $(V, \|\cdot\|_a)$, i.e.

$$(\forall \varepsilon > 0)\ (\exists\, N > 0)\ (\forall\, m, n > N) \quad \|v_n - v_m\|_a < \varepsilon.$$

Since $\|v\|_H \leq \frac{1}{\sqrt{\alpha}}\|v\|_a$, $\|v_n - v_m\|_H < \varepsilon/\sqrt{\alpha}$ and $\{v_n\}$ is therefore also Cauchy in $(H, \|\cdot\|_H)$.

Since H is complete, there exists a $v \in H$ 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

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

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

Faster: note that coercivity and continuity guarantee that

$$\alpha\|v\|_H^2 \leq \|v\|_a^2 \leq C\|v\|_H^2 \quad \text{for all } v \in V.$$

So the norms are equivalent, and hence induce the same notion of convergence and completeness.

The well-posedness of the symmetric coercive bounded linear variational problem follows immediately.

Theorem

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:

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

This problem has a unique stable solution.

The well-posedness of the symmetric coercive bounded linear variational problem follows immediately.

Theorem

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:

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

This problem has a unique stable solution.

Proof.

Our previous result implies that $a(\cdot, \cdot)$ is an inner product on V, and that $(V, \|\cdot\|_a)$ is a Hilbert space. Apply the Riesz Representation Theorem, that every bounded linear functional $F \in V^*$ has a unique representative (in this case u).

Proof.

Stability means that there exists a constant c s.t, with $\|\cdot\|_V := \|\cdot\|_a$,

$$\|u\|_V \leq c\|F\|_{V^*}.$$

This follows by noting that

$$\|u\|_a^2 = a(u,u) = F(u) = \frac{F(u)}{\|u\|_a}\|u\|_a \leq \left(\sup_{v\in V}\frac{F(v)}{\|v\|_a}\right)\|u\|_a = \|F\|_{V^*}\|u\|_a.$$

Thus $\|u\|_V = \|u\|_a \leq \|F\|_{V^*}$ (i.e. the stability constant is $c = 1$). $\square$

Example

The variational problem

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

is well-posed, as $H_0^1(\Omega)$ is a closed subspace of $H^1(\Omega)$, and we will show later that the bilinear form is $H_0^1(\Omega)$-coercive, symmetric, and bounded.

Section 3

The nonsymmetric case

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

Theorem (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:

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

This problem has a unique stable solution.

For the proof, it will be more convenient to treat this linear variational problem as an equation in the dual V^*.

Lemma

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 $A : u \mapsto Au$

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

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

For the proof, it will be more convenient to treat this linear variational problem as an equation in the dual V^*.

Lemma

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 $A : u \mapsto Au$

$$(Au)(v) := a(u,v) \quad \textit{for all } v \in V.$$

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

Proof.

Linearity is straightforward. For boundedness (so that $Au \in V^*$),

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

$\square$

Thus, the variational problem

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

is equivalent to

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

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

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

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

Example

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

$$-\nabla^2 u = f$$

as an equation in the dual of $H_0^1(\Omega)$. This dual space is denoted

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

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

We know from the Riesz Representation Theorem that there is an isometric isomorphism $\mathcal{R} : V^* \to V$ from the dual of a Hilbert space V^* back to V. By composing these operators, we have the problem

$$\text{find } u \in V \text{ such that } \mathcal{R}Au = \mathcal{R}F,$$

where the equality is between *primal* objects, $\mathcal{R}Au \in V$ and $\mathcal{R}F \in V$.

Proof strategy: 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.

Theorem (Contraction mapping theorem)

Given a nonempty Banach space V and a mapping $T : V \to V$ satisfying

$$\|Tv_1 - Tv_2\| \leq M\|v_1 - v_2\|$$

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

$$u = Tu.$$

That is, a contraction $T : V \to V$ on a Banach space V has a unique fixed point $u \in V$.

We now prove the Lax–Milgram Theorem.

Proof.

Cast the variational problem

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

as the primal equality

$$\text{find } u \in V \text{ such that } \mathcal{R}Au = \mathcal{R}F$$

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

$$Tv = v - \rho\,(\mathcal{R}Av - \mathcal{R}F)\,.$$

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

$$Tu = u - \rho(\mathcal{R}Au - \mathcal{R}F) = u,$$

i.e. that $\mathcal{R}Au = \mathcal{R}F$. We now show that such a ρ exists.

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$\|Tv_1 - Tv_2\|_H^2 = \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2$$

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2 \\
&= \|v - \rho(\mathcal{R}Av)\|_H^2 && \text{(lin. of } \mathcal{R}, A\text{)}
\end{aligned}$$

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$
\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2 \\
&= \|v - \rho(\mathcal{R}Av)\|_H^2 && \text{(lin. of } \mathcal{R}, A) \\
&= \|v\|_H^2 - 2\rho(\mathcal{R}Av, v) + \rho^2\|\mathcal{R}Av\|_H^2 && \text{(lin. of i. prod.)}
\end{aligned}
$$

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2 \\
&= \|v - \rho(\mathcal{R}Av)\|_H^2 && \text{(lin. of } \mathcal{R},\ A) \\
&= \|v\|_H^2 - 2\rho(\mathcal{R}Av, v) + \rho^2\|\mathcal{R}Av\|_H^2 && \text{(lin. of i. prod.)} \\
&= \|v\|_H^2 - 2\rho Av(v) + \rho^2 Av(\mathcal{R}Av) && \text{(definition of } \mathcal{R})
\end{aligned}$$

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$
\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2 \\
&= \|v - \rho(\mathcal{R}Av)\|_H^2 && \text{(lin. of } \mathcal{R},\ A\text{)} \\
&= \|v\|_H^2 - 2\rho(\mathcal{R}Av, v) + \rho^2\|\mathcal{R}Av\|_H^2 && \text{(lin. of i. prod.)} \\
&= \|v\|_H^2 - 2\rho Av(v) + \rho^2 Av(\mathcal{R}Av) && \text{(definition of } \mathcal{R}\text{)} \\
&= \|v\|_H^2 - 2\rho a(v, v) + \rho^2 a(v, \mathcal{R}Av) && \text{(definition of } A\text{)}
\end{aligned}
$$

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$
\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2 \\
&= \|v - \rho(\mathcal{R}Av)\|_H^2 && \text{(lin. of } \mathcal{R},\ A\text{)} \\
&= \|v\|_H^2 - 2\rho(\mathcal{R}Av, v) + \rho^2\|\mathcal{R}Av\|_H^2 && \text{(lin. of i. prod.)} \\
&= \|v\|_H^2 - 2\rho Av(v) + \rho^2 Av(\mathcal{R}Av) && \text{(definition of } \mathcal{R}\text{)} \\
&= \|v\|_H^2 - 2\rho a(v, v) + \rho^2 a(v, \mathcal{R}Av) && \text{(definition of } A\text{)} \\
&\le \|v\|_H^2 - 2\rho\alpha\|v\|_H^2 + \rho^2 C\|v\|_H\|\mathcal{R}Av\|_H && \text{(coerc. \& cont.)}
\end{aligned}
$$

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2 \\
&= \|v - \rho(\mathcal{R}Av)\|_H^2 && \text{(lin. of } \mathcal{R},\ A) \\
&= \|v\|_H^2 - 2\rho(\mathcal{R}Av, v) + \rho^2\|\mathcal{R}Av\|_H^2 && \text{(lin. of i. prod.)} \\
&= \|v\|_H^2 - 2\rho Av(v) + \rho^2 Av(\mathcal{R}Av) && \text{(definition of } \mathcal{R}) \\
&= \|v\|_H^2 - 2\rho a(v, v) + \rho^2 a(v, \mathcal{R}Av) && \text{(definition of } A) \\
&\leq \|v\|_H^2 - 2\rho\alpha\|v\|_H^2 + \rho^2 C\|v\|_H\|\mathcal{R}Av\|_H && \text{(coerc. \& cont.)} \\
&\leq (1 - 2\rho\alpha + \rho^2 C^2)\|v\|_H^2 && (A \text{ cts, } \mathcal{R} \text{ isom.})
\end{aligned}$$

Proof.

For any $v_1, v_2 \in V$, let $v = v_1 - v_2$. Then

$$
\begin{aligned}
\|Tv_1 - Tv_2\|_H^2 &= \|v_1 - v_2 - \rho(\mathcal{R}Av_1 - \mathcal{R}Av_2)\|_H^2 \\
&= \|v - \rho(\mathcal{R}Av)\|_H^2 && \text{(lin. of } \mathcal{R},\ A\text{)} \\
&= \|v\|_H^2 - 2\rho(\mathcal{R}Av, v) + \rho^2\|\mathcal{R}Av\|_H^2 && \text{(lin. of i. prod.)} \\
&= \|v\|_H^2 - 2\rho Av(v) + \rho^2 Av(\mathcal{R}Av) && \text{(definition of } \mathcal{R}\text{)} \\
&= \|v\|_H^2 - 2\rho a(v, v) + \rho^2 a(v, \mathcal{R}Av) && \text{(definition of } A\text{)} \\
&\leq \|v\|_H^2 - 2\rho\alpha\|v\|_H^2 + \rho^2 C\|v\|_H\|\mathcal{R}Av\|_H && \text{(coerc. \& cont.)} \\
&\leq (1 - 2\rho\alpha + \rho^2 C^2)\|v\|_H^2 && (A \text{ cts, } \mathcal{R} \text{ isom.)} \\
&= (1 - 2\rho\alpha + \rho^2 C^2)\|v_1 - v_2\|_H^2.
\end{aligned}
$$

Proof.

Thus, if we can find a ρ such that

$$1 - 2\rho\alpha + \rho^2 C^2 < 1,$$

i.e. that

$$\rho(\rho C^2 - 2\alpha) < 0,$$

then we are done. If we choose $\rho \in (0, 2\alpha/C^2)$ then T is a contraction and a unique solution exists.

Proof.

It remains to show stability.

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

and so

$$\|u\|_H \leq \frac{1}{\alpha} \|F\|_{V^*}$$

□

C6.4 Finite Element Methods for PDEs
Lecture 5: More variational formulations

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

This lecture has three goals:

- Study the variational formulation of more problems;
- Look at how to prove well-posedness with the Lax–Milgram theorem;
- See some problems our current theory cannot handle.

We consider the Poisson problem in one dimension with mixed boundary conditions:

$$-u'' = f, \quad u(0) = 0,\ u'(1) = g.$$

Let us investigate how we can use the Lax–Milgram theorem for this.

We consider the Poisson problem in one dimension with mixed boundary conditions:

$$-u'' = f, \quad u(0) = 0,\ u'(1) = g.$$

Let us investigate how we can use the Lax–Milgram theorem for this.

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

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

We consider the Poisson problem in one dimension with mixed boundary conditions:

$$-u'' = f, \quad u(0) = 0,\ u'(1) = g.$$

Let us investigate how we can use the Lax–Milgram theorem for this.

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

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

and integrating again from 0 to x yields

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

We consider the Poisson problem in one dimension with mixed boundary conditions:

$$-u'' = f, \quad u(0) = 0, \ u'(1) = g.$$

Let us investigate how we can use the Lax–Milgram theorem for this.

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

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

and integrating again from 0 to x yields

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

This shows that the problem has a solution. In fact the solution is unique.

Now consider a variational formulation. We define the space

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

This makes sense, because in one dimension H^1 functions are continuous.

Now consider a variational formulation. We define the space

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

This makes sense, because in one dimension H^1 functions are continuous.

Two kinds of boundary conditions:

- The Dirichlet condition $u(0) = 0$ is carried in the definition of the space (strongly enforced).

Now consider a variational formulation. We define the space

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

This makes sense, because in one dimension H^1 functions are continuous.

Two kinds of boundary conditions:

- The Dirichlet condition $u(0) = 0$ is carried in the definition of the space (strongly enforced).
- The Neumann condition $u'(1) = g$ will appear in the variational formulation (weakly enforced).

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

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

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

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

We next integrate by parts:

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

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

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

We next integrate by parts:

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

The third term on the left disappears as $v(0) = 0$. The second term on the left is moved to the right-hand side, and we use that $u'(1) = g$:

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

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

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

We next integrate by parts:

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

The third term on the left disappears as $v(0) = 0$. The second term on the left is moved to the right-hand side, and we use that $u'(1) = g$:

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

Thus, we have a linear variational problem with

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

Theorem

The following bilinear form is coercive on V:

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

Theorem

The following bilinear form is coercive on V:

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

Proof.

The norm on $H^1((0,1))$ is

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

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

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

Expanding definitions, we want to find an α such that

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

Proof.

If we can prove that there exists an $\alpha' > 0$ such that

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

Proof.

If we can prove that there exists an $\alpha' > 0$ such that

$$\begin{aligned} &\|v'\|^2_{L^2((0,1))} \geq \alpha' \|v\|^2_{L^2((0,1))} \\ \Longleftrightarrow\ &\|v'\|^2_{L^2((0,1))} + \alpha' \|v'\|^2_{L^2((0,1))} \geq \alpha' \left(\|v\|^2_{L^2((0,1))} + \|v'\|^2_{L^2((0,1))} \right) \end{aligned}$$

Proof.

If we can prove that there exists an $\alpha' > 0$ such that

$$\begin{aligned} & \|v'\|^2_{L^2((0,1))} \geq \alpha' \|v\|^2_{L^2((0,1))} \\ \Longleftrightarrow\ & \|v'\|^2_{L^2((0,1))} + \alpha' \|v'\|^2_{L^2((0,1))} \geq \alpha' \left(\|v\|^2_{L^2((0,1))} + \|v'\|^2_{L^2((0,1))} \right) \\ \Longleftrightarrow\ & \|v'\|^2_{L^2((0,1))} \geq \frac{\alpha'}{\alpha' + 1} \|v\|^2_{H^1((0,1))} \end{aligned}$$

Proof.

If we can prove that there exists an $\alpha' > 0$ such that

$$
\begin{aligned}
& \|v'\|^2_{L^2((0,1))} \geq \alpha' \|v\|^2_{L^2((0,1))} \\
\iff & \|v'\|^2_{L^2((0,1))} + \alpha' \|v'\|^2_{L^2((0,1))} \geq \alpha' \left(\|v\|^2_{L^2((0,1))} + \|v'\|^2_{L^2((0,1))} \right) \\
\iff & \|v'\|^2_{L^2((0,1))} \geq \frac{\alpha'}{\alpha' + 1} \|v\|^2_{H^1((0,1))}
\end{aligned}
$$

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

Proof.

We can write, for $t \in [0,1]$,

$$v(t) = \int_0^t 1 \cdot v'(x) \, \mathrm{d}x$$

and apply the Cauchy–Schwarz inequality to deduce that

$$|v(t)|^2 \leq \left(\int_0^t 1^2 \, \mathrm{d}x\right)\left(\int_0^t |v'(x)|^2 \, \mathrm{d}x\right) \leq t\|v'\|^2_{L^2((0,1))}.$$

Hence,

$$\|v\|^2_{L^2((0,1))} = \int_0^1 |v(t)|^2 \, \mathrm{d}t \leq \frac{1}{2}\|v'\|^2_{L^2((0,1))},$$

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

Two remarks:

Remark

Note that if we consider a over the whole of $H^1((0,1))$, it is *not* coercive: $v(x) \equiv 1 \in H^1((0,1))$ with $a(v,v) = 0$ but $\|v\|_{H^1((0,1))} = 1 > 0$.

The boundary condition $v(0) = 0$ is essential to the coercivity.

Two remarks:

Remark

Note that if we consider a over the whole of $H^1((0,1))$, it is *not* coercive: $v(x) \equiv 1 \in H^1((0,1))$ with $a(v,v) = 0$ but $\|v\|_{H^1((0,1))} = 1 > 0$.

The boundary condition $v(0) = 0$ is essential to the coercivity.

Remark

Notice 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

The following bilinear form is continuous:

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

Theorem

The following bilinear form is continuous:

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

Proof.

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

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

Theorem

The following bilinear form is continuous:

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

Proof.

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

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

The Lax–Milgram theorem then implies the existence of a unique weak solution to this boundary-value problem, which depends continuously on the data (i.e. the right-hand side $f \in L^2((0,1))$).

Now consider a nonsymmetric problem: for $f \in L^2((0,1))$, solve

$$-u'' + u' + u = f, \quad u'(0) = 0 = u'(1).$$

Now consider a nonsymmetric problem: for $f \in L^2((0,1))$, solve

$$-u'' + u' + u = f, \quad u'(0) = 0 = u'(1).$$

Since we have no Dirichlet conditions, we set $V = H^1((0,1))$.

Now consider a nonsymmetric problem: for $f \in L^2((0,1))$, solve

$$-u'' + u' + u = f, \quad u'(0) = 0 = u'(1).$$

Since we have no Dirichlet conditions, we set $V = H^1((0,1))$.

Testing against $v \in V$ and integrating by parts, we find

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

Thus, our standard variational problem has

$$a(u,v) = \int_0^1 u'v' + u'v + uv \, \mathrm{d}x, \quad F(v) = \int_0^1 fv \, \mathrm{d}x.$$

To prove continuity, observe that

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

so we can take our continuity constant to be $C = 2$.

To prove continuity, observe that

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

so we can take our continuity constant to be $C = 2$.

To prove coercivity, observe that

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

Section 2

Higher dimensions

Break up $\partial\Omega$ into disjoint Γ_D and Γ_N. Consider

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

Break up $\partial\Omega$ into disjoint Γ_D and Γ_N. Consider

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

Define the space

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

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

$$\begin{aligned} \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x &= \int_\Omega f v \, \mathrm{d}x + \int_{\partial\Omega} \nabla u \cdot n v \, \mathrm{d}s, \\ &= \int_\Omega f v \, \mathrm{d}x + \int_{\Gamma_N} g v \, \mathrm{d}s. \end{aligned}$$

The continuity proof works as in one dimension. For coercivity, we need a result from the theory of function spaces.

The continuity proof works as in one dimension. For coercivity, we need a result from the theory of function spaces.

Theorem (Poincaré–Friedrichs inequality)

Let Ω be a bounded Lipschitz domain, and suppose $\Gamma_D \subset \partial\Omega$ is closed and has nonzero measure. Let

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

Then there is a constant $K > 0$ depending only on Ω and Γ_D such that

$$\|u\|^2_{L^2(\Omega)} = \int_\Omega u^2 \,\mathrm{d}x \leq K \int_\Omega |\nabla u|^2 \,\mathrm{d}x = K\|\nabla u\|^2_{L^2(\Omega)}$$

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

Write

$$|u|_{H^1(\Omega)}^2 := \int_\Omega |\nabla u|^2 \, \mathrm{d}x = \|\nabla u\|_{L^2(\Omega)}^2.$$

In general this is a *seminorm*: $|u|_{H^1(\Omega)} = 0 \;\not\Longrightarrow\; u = 0$.

Write

$$|u|^2_{H^1(\Omega)} := \int_\Omega |\nabla u|^2 \, \mathrm{d}x = \|\nabla u\|^2_{L^2(\Omega)}.$$

In general this is a *seminorm*: $|u|_{H^1(\Omega)} = 0 \not\Rightarrow u = 0$.

On

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

manipulating the Poincaré–Friedrichs inequality yields

$$\frac{1}{K+1}\|u\|^2_{H^1(\Omega)} \le |u|^2_{H^1(\Omega)} \le \|u\|^2_{H^1(\Omega)}.$$

So if one has a Dirichlet condition on a part Γ_D of the boundary $\partial\Omega$, then $|u|_{H^1(\Omega)}$ is an *equivalent norm* to $\|u\|_{H^1(\Omega)}$.

Write

$$|u|_{H^1(\Omega)}^2 := \int_\Omega |\nabla u|^2 \, \mathrm{d}x = \|\nabla u\|_{L^2(\Omega)}^2.$$

In general this is a *seminorm*: $|u|_{H^1(\Omega)} = 0 \not\Rightarrow u = 0$.

On

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

manipulating the Poincaré–Friedrichs inequality yields

$$\frac{1}{K+1}\|u\|_{H^1(\Omega)}^2 \leq |u|_{H^1(\Omega)}^2 \leq \|u\|_{H^1(\Omega)}^2.$$

So if one has a Dirichlet condition on a part Γ_D of the boundary $\partial\Omega$, then $|u|_{H^1(\Omega)}$ is an *equivalent norm* to $\|u\|_{H^1(\Omega)}$.

Remark

If Ω is contained within an n-dimensional cube of side L, then L provides a (possibly non-optimal) Poincaré constant. (Braess, 2007)

What happens in the case of an inhomogeneous Dirichlet condition?

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

What happens in the case of an inhomogeneous Dirichlet condition?

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

Consider

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

Then $\hat{u}$ satisfies

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

or variationally

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

for all $v \in V$.

What about a pure Neumann condition?

$$-\nabla^2 u = f \text{ in } \Omega$$
$$\nabla u \cdot n = g \text{ on } \partial\Omega.$$

What about a pure Neumann condition?

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

If u satisfies the equation, so does $u + c$, $c \in \mathbb{R}$! Lack of uniqueness!

What about a pure Neumann condition?

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

If u satisfies the equation, so does $u + c$, $c \in \mathbb{R}$! Lack of uniqueness!

Existence: for a solution to exist f and g must be compatible:

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

If we take the variational formulation with $V = H^1(\Omega)$, we get

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x + \int_{\partial\Omega} g v \, \mathrm{d}s,$$

We cannot apply Lax–Milgram because the bilinear form is not V-coercive.

If we take the variational formulation with $V = H^1(\Omega)$, we get

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x + \int_{\partial\Omega} g v \, \mathrm{d}s,$$

We cannot apply Lax–Milgram because the bilinear form is not V-coercive.

To eliminate the nullspace of constants, consider the solution space that is L^2-orthogonal to it:

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

The bilinear form *is* coercive over V by the following Poincaré–Wirtinger inequality: there exists a positive constant C, such that for all $v \in H^1(\Omega)$,

$$\left\| v - \frac{1}{|\Omega|} \int_\Omega v \, \mathrm{d}x \right\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)}.$$

Let us briefly consider two other kinds of boundary conditions. Robin conditions relate $\nabla u \cdot n$ and u:

$$-\nabla^2 u = f \text{ in } \Omega,$$
$$\nabla u \cdot n + \beta u = g \text{ on } \partial\Omega.$$

Taking $V = H^1(\Omega)$, this yields the variational problem

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \beta \int_{\partial\Omega} uv \, \mathrm{d}s = \int_\Omega fv \, \mathrm{d}x + \int_{\partial\Omega} gv \, \mathrm{d}s.$$

Proving the properties required for the application of the Lax–Milgram theorem requires knowledge of trace theorems, which we haven't discussed.

At this point, one may get the impression that there is only one variational formulation for a problem. This is not true, and different formulations have advantages and disadvantages.

At this point, one may get the impression that there is only one variational formulation for a problem. This is not true, and different formulations have advantages and disadvantages.

Suppose we want to know the *flux* in the Poisson equation accurately. We can solve the *mixed* formulation: find $\sigma : \Omega \to \mathbb{R}^n$, $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} \sigma &= -\nabla u \text{ in } \Omega, \\ \nabla \cdot \sigma &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned}$$

Solving this problem will give an accurate approximation of the flux, and allow for the easy implementation of more complicated constitutive laws.

$$\sigma = -\nabla u \text{ in } \Omega, \quad \nabla \cdot \sigma = f \text{ in } \Omega.$$

Let us multiply the first equation by a vector-valued test function v, and the second by a scalar-valued function w:

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x + \int_\Omega \nabla u \cdot v = 0,$$
$$\int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = \int_\Omega f w \, \mathrm{d}x.$$

$$\sigma = -\nabla u \text{ in } \Omega, \quad \nabla \cdot \sigma = f \text{ in } \Omega.$$

Let us multiply the first equation by a vector-valued test function v, and the second by a scalar-valued function w:

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x + \int_\Omega \nabla u \cdot v = 0,$$
$$\int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = \int_\Omega f w \, \mathrm{d}x.$$

Since σ needs to have a divergence, and we want v and σ to come from the same space, we integrate by parts in the first equation. For symmetry, we negate the second equation:

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega u \nabla \cdot v + \int_{\partial\Omega} u(v \cdot n) \, \mathrm{d}s = 0,$$
$$-\int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = -\int_\Omega f w \, \mathrm{d}x.$$

$$\sigma = -\nabla u \text{ in } \Omega, \quad \nabla \cdot \sigma = f \text{ in } \Omega.$$

Let us multiply the first equation by a vector-valued test function v, and the second by a scalar-valued function w:

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x + \int_\Omega \nabla u \cdot v = 0,$$
$$\int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = \int_\Omega f w \, \mathrm{d}x.$$

Since σ needs to have a divergence, and we want v and σ to come from the same space, we integrate by parts in the first equation. For symmetry, we negate the second equation:

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega u \nabla \cdot v + \int_{\partial\Omega} u(v \cdot n) \, \mathrm{d}s = 0,$$
$$-\int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = -\int_\Omega f w \, \mathrm{d}x.$$

Impose the Dirichlet condition weakly by omitting the boundary integral.

What function spaces do we need to make sense of the resulting problem:

$$\int_\Omega \sigma \cdot v \ \mathrm{d}x - \int_\Omega u \nabla \cdot v = 0,$$
$$-\int_\Omega (\nabla \cdot \sigma) w \ \mathrm{d}x = -\int_\Omega f w \ \mathrm{d}x.$$

We do not need any derivatives on u or w, so $u \in L^2(\Omega)$.

What function spaces do we need to make sense of the resulting problem:

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega u \nabla \cdot v = 0,$$
$$-\int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = -\int_\Omega f w \, \mathrm{d}x.$$

We do not need any derivatives on u or w, so $u \in L^2(\Omega)$.

For σ and v, we need $\sigma \in L^2(\Omega; \mathbb{R}^n)$ and for $\nabla \cdot \sigma \in L^2(\Omega)$. This is the space $H(\mathrm{div}, \Omega)$:

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

What function spaces do we need to make sense of the resulting problem:

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega u \nabla \cdot v = 0,$$
$$-\int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = -\int_\Omega f w \, \mathrm{d}x.$$

We do not need any derivatives on u or w, so $u \in L^2(\Omega)$.

For σ and v, we need $\sigma \in L^2(\Omega; \mathbb{R}^n)$ and for $\nabla \cdot \sigma \in L^2(\Omega)$. This is the space $H(\mathrm{div}, \Omega)$:

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

Its inner product is

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

A nice property of this variational problem is that we can add the two equations together. The problem is the same as:

Find $(\sigma, u) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$B(\sigma, u; v, w) := \int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega (\nabla \cdot v) u - \int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = -\int_\Omega f w \, \mathrm{d}x$$

for all $(v, w) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$.

A nice property of this variational problem is that we can add the two equations together. The problem is the same as:

Find $(\sigma, u) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$B(\sigma, u; v, w) := \int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega (\nabla \cdot v) u - \int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = -\int_\Omega f w \, \mathrm{d}x$$

for all $(v, w) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$.

The Lax–Milgram theorem will certainly not apply, as coercivity fails:

$$B(0, u; 0, u) = 0 \text{ for all } u \in L^2(\Omega).$$

A nice property of this variational problem is that we can add the two equations together. The problem is the same as:

Find $(\sigma, u) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$B(\sigma, u; v, w) := \int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega (\nabla \cdot v) u - \int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = - \int_\Omega f w \, \mathrm{d}x$$

for all $(v, w) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$.

The Lax–Milgram theorem will certainly not apply, as coercivity fails:

$$B(0, u; 0, u) = 0 \text{ for all } u \in L^2(\Omega).$$

We will study the well-posedness of this problem with a more general theory later.

C6.4 Finite Element Methods for PDEs
Lecture 6: Differentiation and energy minimisation

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

In this lecture we will see a fundamental connection between symmetric linear variational problems:

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

and energy minimisation:

$$u = \underset{v \in V}{\operatorname{argmin}}\ J(v).$$

In this lecture we will see a fundamental connection between symmetric linear variational problems:

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

and energy minimisation:

$$u = \underset{v \in V}{\operatorname{argmin}}\ J(v).$$

This is one reason why the variational formulation is so useful, and will also lead to an insight into Galerkin approximation.

We use derivatives of a function at a point u to give us a *local model* of how it behaves in the neighbourhood of u. We want to do the same in Banach spaces.

We use derivatives of a function at a point u to give us a *local model* of how it behaves in the neighbourhood of u. We want to do the same in Banach spaces.

Given a function $J : V \to W$, V, W Banach spaces, how can we differentiate J? As a concrete example, think of $J : H_0^1(\Omega) \to \mathbb{R}$ defined by

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

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

We use derivatives of a function at a point u to give us a *local model* of how it behaves in the neighbourhood of u. We want to do the same in Banach spaces.

Given a function $J : V \to W$, V, W Banach spaces, how can we differentiate J? As a concrete example, think of $J : H_0^1(\Omega) \to \mathbb{R}$ defined by

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

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

We will introduce progressively stronger notions of differentiation for this.

Definition (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

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

if the limit exists.

Definition (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

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

if the limit exists.

Definition (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.

We will want more than just all directional derivatives exist. We will want

- that the derivative is *linear and bounded* in the perturbation direction (Gâteaux);
- that the derivative is *a good approximation* of the function nearby (Fréchet).

We will want more than just all directional derivatives exist. We will want

- that the derivative is *linear and bounded* in the perturbation direction (Gâteaux);
- that the derivative is *a good approximation* of the function nearby (Fréchet).

Definition (Gâteaux differentiable)

If J is directionally differentiable at u, and there exists a bounded linear map $J'(u) : V \to W$ such that

$$J'(u; v) = J'(u)v,$$

then J is Gâteaux differentiable at u with derivative $J'(u)$.

We will want more than just all directional derivatives exist. We will want

- that the derivative is *linear and bounded* in the perturbation direction (Gâteaux);
- that the derivative is *a good approximation* of the function nearby (Fréchet).

Definition (Gâteaux differentiable)

If J is directionally differentiable at u, and there exists a bounded linear map $J'(u) : V \to W$ such that

$$J'(u; v) = J'(u)v,$$

then J is Gâteaux differentiable at u with derivative $J'(u)$.

Example

If $W = \mathbb{R}$, then $J'(u) \in V^*$ for each $u \in V$.

Definition (Fréchet differentiable)

Suppose $J : V \to W$ is Gâteaux differentiable at a point $u \in V$ and that the derivative J' satisfies

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

Then J is Fréchet differentiable at u.

Definition (Fréchet differentiable)

Suppose $J : V \to W$ is Gâteaux differentiable at a point $u \in V$ and that the derivative J' satisfies

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

Then J is Fréchet differentiable at u.

This allows us to approximate what J does near u.

With our example

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

let us calculate $J'(u;v)$ for given $u, v \in H_0^1(\Omega)$.

With our example

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

let us calculate $J'(u;v)$ for given $u, v \in H_0^1(\Omega)$.

$$J'(u;v) = \lim_{\varepsilon\to 0^+} \frac{1}{2\varepsilon}\int_\Omega |\nabla u + \varepsilon\nabla v|^2 - |\nabla u|^2 \,\mathrm{d}x - \frac{1}{\varepsilon}\int_\Omega f(u + \varepsilon v - u)\,\mathrm{d}x$$

With our example

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

let us calculate $J'(u;v)$ for given $u, v \in H^1_0(\Omega)$.

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

With our example

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

let us calculate $J'(u;v)$ for given $u, v \in H_0^1(\Omega)$.

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

With our example

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

let us calculate $J'(u;v)$ for given $u, v \in H_0^1(\Omega)$.

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

With our example

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

let us calculate $J'(u;v)$ for given $u, v \in H_0^1(\Omega)$.

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

So if $J'(u;v) = 0$ for all $v \in H_0^1(\Omega)$, u satisfies the weak formulation of the Poisson equation!

For symmetric coercive problems, there is a strong relationship between linear variational problems (LVPs) and minimisation.

Theorem (Energy minimisation)

Suppose a is symmetric, coercive, and bounded, and $F \in V^*$. Let u be the unique solution to

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

Then u is the unique solution to

$$u = \operatorname*{argmin}_{v \in V} J(v) := \frac{1}{2} a(v,v) - F(v).$$

Proof.

Let $v \in V$. We want to show $J(v) \geq J(u)$. Calculating,

$$J(v) - J(u) = \frac{1}{2}a(v, v) - F(v) - \frac{1}{2}a(u, u) + F(u)$$

□

Proof.

Let $v \in V$. We want to show $J(v) \geq J(u)$. Calculating,

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2}a(v,v) - F(v) - \frac{1}{2}a(u,u) + F(u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - F(v-u) \end{aligned}$$

□

Proof.

Let $v \in V$. We want to show $J(v) \geq J(u)$. Calculating,

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2}a(v,v) - F(v) - \frac{1}{2}a(u,u) + F(u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - F(v-u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - a(u, v-u) \end{aligned}$$

$\square$

Proof.

Let $v \in V$. We want to show $J(v) \geq J(u)$. Calculating,

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2}a(v,v) - F(v) - \frac{1}{2}a(u,u) + F(u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - F(v-u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - a(u,v-u) \\ &= \frac{1}{2}a(v-u,v-u). \end{aligned}$$

□

Proof.

Let $v \in V$. We want to show $J(v) \geq J(u)$. Calculating,

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2}a(v,v) - F(v) - \frac{1}{2}a(u,u) + F(u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - F(v-u) \\ &= \frac{1}{2}a(v,v) - \frac{1}{2}a(u,u) - a(u,v-u) \\ &= \frac{1}{2}a(v-u,v-u). \end{aligned}$$

Because a is coercive,

$$J(v) - J(u) \geq \frac{\alpha}{2}\|v-u\|_V^2 \geq 0 \text{ for all } v \in V.$$

$\square$

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

$$J(\tilde{u}) - J(u) = 0 \geq \frac{\alpha}{2}\|\tilde{u} - u\|_V^2 \geq 0$$

and hence $\tilde{u} = u$.

What about the other way around?

Theorem

Let $u \in V$ be a minimiser of $J : V \to \mathbb{R}$. Then u is a solution of

$$J'(u; v) = 0 \text{ for all } v \in V.$$

What about the other way around?

Theorem

Let $u \in V$ be a minimiser of $J : V \to \mathbb{R}$. Then u is a solution of

$$J'(u; v) = 0 \text{ for all } v \in V.$$

Proof.

Since u is a minimiser of J, $J(u + \varepsilon v) \geq J(u)$ for all $\varepsilon > 0, v \in V$. This implies

$$\frac{J(u + \varepsilon v) - J(u)}{\varepsilon} \geq 0.$$

Taking the limit as $\varepsilon \to 0^+$ yields $J'(u; v) \geq 0$ for all $v \in V$.

What about the other way around?

Theorem

Let $u \in V$ be a minimiser of $J : V \to \mathbb{R}$. Then u is a solution of

$$J'(u; v) = 0 \text{ for all } v \in V.$$

Proof.

Since u is a minimiser of J, $J(u + \varepsilon v) \geq J(u)$ for all $\varepsilon > 0, v \in V$. This implies

$$\frac{J(u + \varepsilon v) - J(u)}{\varepsilon} \geq 0.$$

Taking the limit as $\varepsilon \to 0^+$ yields $J'(u; v) \geq 0$ for all $v \in V$.

Replacing v with $-v$, we have $J'(u; -v) = -J'(u; v) \geq 0$,
i.e. $J'(u; v) \leq 0$. So $J'(u; v) = 0$. □

In our case, $J'(u, v) = a(u, v) - F(v)$ for all $v \in V$, and therefore $J'(u, v) = 0$ for all $v \in V$ if and only if $a(u, v) = F(v)$ for all $v \in V$.

In our case, $J'(u,v) = a(u,v) - F(v)$ for all $v \in V$, and therefore $J'(u,v) = 0$ for all $v \in V$ if and only if $a(u,v) = F(v)$ for all $v \in V$.

More generally, the variational problem

$$\text{find } u \in V \text{ such that } J'(u;v) = 0 \text{ for all } v \in V$$

might be linear, nonlinear, coercive, or not! In general it describes *stationary points* of J, not just minimisers.

In our case, $J'(u,v) = a(u,v) - F(v)$ for all $v \in V$, and therefore $J'(u,v) = 0$ for all $v \in V$ if and only if $a(u,v) = F(v)$ for all $v \in V$.

More generally, the variational problem

$$\text{find } u \in V \text{ such that } J'(u;v) = 0 \text{ for all } v \in V$$

might be linear, nonlinear, coercive, or not! In general it describes *stationary points* of J, not just minimisers.

Only symmetric problems can arise from energy minimisation. If a variational problem is nonsymmetric it does not enjoy this structure.

Section 3

Galerkin approximation

We saw in Lecture 1 that a finite element approximation is a *Galerkin* approximation: given a closed subspace $V_h \subset V$, we approximate the solution of

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

We saw in Lecture 1 that a finite element approximation is a *Galerkin* approximation: given a closed subspace $V_h \subset V$, we approximate the solution of

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

with the Galerkin approximation over V_h:

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

Assume that a is symmetric, coercive, and bounded. Then these two variational problems can be restated as (for $J(v) := \frac{1}{2}a(v,v) - F(v)$):

$$u = \underset{v \in V}{\operatorname{argmin}}\ J(v),$$

Assume that a is symmetric, coercive, and bounded. Then these two variational problems can be restated as (for $J(v) := \frac{1}{2}a(v,v) - F(v)$):

$$u = \underset{v \in V}{\operatorname{argmin}}\ J(v),$$

and

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

Assume that a is symmetric, coercive, and bounded. Then these two variational problems can be restated as (for $J(v) := \frac{1}{2}a(v,v) - F(v)$):

$$u = \underset{v \in V}{\operatorname{argmin}}\ J(v),$$

and

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

So

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

Assume that a is symmetric, coercive, and bounded. Then these two variational problems can be restated as (for $J(v) := \frac{1}{2}a(v,v) - F(v)$):

$$u = \operatorname*{argmin}_{v \in V} J(v),$$

and

$$u_h = \operatorname*{argmin}_{v_h \in V_h} J(v_h).$$

So

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

The finite element method provides *the best approximation* in this energetic sense, when the problem has a nice quadratic energy functional.

Example 1. Let us set

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

and look for minimisers of

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

Example 1. Let us set

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

and look for minimisers of

$$J(u) = \frac{1}{2}\int_\Omega \nabla u \cdot \nabla u \, \mathrm{d}x + \frac{1}{2}\int_\Omega u^2 \, \mathrm{d}x - \int_\Omega fu \, \mathrm{d}x - \int_{\Gamma_N} gu \, \mathrm{d}s.$$

Calculating, we find

$$J'(u; v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \int_\Omega uv \, \mathrm{d}x - \int_\Omega fv \, \mathrm{d}x - \int_{\Gamma_N} gv \, \mathrm{d}s$$

Setting $J'(u; v) = 0$, we have a LVP $a(u, v) = F(v)$ for all $v \in V$ with

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \int_\Omega uv \, \mathrm{d}x, \quad F(v) = \int_\Omega fv \, \mathrm{d}x + \int_{\Gamma_N} gv \, \mathrm{d}s.$$

This is called the (good) *Helmholtz* problem:

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \; \mathrm{d}x + \int_\Omega uv \; \mathrm{d}x, \quad F(v) = \int_\Omega fv \; \mathrm{d}x + \int_{\Gamma_N} gv \; \mathrm{d}s.$$

This is called the (good) *Helmholtz* problem:

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \; \mathrm{d}x + \int_\Omega uv \; \mathrm{d}x, \quad F(v) = \int_\Omega fv \; \mathrm{d}x + \int_{\Gamma_N} gv \; \mathrm{d}s.$$

In strong form, we have

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

The bilinear form $a(u,v) = (u,v)_{H^1(\Omega)}$, so the form is continuous by the Cauchy–Schwarz inequality with continuity constant $C = 1$.

The bilinear form $a(u, v) = (u, v)_{H^1(\Omega)}$, so the form is continuous by the Cauchy–Schwarz inequality with continuity constant $C = 1$.

Similarly, as

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

the bilinear form is coercive with coercivity constant $\alpha = 1$.

Let us look at how to compute minimisers of this Helmholtz problem.

```
from firedrake import *

mesh = UnitSquareMesh(128, 128, quadrilateral=True)
(left, right) = (1, 2)
V = FunctionSpace(mesh, "CG", 1)
(x, y) = SpatialCoordinate(mesh)

f = Constant(1), g = Constant(0)
bc = DirichletBC(V, 0, (left, right))

u = Function(V)

J = (0.5 * inner(grad(u), grad(u))*dx + 0.5 * inner(u, u)
    - inner(f, u)*dx - inner(g, u)*ds)
G = derivative(J, u, TestFunction(V))

solve(G == 0, u, bc)
```

Example 2. By far the most important problem of this form is the equations of *linear elasticity*. The finite element method was invented by engineers looking to solve linear elasticity problems; its mathematical foundations and generalisations came much later.

Example 2. By far the most important problem of this form is the equations of *linear elasticity*. The finite element method was invented by engineers looking to solve linear elasticity problems; its mathematical foundations and generalisations came much later.

Let $\Omega \subset \mathbb{R}^n$ 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 $\phi : \bar{\Omega} \to \mathbb{R}^n$ via

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

Example 2. By far the most important problem of this form is the equations of *linear elasticity*. The finite element method was invented by engineers looking to solve linear elasticity problems; its mathematical foundations and generalisations came much later.

Let $\Omega \subset \mathbb{R}^n$ 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 $\phi : \bar{\Omega} \to \mathbb{R}^n$ via

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

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

$$\phi(x) = x + u(x),$$

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

For an isotropic homogeneous body, the displacement $u : \bar{\Omega} \to \mathbb{R}^n$ minimises the potential energy

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

Here f is the body loading (e.g. gravity), g is the surface traction, and

$$\varepsilon(u) = \frac{1}{2} \left(\nabla u + (\nabla u)^{\mathrm{T}} \right),$$

i.e. the symmetric part of the gradient (Jacobian matrix) $\nabla u : \Omega \to \mathbb{R}^{n \times n}$ of the displacement. The parameters $\mu, \lambda > 0$ are material-dependent.

For an isotropic homogeneous body, the displacement $u : \bar{\Omega} \to \mathbb{R}^n$ minimises the potential energy

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

Here f is the body loading (e.g. gravity), g is the surface traction, and

$$\varepsilon(u) = \frac{1}{2}\left(\nabla u + (\nabla u)^{\mathrm{T}}\right),$$

i.e. the symmetric part of the gradient (Jacobian matrix) $\nabla u : \Omega \to \mathbb{R}^{n\times n}$ of the displacement. The parameters $\mu, \lambda > 0$ are material-dependent.

This yields a linear variational problem with

$$a(u,v) = \int_\Omega 2\mu\varepsilon(u) : \varepsilon(v) + \lambda(\nabla \cdot u)(\nabla \cdot v) \, \mathrm{d}x, \; F(v) = \int_\Omega f \cdot v \, \mathrm{d}x + \int_{\Gamma_N} g \cdot v \, \mathrm{d}s.$$

In strong form, we have

$$\begin{aligned} -2\mu\nabla\cdot\varepsilon(u) - \lambda\nabla(\nabla\cdot u) &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma_D, \\ 2\mu\varepsilon(u)\cdot n + \lambda\left(\operatorname{tr}\varepsilon(u)\right)n &= g \text{ on } \Gamma_N. \end{aligned}$$

In strong form, we have

$$\begin{aligned} -2\mu\nabla\cdot\varepsilon(u) - \lambda\nabla(\nabla\cdot u) &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \Gamma_D, \\ 2\mu\varepsilon(u)\cdot n + \lambda\,(\operatorname{tr}\varepsilon(u))\,n &= g \text{ on } \Gamma_N. \end{aligned}$$

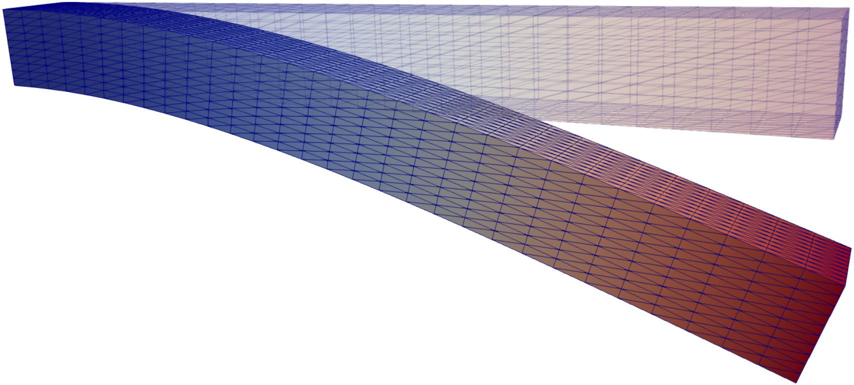
The bilinear form is continuous and coercive. In n dimensions, the continuity constant is

$$C = 2\mu + n\lambda.$$

Coercivity is guaranteed by *Korn's inequality*; if $\Gamma_D = \partial\Omega$, the coercivity constant is

$$\alpha = \mu.$$

```python
from firedrake import *

mesh = BoxMesh(30, 10, 10, 10, 1, 1)
V = VectorFunctionSpace(mesh, "CG", 1)

g = Constant((0, 0, -5e7))
bc = DirichletBC(V, 0, 1)

u = Function(V)
(mu, lam) = (27.4e9, 64.0e9)

J = (0.5 * 2*mu * inner(sym(grad(u)), sym(grad(u)))*dx
    + 0.5 * lam  * inner(div(u), div(u)) * dx
    - inner(g, u)*ds(2))
G = derivative(J, u, TestFunction(V))

solve(G == 0, u, bc)
```



C6.4 Finite Element Methods for PDEs
Lecture 7: Galerkin approximation

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

Given a linear variational problem

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

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

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

Given a linear variational problem

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

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

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

We first consider its approximation properties over *arbitrary* subspaces V_h.
In subsequent lectures we shall consider V_h constructed via finite elements.

Corollary

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

Corollary

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

Proof.

As $V_h \subset V$, $a : V_h \times V_h \to \mathbb{R}$ is bounded and coercive on V_h, with the same continuity and coercivity constants.

Corollary

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

Proof.

As $V_h \subset V$, $a : V_h \times V_h \to \mathbb{R}$ is bounded and coercive on V_h, with the same continuity and coercivity constants. $F : V_h \to \mathbb{R}$ is linear and bounded.

Corollary

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

Proof.

As $V_h \subset V$, $a : V_h \times V_h \to \mathbb{R}$ is bounded and coercive on V_h, with the same continuity and coercivity constants. $F : V_h \to \mathbb{R}$ is linear and bounded. Because V_h is, by assumption, closed and therefore a Hilbert subspace of the Hilbert space V, by the Lax–Milgram Theorem the variational problem defining the Galerkin approximation is well-posed. $\square$

Corollary

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

Proof.

As $V_h \subset V$, $a : V_h \times V_h \to \mathbb{R}$ is bounded and coercive on V_h, with the same continuity and coercivity constants. $F : V_h \to \mathbb{R}$ is linear and bounded. Because V_h is, by assumption, closed and therefore a Hilbert subspace of the Hilbert space V, by the Lax–Milgram Theorem the variational problem defining the Galerkin approximation is well-posed. $\square$

For coercive problems, well-posedness is inherited. *This is not true for noncoercive problems.* This makes discretising noncoercive problems much harder.

Once we choose a basis $\{\phi_i\}_{i=1}^N$ of V_h, the linear system is

$$Ax = b,$$

where $x = (x_1, \ldots, x_N)^{\mathrm{T}} \in \mathbb{R}^N$, $b = (b_1, \ldots, b_N)^{\mathrm{T}} \in \mathbb{R}^N$,

$$u_h = \sum_{i=1}^{N} x_i \phi_i, \quad b_j = F(\phi_j), \quad A_{ji} = a(\phi_i, \phi_j), \qquad i, j = 1, \ldots, N.$$

Once we choose a basis $\{\phi_i\}_{i=1}^N$ of V_h, the linear system is

$$Ax = b,$$

where $x = (x_1, \ldots, x_N)^{\mathrm{T}} \in \mathbb{R}^N$, $b = (b_1, \ldots, b_N)^{\mathrm{T}} \in \mathbb{R}^N$,

$$u_h = \sum_{i=1}^N x_i \phi_i, \quad b_j = F(\phi_j), \quad A_{ji} = a(\phi_i, \phi_j), \qquad i, j = 1, \ldots, N.$$

The linear system we must solve for our Galerkin approximation inherits useful properties of the underlying problem.

Once we choose a basis $\{\phi_i\}_{i=1}^N$ of V_h, the linear system is

$$Ax = b,$$

where $x = (x_1, \dots, x_N)^{\mathrm{T}} \in \mathbb{R}^N$, $b = (b_1, \dots, b_N)^{\mathrm{T}} \in \mathbb{R}^N$,

$$u_h = \sum_{i=1}^{N} x_i \phi_i, \quad b_j = F(\phi_j), \quad A_{ji} = a(\phi_i, \phi_j), \qquad i, j = 1, \dots, N.$$

The linear system we must solve for our Galerkin approximation inherits useful properties of the underlying problem.

If a is symmetric, so is A:

$$A_{ji} = a(\phi_i, \phi_j) = a(\phi_j, \phi_i) = A_{ij}, \quad i, j = 1, \dots, N.$$

Once we choose a basis $\{\phi_i\}_{i=1}^N$ of V_h, the linear system is

$$Ax = b,$$

where $x = (x_1, \ldots, x_N)^{\mathrm{T}} \in \mathbb{R}^N$, $b = (b_1, \ldots, b_N)^{\mathrm{T}} \in \mathbb{R}^N$,

$$u_h = \sum_{i=1}^N x_i \phi_i, \quad b_j = F(\phi_j), \quad A_{ji} = a(\phi_i, \phi_j), \qquad i, j = 1, \ldots, N.$$

The linear system we must solve for our Galerkin approximation inherits useful properties of the underlying problem.

If a is symmetric, so is A:

$$A_{ji} = a(\phi_i, \phi_j) = a(\phi_j, \phi_i) = A_{ij}, \quad i, j = 1, \ldots, N.$$

If a is coercive, then A is positive definite: for all $c \in \mathbb{R}^N \setminus \{0\}$:

$$c^\top A c = a\left(\sum_{i=1}^N c_i \phi_i, \sum_{i=1}^N c_i \phi_i\right) \geq \alpha \left\| \sum_{i=1}^N c_i \phi_i \right\|_V^2 > 0.$$

We know that the solution u satisfies

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

and thus in particular

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

We know that the solution u satisfies

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

and thus in particular

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

The Galerkin approximation $u_h \in V_h$ satisfies

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

We know that the solution u satisfies

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

and thus in particular

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

The Galerkin approximation $u_h \in V_h$ satisfies

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

Subtracting, we find

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

We know that the solution u satisfies

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

and thus in particular

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

The Galerkin approximation $u_h \in V_h$ satisfies

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

Subtracting, we find

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

This is called *Galerkin orthogonality*.

Let us assume that a is coercive and bounded, but not symmetric.

Lemma (Céa's Lemma)

The Galerkin approximation $u_h \in V_h$ to $u \in V$ is quasi-optimal, in that it satisfies

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

where α and C are the coercivity and continuity constant, respectively.

Let us assume that a is coercive and bounded, but not symmetric.

Lemma (Céa's Lemma)

The Galerkin approximation $u_h \in V_h$ to $u \in V$ is quasi-optimal, in that it satisfies

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

where α and C are the coercivity and continuity constant, respectively.

Proof.

For any $v_h \in V_h$,

$$\alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h)$$

$\square$

Let us assume that a is coercive and bounded, but not symmetric.

Lemma (Céa's Lemma)

The Galerkin approximation $u_h \in V_h$ to $u \in V$ is quasi-optimal, in that it satisfies

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

where α and C are the coercivity and continuity constant, respectively.

Proof.

For any $v_h \in V_h$,

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

□

Let us assume that a is coercive and bounded, but not symmetric.

Lemma (Céa's Lemma)

The Galerkin approximation $u_h \in V_h$ to $u \in V$ is quasi-optimal, in that it satisfies

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

where α and C are the coercivity and continuity constant, respectively.

Proof.

For any $v_h \in V_h$,

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

□

Let us assume that a is coercive and bounded, but not symmetric.

Lemma (Céa's Lemma)

The Galerkin approximation $u_h \in V_h$ to $u \in V$ is quasi-optimal, in that it satisfies

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

where α and C are the coercivity and continuity constant, respectively.

Proof.

For any $v_h \in V_h$,

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

$\square$

Let us assume that a is coercive and bounded, but not symmetric.

Lemma (Céa's Lemma)

The Galerkin approximation $u_h \in V_h$ to $u \in V$ is quasi-optimal, in that it satisfies

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

where α and C are the coercivity and continuity constant, respectively.

Proof.

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 result. $\square$

Remark

This quasi-optimality result relates (the error in the PDE approximation) with (the approximating power of the space V_h). This decouples the error analysis from the specific PDE and turns the focus to constructing V_h with good approximation properties.

Remark

This quasi-optimality result relates (the error in the PDE approximation) with (the approximating power of the space V_h). This decouples the error analysis from the specific PDE and turns the focus to constructing V_h with good approximation properties.

This leads to the question: given $u \in V$, what is

$$\min_{v_h \in V_h} \|u - v_h\|_V ?$$

In the finite element context, the answer will depend on the smoothness of u, the mesh size h, and the polynomial degree p.

Remark

This quasi-optimality result relates (the error in the PDE approximation) with (the approximating power of the space V_h). This decouples the error analysis from the specific PDE and turns the focus to constructing V_h with good approximation properties.

This leads to the question: given $u \in V$, what is

$$\min_{v_h \in V_h} \|u - v_h\|_V?$$

In the finite element context, the answer will depend on the smoothness of u, the mesh size h, and the polynomial degree p.

Remark

The ratio C/α is crucial. If $C/\alpha = 5$, things are fine. But if $C/\alpha = 1000$, our discretisation *will not* be very useful.

Now let us also assume that a is symmetric.

Now let us also assume that a is symmetric.

Recall that a defines a norm $\|v\|_a := \sqrt{a(v,v)}$ on V, with

$$\alpha\|v\|_V^2 \le \|v\|_a^2 \le C\|v\|_V^2,$$

where the continuity and coercivity constants are measured in the V norm.

Now let us also assume that a is symmetric.

Recall that a defines a norm $\|v\|_a := \sqrt{a(v,v)}$ on V, with

$$\alpha\|v\|_V^2 \leq \|v\|_a^2 \leq C\|v\|_V^2,$$

where the continuity and coercivity constants are measured in the V norm.

When we measure the continuity and coercivity constants *in the energy norm*, we get that $C = 1$ (by Cauchy–Schwarz) and $\alpha = 1$ (by definition):

$$a(u,v) \leq 1 \cdot \|u\|_a \, \|v\|_a, \qquad a(v,v) = \|v\|_a^2 \geq 1 \cdot \|v\|_a^2.$$

Apply Céa's Lemma in the energy norm $\|\cdot\|_a$, with $C = \alpha = 1$:

$$\begin{aligned} \|u - u_h\|_a &\leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_a \\ &= \min_{v_h \in V_h} \|u - v_h\|_a. \end{aligned}$$

Apply Céa's Lemma in the energy norm $\|\cdot\|_a$, with $C = \alpha = 1$:

$$\begin{aligned}\|u - u_h\|_a &\leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_a \\ &= \min_{v_h \in V_h} \|u - v_h\|_a.\end{aligned}$$

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

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

Apply Céa's Lemma in the energy norm $\|\cdot\|_a$, with $C = \alpha = 1$:

$$\begin{aligned}\|u - u_h\|_a &\le \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_a \\ &= \min_{v_h \in V_h} \|u - v_h\|_a.\end{aligned}$$

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

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

The Galerkin approximation u_h is the *orthogonal projection* of u onto V_h in the a-inner product $a(\cdot, \cdot)$.

What if we want to measure our error for a symmetric problem in the V-norm?

What if we want to measure our error for a symmetric problem in the V-norm?

Using the norm-equivalence

$$\alpha\|v\|_V^2 \leq \|v\|_a^2 \leq C\|v\|_V^2,$$

we have

$$\|u - u_h\|_V \leq \frac{1}{\sqrt{\alpha}}\|u - u_h\|_a$$

What if we want to measure our error for a symmetric problem in the V-norm?

Using the norm-equivalence

$$\alpha\|v\|_V^2 \leq \|v\|_a^2 \leq C\|v\|_V^2,$$

we have

$$\begin{aligned} \|u - u_h\|_V &\leq \frac{1}{\sqrt{\alpha}}\|u - u_h\|_a \\ &= \frac{1}{\sqrt{\alpha}} \min_{v_h \in V_h} \|u - v_h\|_a \end{aligned}$$

What if we want to measure our error for a symmetric problem in the V-norm?

Using the norm-equivalence

$$\alpha\|v\|_V^2 \le \|v\|_a^2 \le C\|v\|_V^2,$$

we have

$$\begin{aligned}\|u-u_h\|_V &\le \frac{1}{\sqrt{\alpha}}\|u-u_h\|_a \\ &= \frac{1}{\sqrt{\alpha}}\min_{v_h\in V_h}\|u-v_h\|_a \\ &\le \sqrt{\frac{C}{\alpha}}\min_{v_h\in V_h}\|u-v_h\|_V.\end{aligned}$$

What if we want to measure our error for a symmetric problem in the V-norm?

Using the norm-equivalence

$$\alpha\|v\|_V^2 \leq \|v\|_a^2 \leq C\|v\|_V^2,$$

we have

$$\begin{aligned} \|u-u_h\|_V &\leq \frac{1}{\sqrt{\alpha}}\|u-u_h\|_a \\ &= \frac{1}{\sqrt{\alpha}} \min_{v_h \in V_h} \|u-v_h\|_a \\ &\leq \sqrt{\frac{C}{\alpha}} \min_{v_h \in V_h} \|u-v_h\|_V. \end{aligned}$$

Thus, in the case of a symmetric bilinear form $a(\cdot,\cdot)$ we have improved the constant of quasi-optimality by a square root!

Section 5

Linear elasticity: the nearly incompressible case

Let us consider linear elasticity again for a problem on $\Omega \subset \mathbb{R}^n$. We seek

$$u = \underset{v \in [H_0^1(\Omega)]^n}{\text{argmin}} \frac{1}{2} \int_\Omega 2\mu\varepsilon(u) : \varepsilon(u) + \lambda(\nabla \cdot u)^2 \, \mathrm{d}x - \int_\Omega f \cdot u \, \mathrm{d}x.$$

Let us consider linear elasticity again for a problem on $\Omega \subset \mathbb{R}^n$. We seek

$$u = \operatorname*{argmin}_{v \in [H_0^1(\Omega)]^n} \frac{1}{2} \int_\Omega 2\mu\varepsilon(u) : \varepsilon(u) + \lambda(\nabla \cdot u)^2 \, \mathrm{d}x - \int_\Omega f \cdot u \, \mathrm{d}x.$$

When engineers implemented the finite element method for this problem, they observed something puzzling: it worked well for steel and concrete, but did not work for rubber. Why not?

Let us consider linear elasticity again for a problem on $\Omega \subset \mathbb{R}^n$. We seek

$$u = \operatorname*{argmin}_{v \in [H_0^1(\Omega)]^n} \frac{1}{2} \int_\Omega 2\mu\varepsilon(u) : \varepsilon(u) + \lambda(\nabla \cdot u)^2 \, \mathrm{d}x - \int_\Omega f \cdot u \, \mathrm{d}x.$$

When engineers implemented the finite element method for this problem, they observed something puzzling: it worked well for steel and concrete, but did not work for rubber. Why not?

We claimed in Lecture 6 that its coercivity and continuity constants are

$$C = 2\mu + n\lambda, \quad \alpha = \mu, \qquad \text{so} \quad \frac{C}{\alpha} = 2 + n\frac{\lambda}{\mu}.$$

Let us investigate the practical consequences of this.

Consider the following different materials for $n = 3$:

Material	μ	λ	C	α	$\sqrt{C/\alpha}$
steel	75	112	486	75	2.55

Table: Lamé parameters for different materials. All units for μ, λ, C, α are multiplied by 10^9.

Consider the following different materials for $n = 3$:

Material	μ	λ	C	α	$\sqrt{C/\alpha}$
steel	75	112	486	75	2.55
concrete	18	27	117	18	2.55

Table: Lamé parameters for different materials. All units for μ, λ, C, α are multiplied by 10^9.

Consider the following different materials for $n = 3$:

Material	μ	λ	C	α	$\sqrt{C/\alpha}$
steel	75	112	486	75	2.55
concrete	18	27	117	18	2.55
rubber A	0.018	0.9	2.75	0.018	12.2

Table: Lamé parameters for different materials. All units for μ, λ, C, α are multiplied by 10^9.

Consider the following different materials for $n = 3$:

Material	μ	λ	C	α	$\sqrt{C/\alpha}$
steel	75	112	486	75	2.55
concrete	18	27	117	18	2.55
rubber A	0.018	0.9	2.75	0.018	12.2
rubber B	0.018	9.0	27.5	0.018	38.7

Table: Lamé parameters for different materials. All units for μ, λ, C, α are multiplied by 10^9.

Consider the following different materials for $n = 3$:

Material	μ	λ	C	α	$\sqrt{C/\alpha}$
steel	75	112	486	75	2.55
concrete	18	27	117	18	2.55
rubber A	0.018	0.9	2.75	0.018	12.2
rubber B	0.018	9.0	27.5	0.018	38.7
rubber C	0.018	90	275	0.018	122.4

Table: Lamé parameters for different materials. All units for μ, λ, C, α are multiplied by 10^9.

Consider the following different materials for $n = 3$:

Material	μ	λ	C	α	$\sqrt{C/\alpha}$
steel	75	112	486	75	2.55
concrete	18	27	117	18	2.55
rubber A	0.018	0.9	2.75	0.018	12.2
rubber B	0.018	9.0	27.5	0.018	38.7
rubber C	0.018	90	275	0.018	122.4

Table: Lamé parameters for different materials. All units for μ, λ, C, α are multiplied by 10^9.

As $\lambda \to \infty$, for $\mu > 0$ fixed, the Galerkin approximation breaks down. This is because $C = C(\mu, \lambda) \to \infty$, while $\alpha = \alpha(\mu)$. This is called *locking*.

Consider the following different materials for $n = 3$:

Material	μ	λ	C	α	$\sqrt{C/\alpha}$
steel	75	112	486	75	2.55
concrete	18	27	117	18	2.55
rubber A	0.018	0.9	2.75	0.018	12.2
rubber B	0.018	9.0	27.5	0.018	38.7
rubber C	0.018	90	275	0.018	122.4

Table: Lamé parameters for different materials. All units for μ, λ, C, α are multiplied by 10^9.

As $\lambda \to \infty$, for $\mu > 0$ fixed, the Galerkin approximation breaks down. This is because $C = C(\mu, \lambda) \to \infty$, while $\alpha = \alpha(\mu)$. This is called *locking*.

The parameter λ penalises $\|\nabla \cdot u\|^2_{L^2(\Omega)}$; as $\lambda \to \infty$, the displacement is not allowed to change the volume. The different rubber samples are becoming *nearly incompressible*.

Note that this problem of locking is not specific to any particular discretisation; it is that *the formulation of the problem is becoming ill-conditioned* as $C/\alpha \to \infty$.

Note that this problem of locking is not specific to any particular discretisation; it is that *the formulation of the problem is becoming ill-conditioned* as $C/\alpha \to \infty$.

What can we do?

Note that this problem of locking is not specific to any particular discretisation; it is that *the formulation of the problem is becoming ill-conditioned* as $C/\alpha \to \infty$.

What can we do? Use a different formulation!

We introduce an auxiliary variable

$$p := \lambda \nabla \cdot u$$

which in weak form becomes

$$\int_\Omega q \nabla \cdot u \; \mathrm{d}x = \frac{1}{\lambda} \int_\Omega pq \; \mathrm{d}x.$$

We introduce an auxiliary variable

$$p := \lambda \nabla \cdot u$$

which in weak form becomes

$$\int_\Omega q \nabla \cdot u \, \mathrm{d}x = \frac{1}{\lambda} \int_\Omega pq \, \mathrm{d}x.$$

We then consider: find $(u, p) \in H_0^1(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ such that

$$\int_\Omega 2\mu \varepsilon(u) : \varepsilon(v) \, \mathrm{d}x + \int_\Omega p \nabla \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q \nabla \cdot u \, \mathrm{d}x - \frac{1}{\lambda} \int_\Omega pq \, \mathrm{d}x = 0,$$

for all $(v, q) \in H_0^1(\Omega; \mathbb{R}^2) \times L^2(\Omega)$.

We introduce an auxiliary variable

$$p := \lambda \nabla \cdot u$$

which in weak form becomes

$$\int_\Omega q \nabla \cdot u \, \mathrm{d}x = \frac{1}{\lambda} \int_\Omega pq \, \mathrm{d}x.$$

We then consider: find $(u, p) \in H^1_0(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ such that

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v) \, \mathrm{d}x + \int_\Omega p \nabla \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q \nabla \cdot u \, \mathrm{d}x - \frac{1}{\lambda} \int_\Omega pq \, \mathrm{d}x = 0,$$

for all $(v, q) \in H^1_0(\Omega; \mathbb{R}^2) \times L^2(\Omega)$.

This remains uniformly well-posed as $\lambda \to \infty$, even for $\lambda = \infty$!

Does this, so called mixed formulation, have an energetic structure?

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v) \, \mathrm{d}x + \int_\Omega p\nabla \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x$$
$$\int_\Omega q\nabla \cdot u \, \mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq \, \mathrm{d}x = 0$$

Does this, so called mixed formulation, have an energetic structure?

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\ \mathrm{d}x + \int_\Omega p\nabla \cdot v\ \mathrm{d}x = \int_\Omega f \cdot v\ \mathrm{d}x$$
$$\int_\Omega q\nabla \cdot u\ \mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\ \mathrm{d}x = 0$$

are the Euler–Lagrange stationarity conditions for the Lagrangian

$$L(u,p) = \frac{1}{2}\int_\Omega 2\mu\varepsilon(u) : \varepsilon(u)\,\mathrm{d}x + \int_\Omega p\nabla \cdot u\,\mathrm{d}x - \frac{1}{2\lambda}\int p^2\,\mathrm{d}x - \int_\Omega f \cdot u\,\mathrm{d}x.$$

This becomes the familiar linear elastic energy under $p \mapsto \lambda\nabla \cdot u$.

Does this, so called mixed formulation, have an energetic structure?

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\ \mathrm{d}x + \int_\Omega p\nabla\cdot v\ \mathrm{d}x = \int_\Omega f\cdot v\ \mathrm{d}x$$
$$\int_\Omega q\nabla\cdot u\ \mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\ \mathrm{d}x = 0$$

are the Euler–Lagrange stationarity conditions for the Lagrangian

$$L(u,p) = \frac{1}{2}\int_\Omega 2\mu\varepsilon(u) : \varepsilon(u)\,\mathrm{d}x + \int_\Omega p\nabla\cdot u\,\mathrm{d}x - \frac{1}{2\lambda}\int p^2\,\mathrm{d}x - \int_\Omega f\cdot u\,\mathrm{d}x.$$

This becomes the familiar linear elastic energy under $p \mapsto \lambda\nabla\cdot u$.

We will see later that the solution (u,p) is a saddle point of L.

Note that the bilinear form

$$B(u,p;v,q) = \int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x + \int_\Omega p\nabla\cdot v\,\mathrm{d}x + \int_\Omega q\nabla\cdot u\,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x$$

Note that the bilinear form

$$B(u,p;v,q) = \int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x + \int_\Omega p\nabla\cdot v\,\mathrm{d}x + \int_\Omega q\nabla\cdot u\,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x$$

is not coercive:

$$B(0,p;0,p) = -\frac{1}{\lambda}\|p\|^2_{L^2(\Omega)} < 0 \text{ for } p \neq 0.$$

Note that the bilinear form

$$B(u,p;v,q) = \int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x + \int_\Omega p\nabla\cdot v\,\mathrm{d}x + \int_\Omega q\nabla\cdot u\,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x$$

is not coercive:

$$B(0,p;0,p) = -\frac{1}{\lambda}\|p\|^2_{L^2(\Omega)} < 0 \text{ for } p \neq 0.$$

We will see in subsequent lectures how to analyse such systems. We have exchanged a simple-to-discretise but unstable formulation for a stable but one that is harder to discretise.

```
from firedrake import *
mesh = BoxMesh(30, 10, 10, 10, 1, 1)
V = VectorFunctionSpace(mesh, "CG", 2)
Q = FunctionSpace(mesh, "CG", 1)
Z = V*Q

g = Constant((0, 0, -5e7))
bc = DirichletBC(Z.sub(0), 0, 1)

z = Function(Z)
(u, p) = split(z)
(mu, lam) = (27.4e9, 64.0e9)

L = (mu * inner(sym(grad(u)), sym(grad(u)))*dx
    + p * div(u) * dx - 1/(2*lam) * p**2 * dx
    - inner(g, u)*ds(2))
G = derivative(L, z, TestFunction(Z))
solve(G == 0, z, bc)
```

C6.4 Finite Element Methods for PDEs
Lecture 8: Constructing function spaces with finite elements

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

In the last lecture, we saw Céa's Lemma: for a coercive bounded a, the error in Galerkin approximation is bounded by

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

In the last lecture, we saw Céa's Lemma: for a coercive bounded a, the error in Galerkin approximation is bounded by

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

How do we construct discrete spaces $V_h \subset V$ with good approximating properties on general domains?

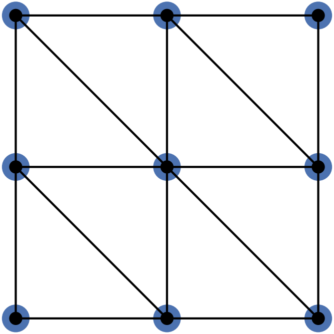
In the last lecture, we saw Céa's Lemma: for a coercive bounded a, the error in Galerkin approximation is bounded by

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

How do we construct discrete spaces $V_h \subset V$ with good approximating properties on general domains?

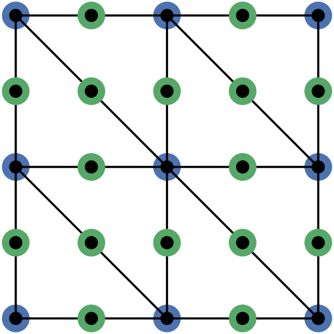
The finite element method!

Key idea: use piecewise polynomials on a mesh of Ω.



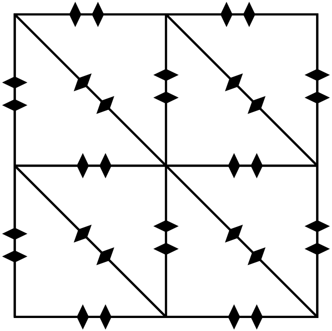
Data stored to represent a piecewise linear function, $V_h \subset H^1(\Omega)$.

Key idea: use piecewise polynomials on a mesh of Ω.



Data stored to represent a piecewise quadratic function, $V_h \subset H^1(\Omega)$.

Key idea: use piecewise polynomials on a mesh of Ω.



Data stored to represent a piecewise linear vector function, $V_h \subset H(\mathrm{div}, \Omega)$.

Let us consider what happens on a single cell first, then stitch them together to enforce the continuity properties we need to conform V_h to the Sobolev space V we want.

Let us consider what happens on a single cell first, then stitch them together to enforce the continuity properties we need to conform V_h to the Sobolev space V we want.

Definition (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;

Let us consider what happens on a single cell first, then stitch them together to enforce the continuity properties we need to conform V_h to the Sobolev space V we want.

Definition (Finite element)

A finite element is a triple $(K, \mathcal{V}, \mathcal{L})$ where

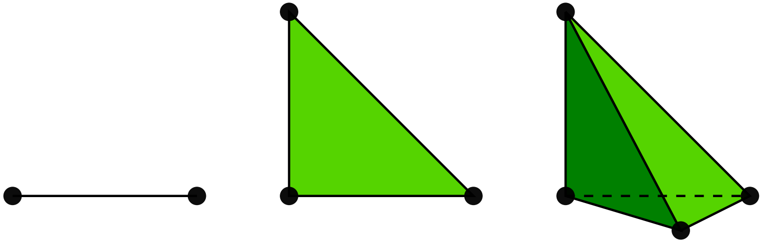
- The cell K is a bounded, closed subset of $\mathbb{R}^n$ with nonempty connected interior and piecewise smooth boundary;
- The space $\mathcal{V} = \mathcal{V}(K)$ is a finite-dimensional function space on K of dimension d;

Let us consider what happens on a single cell first, then stitch them together to enforce the continuity properties we need to conform V_h to the Sobolev space V we want.

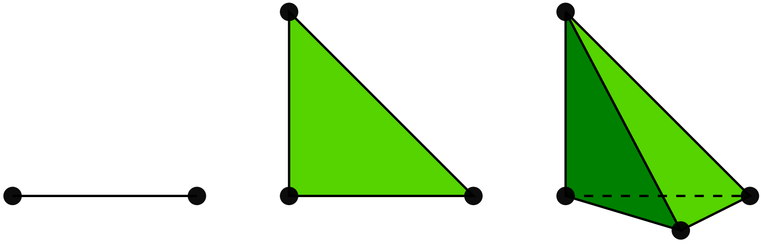
Definition (Finite element)

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 $\mathcal{V}$.



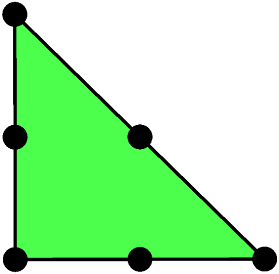
The linear Lagrange finite element CG_1 in one, two and three dimensions. The black circles denote pointwise evaluation.



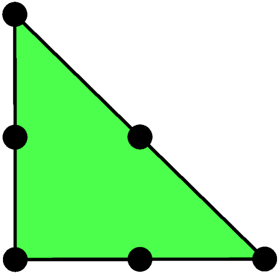
The linear Lagrange finite element CG_1 in one, two and three dimensions. The black circles denote pointwise evaluation.

Example (CG_1 in 2D)

$K = \triangle$, $\mathcal{V} = \mathrm{span}(1, x, y)$, $\mathcal{L} = \{\ell_1, \ell_2, \ell_3\}$, $\ell_i : v \mapsto v(x_i)$.



The quadratic Lagrange finite element CG_2 in two dimensions.



The quadratic Lagrange finite element CG_2 in two dimensions.

Example (CG_2 in 2D)

$K = \triangle$, $\mathcal{V} = \mathrm{span}(1, x, y, x^2, y^2, xy)$, $\mathcal{L} = \{\ell_1, \ldots, \ell_6\}$, each ℓ_i evaluates the function at a vertex or edge midpoint.

Definition (Polynomial spaces)

Denote the space of polynomials of *total* degree q on $K \subset \mathbb{R}^n$:

$$\mathcal{P}_q(K) := \operatorname{span}\left\{ x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}|_K : \sum_{i=1}^{n} \alpha_i \le q, \alpha_i \ge 0 \text{ for all } i = 1, \dots, n \right\}$$

and of *maximal* degree q on K:

$$\mathcal{Q}_q(K) := \operatorname{span}\left\{ x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}|_K : \alpha_i \le q, \alpha_i \ge 0 \text{ for all } i = 1, \dots, n \right\}.$$

Definition (Polynomial spaces)

Denote the space of polynomials of *total* degree q on $K \subset \mathbb{R}^n$:

$$\mathcal{P}_q(K) := \operatorname{span}\left\{ x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}|_K : \sum_{i=1}^{n} \alpha_i \le q, \alpha_i \ge 0 \text{ for all } i = 1, \dots, n \right\}$$

and of *maximal* degree q on K:

$$\mathcal{Q}_q(K) := \operatorname{span}\left\{ x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}|_K : \alpha_i \le q, \alpha_i \ge 0 \text{ for all } i = 1, \dots, n \right\}.$$

Example

For $\triangle, \square \subset \mathbb{R}^2$,

$$\mathcal{P}_2(\triangle) = \operatorname{span}\{1, x, x^2, y, y^2, xy\},$$
$$\mathcal{Q}_2(\square) = \operatorname{span}\{1, x, x^2, y, y^2, xy, x^2y, xy^2, x^2y^2\}.$$

Lemma (Verifying finite elements)

Let $\mathcal{V}$ be a d-dimensional vector space and let $\mathcal{L} = \{\ell_1, \dots, \ell_d\}$ 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 $i = 1, \dots, d$, then $v = 0$.

Lemma (Verifying finite elements)

Let $\mathcal{V}$ be a d-dimensional vector space and let $\mathcal{L} = \{\ell_1, \dots, \ell_d\}$ 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 $i = 1, \dots, d$, then $v = 0$.

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 $\mathcal{V}$ that satisfies this is the zero element.

Lemma (Verifying finite elements)

Let $\mathcal{V}$ be a d-dimensional vector space and let $\mathcal{L} = \{\ell_1, \dots, \ell_d\}$ 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 $i = 1, \dots, d$, then $v = 0$.

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 $\mathcal{V}$ that satisfies this is the zero element.

Definition

We say that $\mathcal{L} = \{\ell_1, \dots, \ell_d\}$ *determines* $\mathcal{V}$ if, given $v \in \mathcal{V}$,

$$\ell_i(v) = 0 \text{ for all } i = 1, \dots, d \quad \Longrightarrow \quad v = 0.$$

We also say that $\mathcal{L}$ is *unisolvent*.

Example

For $\mathrm{CG}_1(\triangle)$, 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 (nodal basis)

The basis $\{\phi_1, \ldots, \phi_d\}$ of $\mathcal{V}$ dual to $\mathcal{L}$, i.e. with the property that

$$\ell_i(\phi_j) = \delta_{ij}, \qquad i, j = 1, \ldots, d,$$

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

Example (CG_1 in 1D)

Let $K = [0, 1]$, $\mathcal{V} = \mathcal{P}_1(K)$, and $\mathcal{L}$ be pointwise evaluation at the endpoints. Then the nodal basis is given by

$$\phi_1(x) = 1 - x, \quad \phi_2(x) = x.$$

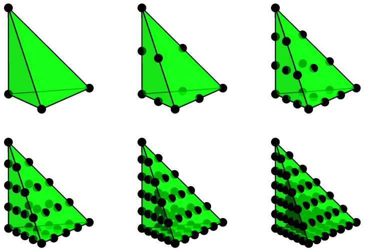
Indeed, as $\ell_1(v) = v(0)$ and $\ell_2(v) = v(1)$ clearly $\ell_i(\phi_j) = \delta_{ij}$, $i, j = 1, 2$.

Example (CG_1 in 2D)

Let K be the triangle with vertices at $(0,0)$, $(1,0)$, $(0,1)$. Let $\mathcal{V} = \mathcal{P}_1(K)$, and $\mathcal{L}$ be pointwise evaluation at the vertices. Then the nodal basis is given, for $x = (x_1, x_2)$, by

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

Indeed, as $\ell_1(v) = v(0,0)$, $\ell_2(v) = v(1,0)$, $\ell_3(v) = v(0,1)$, clearly $\ell_i(\phi_j) = \delta_{ij}$ for $i, j = 1, 2, 3$.



The Lagrange CG_q elements on tetrahedra for $q = 1, \ldots, 6$.

Definition (Lagrange element on a simplex)

The Lagrange element CG_q of dimension n and degree $q \geq 1$ is

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

where x_i, $i = 1, \ldots, f_n(q)$ is an enumeration of points in the element.

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

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

Definition (Interpolant on an element)

Let $(K, \mathcal{V}, \mathcal{L})$ be a finite element. For a suitable function space H, define the interpolant $\mathcal{I}_K : H \to \mathcal{V}$ via

$$\begin{aligned} \mathcal{I}_K &: u \mapsto \mathcal{I}_K u \\ \ell_i(\mathcal{I}_K u) &= \ell_i(u) \quad \text{for all } \ell_i \in \mathcal{L}, \quad i = 1, \dots, d. \end{aligned}$$

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

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

Definition (Interpolant on an element)

Let $(K, \mathcal{V}, \mathcal{L})$ be a finite element. For a suitable function space H, define the interpolant $\mathcal{I}_K : H \to \mathcal{V}$ via

$$\mathcal{I}_K : u \mapsto \mathcal{I}_K u$$
$$\ell_i(\mathcal{I}_K u) = \ell_i(u) \quad \text{for all } \ell_i \in \mathcal{L}, \quad i = 1, \ldots, d.$$

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

In the nodal basis, the interpolation operator is particularly simple:

$$\mathcal{I}_K u(x) = \sum_{i=1}^{d} \ell_i(u)\phi_i(x), \quad x \in K.$$

Section 2

Meshes and the local-to-global mapping

To define a global function space and basis

$$V_h = \mathrm{span}\{\phi_1, \dots, \phi_N\} \subset V,$$

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.

To define a global function space and basis

$$V_h = \mathrm{span}\{\phi_1, \ldots, \phi_N\} \subset V,$$

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 that Ω is polytopic, so that it can be decomposed into cells exactly. (Otherwise we have to worry about geometric approximation errors also.)

Definition (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. $\cup_i K_i = \overline{\Omega}$.

Definition (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

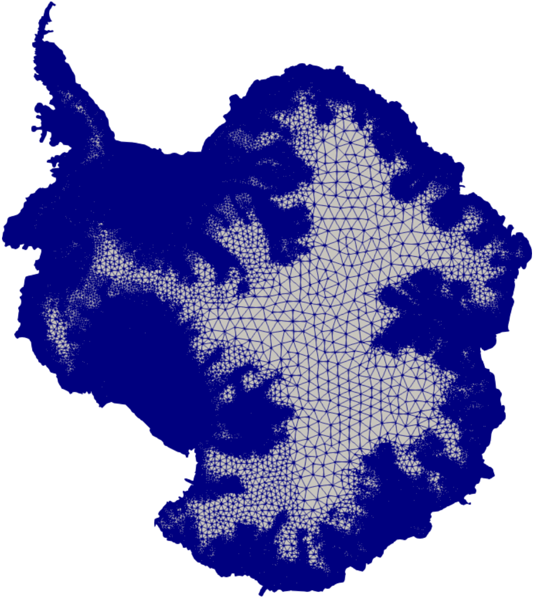
1. $\cup_i K_i = \overline{\Omega}$.
2. If $K_i \cap K_j$ for $i \neq j$ is exactly one point, it is a common vertex of K_i and K_j.

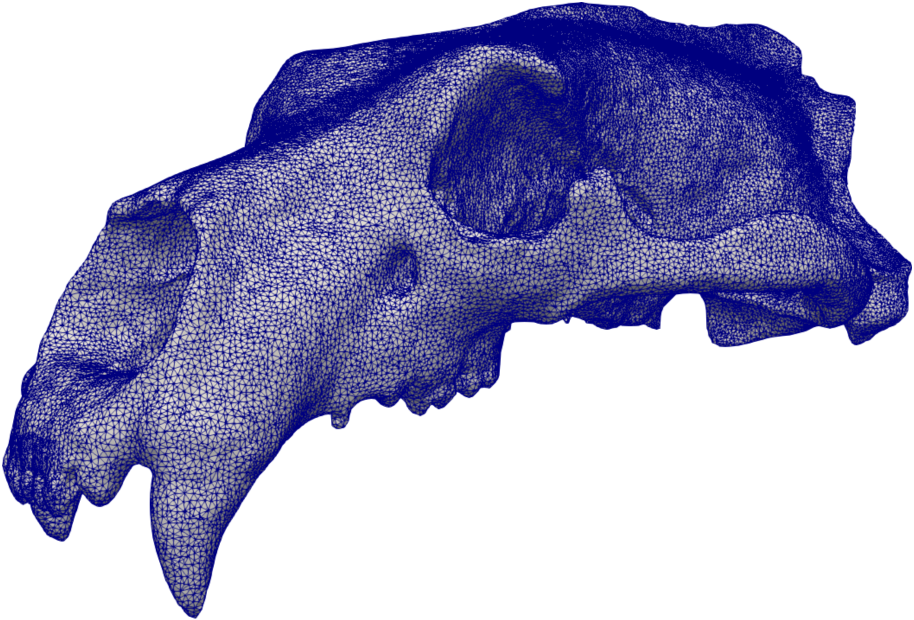
Definition (mesh)

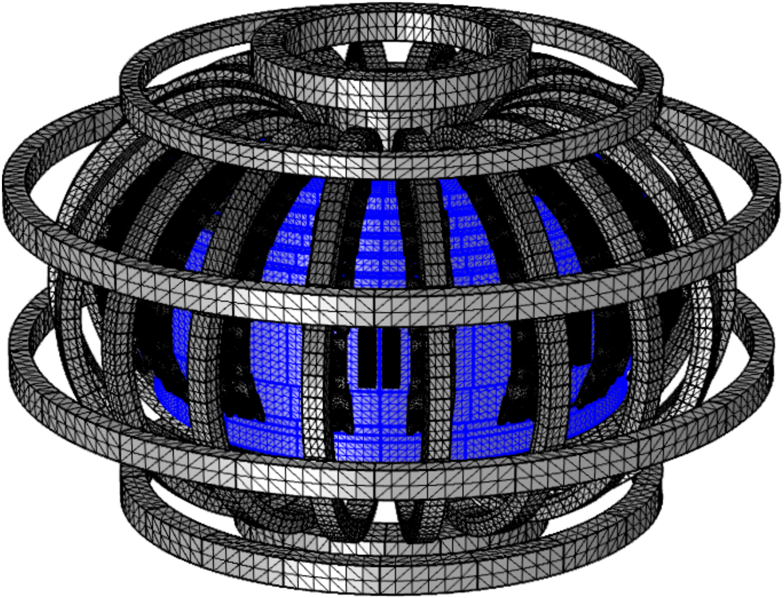
A *mesh* $\mathcal{M}$ is a geometric decomposition of a domain Ω into a finite set of *cells* $\mathcal{M} = \{K_i\}$ such that

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 empty or 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 of computational geometry in its own right.







We equip each cell $K \in \mathcal{M}$ with a finite element, so we have a set of finite elements $\{(K, \mathcal{V}_K, \mathcal{L}_K) : K \in \mathcal{M}\}$. To avoid technicalities we shall assume that we equip each cell with the same type of element.

We equip each cell $K \in \mathcal{M}$ with a finite element, so we have a set of finite elements $\{(K, \mathcal{V}_K, \mathcal{L}_K) : K \in \mathcal{M}\}$. To avoid technicalities we shall assume that we equip each cell with the same type of element.

We can thus give our first specification of a finite element space. Suppose we solve a variational problem over V. Then we take $V_h \subset V$ defined by

$$V_h := \{v \in V : v|_K \in \mathcal{V}_K \quad \forall\, K \in \mathcal{M}\}.$$

We equip each cell $K \in \mathcal{M}$ with a finite element, so we have a set of finite elements $\{(K, \mathcal{V}_K, \mathcal{L}_K) : K \in \mathcal{M}\}$. To avoid technicalities we shall assume that we equip each cell with the same type of element.

We can thus give our first specification of a finite element space. Suppose we solve a variational problem over V. Then we take $V_h \subset V$ defined by

$$V_h := \{v \in V : v|_K \in \mathcal{V}_K \quad \forall\, K \in \mathcal{M}\}.$$

Let us consider $V = H^1(\Omega)$. To enforce that $V_h \subset H^1(\Omega)$, we need to make sure that functions in the space are continuous. How do we do that?

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

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

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:

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

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

$$\iota_K : \{1, \dots, d(K)\} \to \{1, \dots, N\}$$

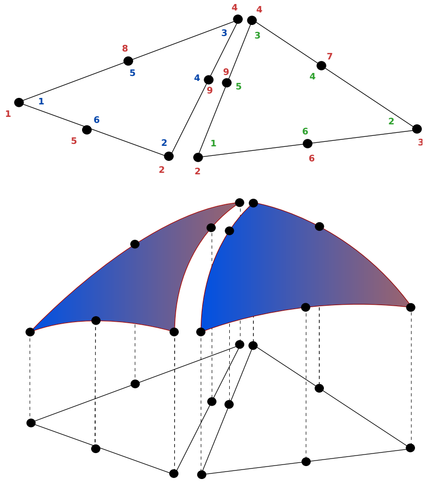
which specifies how the *local* degrees of freedom $\ell_i^K(v)$ relate to the *global* degrees of freedom. Each local degree of freedom corresponds to a global degree of freedom, under the action of the local-to-global map:

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

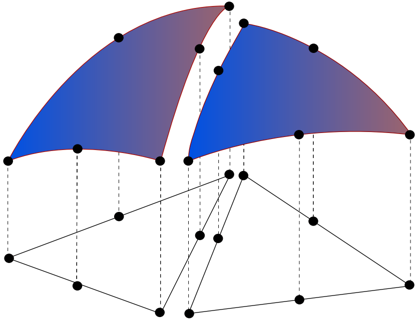
If two different degrees of freedom $\ell_i^K, \ell_j^{K'}$ on two different cells K, K' both map to the same global degree of freedom, we demand

$$\ell_i^K(v|_K) = \ell_j^{K'}(v|_{K'}),$$

i.e. matching degrees of freedom agree.



The local-to-global mapping for a mesh of two triangles, both equipped with CG_2. 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: $V_h \subset H^1(\Omega)$.



By *not* mapping matching local degrees of freedom at the common edge to the same global degree of freedom, a discontinuous approximation results: $V_h \subset L^2(\Omega)$.

The matching properties of the local-to-global map determine the global continuity of the function space, and hence which Sobolev space it *conforms* to.

Definition (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.

The matching properties of the local-to-global map determine the global continuity of the function space, and hence which Sobolev space it *conforms* to.

Definition (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.

In this course we will only consider conforming discretisations, although nonconforming ones are important, common, and useful.

Once we have the local-to-global map, we gather all of the global degrees of freedom

$$\mathcal{L} = \{\ell_1, \ldots, \ell_N\} = \bigcup_{K \in \mathcal{M}} \{\ell_{\iota_K(i)}, \quad i = 1, \ldots, d(K)\}.$$

Once we have the local-to-global map, we gather all of the global degrees of freedom

$$\mathcal{L} = \{\ell_1, \dots, \ell_N\} = \bigcup_{K \in \mathcal{M}} \{\ell_{\iota_K(i)}, \quad i = 1, \dots, d(K)\}.$$

and use its associated nodal basis as our basis for V_h:

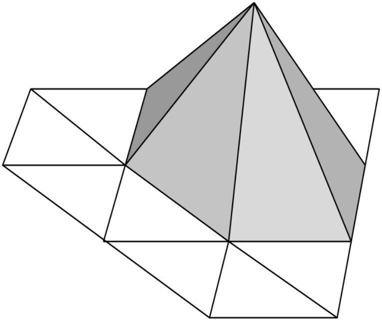
$$V_h = \mathrm{span}\{\phi_1, \dots, \phi_N\}, \quad \ell_i(\phi_j) = \delta_{ij}, \quad \phi_i|_K \in \mathcal{V}_K \quad \forall\, K \in \mathcal{M}.$$

Once we have the local-to-global map, we gather all of the global degrees of freedom

$$\mathcal{L} = \{\ell_1, \dots, \ell_N\} = \bigcup_{K \in \mathcal{M}} \{\ell_{\iota_K(i)}, \quad i = 1, \dots, d(K)\}.$$

and use its associated nodal basis as our basis for V_h:

$$V_h = \operatorname{span}\{\phi_1, \dots, \phi_N\}, \quad \ell_i(\phi_j) = \delta_{ij}, \quad \phi_i|_K \in \mathcal{V}_K \quad \forall\, K \in \mathcal{M}.$$



Reproduced from Braess (2007).

Now that we have a global function space, we can construct a global interpolation operator.

Definition (global interpolation operator)

Let V_h be a finite element function space constructed by equipping a mesh $\mathcal{M}$ with finite elements. The interpolation operator $\mathcal{I}_h : H \subset V \to V_h$ is then defined by

$$(\mathcal{I}_h u)|_K = \mathcal{I}_K u,$$

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

Now that we have a global function space, we can construct a global interpolation operator.

Definition (global interpolation operator)

Let V_h be a finite element function space constructed by equipping a mesh $\mathcal{M}$ with finite elements. The interpolation operator $\mathcal{I}_h : H \subset V \to V_h$ is then defined by

$$(\mathcal{I}_h u)|_K = \mathcal{I}_K u,$$

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

Example

If we take CG_q in 2D or 3D, we have $V = H^1(\Omega)$ and $H = H^2(\Omega)$.

C6.4 Finite Element Methods for PDEs
Lecture 9: Local and global assembly

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

We are solving the Galerkin approximation

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

over $V_h := \text{span}\{\phi_1, \ldots, \phi_N\}$.

In this lecture we study the central algorithm executed by a finite element code, the *assembly* algorithm for computing A and b of

$$Ax = b.$$

We are solving the Galerkin approximation

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

over $V_h := \operatorname{span}\{\phi_1, \ldots, \phi_N\}$.

In this lecture we study the central algorithm executed by a finite element code, the *assembly* algorithm for computing A and b of

$$Ax = b.$$

Recall that

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

The naïve algorithm for assembly:

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

has two major disadvantages:

The naïve algorithm for assembly:

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

has two major disadvantages:

- Each ϕ_i has *local* support. For most pairs i, j, $a(\phi_i, \phi_j) = 0$.

The naïve algorithm for assembly:

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

has two major disadvantages:

- Each ϕ_i has *local* support. For most pairs i, j, $a(\phi_i, \phi_j) = 0$.
- Each evaluation of a requires integrating over Ω, i.e. a loop over cells. The calculations required to integrate over each cell are repeated many times.

A better idea:

- Loop over each cell of the mesh once.
- Calculate all contributions of each cell to all entries that it influences.

A better idea:

- Loop over each cell of the mesh once.
- Calculate all contributions of each cell to all entries that it influences.

Notation: $(K, \mathcal{V}_K, \mathcal{L}_K)$, $d = \dim(\mathcal{V}_K)$, ι_K, $\phi_i^K = \phi_{\iota_K(i)}\big|_K$.

A better idea:

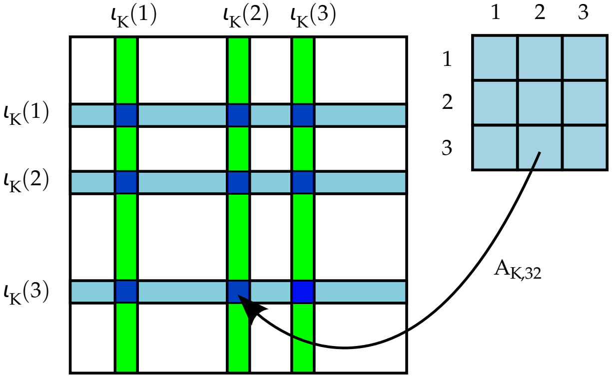
- Loop over each cell of the mesh once.
- Calculate all contributions of each cell to all entries that it influences.

Notation: $(K, \mathcal{V}_K, \mathcal{L}_K)$, $d = \dim(\mathcal{V}_K)$, ι_K, $\phi_i^K = \phi_{\iota_K(i)}\big|_K$.

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

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

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

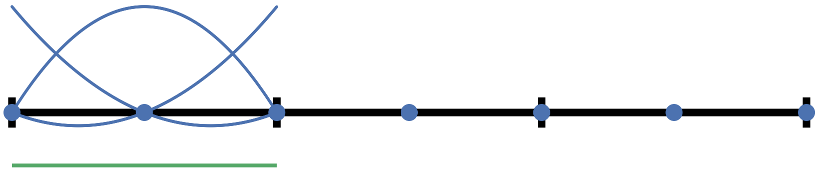


Finite element assembly. We loop over each cell K of the mesh and assemble the local stiffness matrix A_K (top right). We add this matrix to the submatrix of the global stiffness matrix A formed by taking the rows and columns associated with the local-to-global map ι_K.

Assembly of a CG_2 discretisation in 1D.

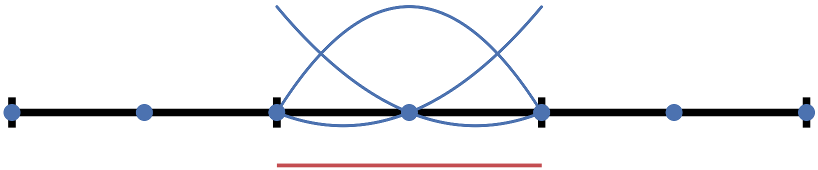
$$
\begin{bmatrix}
\times & \times & \times & 0 & 0 & 0 & 0 \\
\times & \times & \times & 0 & 0 & 0 & 0 \\
\times & \times & \times & \times & \times & 0 & 0 \\
0 & 0 & \times & \times & \times & 0 & 0 \\
0 & 0 & \times & \times & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times & \times \\
0 & 0 & 0 & 0 & \times & \times & \times
\end{bmatrix}
$$

Assembly of a CG_2 discretisation in 1D.



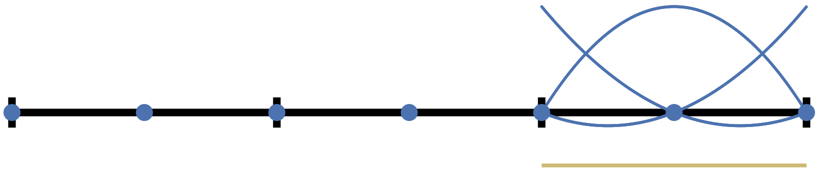
$$\begin{bmatrix} \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times & \times \end{bmatrix}$$

Assembly of a CG_2 discretisation in 1D.



$$\begin{bmatrix} \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times & \times \end{bmatrix}$$

Assembly of a CG_2 discretisation in 1D.



$$\begin{bmatrix} \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & 0 & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & 0 & 0 \\ 0 & 0 & \times & \times & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times & \times \\ 0 & 0 & 0 & 0 & \times & \times & \times \end{bmatrix}$$

Section 2

Assembling the local matrix

How do we assemble the local matrix on a cell?

$$\int_K \phi_i(x)\phi_j(x)\ \mathrm{d}x \quad \text{or} \quad \int_K \nabla\phi_i(x)\cdot\nabla\phi_j(x)\ \mathrm{d}x \quad \text{or} \ \ldots$$

How do we assemble the local matrix on a cell?

$$\int_K \phi_i(x)\phi_j(x)\ \mathrm{d}x \quad \text{or} \quad \int_K \nabla\phi_i(x)\cdot\nabla\phi_j(x)\ \mathrm{d}x \quad \text{or} \ \ldots$$

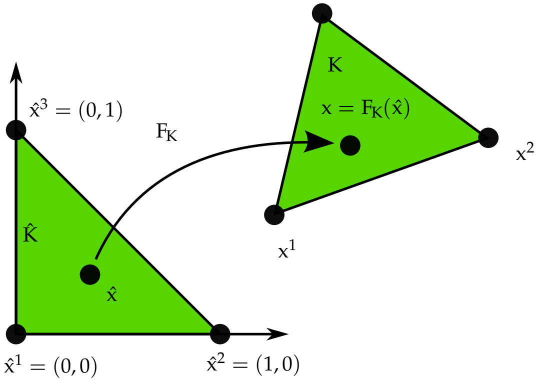
We calculate the integral with a *quadrature rule* on a *reference element*.

Introduce a reference element

$$(\hat{K}, \hat{\mathcal{V}}, \hat{\mathcal{L}})$$

and a set of diffeomorphisms

$$\{F_K : K \in \mathcal{M}\} \text{ such that } K = F_K(\hat{K}) \text{ for all } K \in \mathcal{M}.$$



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

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

and a set of degrees of freedom on K via

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

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

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

and a set of degrees of freedom on K via

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

By construction, we obtain a nodal basis for $\mathcal{V}(K)$ from one on $\hat{\mathcal{V}}$. Suppose $\{\hat{\phi}_i\}_{i=1}^{d}$ satisfies

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

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

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

For this simple approach of mapping to a reference element to work, we need

$$\mathcal{V}(K) = \mathcal{V}_K, \quad \mathcal{L}(K) = \mathcal{L}_K,$$

i.e. the finite element constructed via transformation is the same as that constructed directly on the cell.

For this simple approach of mapping to a reference element to work, we need

$$\mathcal{V}(K) = \mathcal{V}_K, \quad \mathcal{L}(K) = \mathcal{L}_K,$$

i.e. the finite element constructed via transformation is the same as that constructed directly on the cell.

This is true for Lagrange finite elements, and more complicated maps make it true for other finite elements we will meet.

Example 1. Suppose we need to calculate

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

Example 1. Suppose we need to calculate

$$\int_K \phi_i(x)\phi_j(x)\ \mathrm{d}x.$$

We transform coordinates in the integral, at the cost of the determinant of the Jacobian of the mapping:

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

where $J_K(\hat{x})$ is the Jacobian of $F_K(\hat{x})$.

Example 1. Suppose we need to calculate

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

We transform coordinates in the integral, at the cost of the determinant of the Jacobian of the mapping:

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

where $J_K(\hat{x})$ is the Jacobian of $F_K(\hat{x})$.

We then approximate the integral with a *quadrature rule*:

$$\int_{\hat{K}} f(\hat{x})\,\mathrm{d}\hat{x} \approx \sum_{i=1}^{q} w_i f(\hat{x}_i).$$

Definition (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

$$\int_{\hat{K}} f(\hat{x}) \, \mathrm{d}\hat{x} \approx \sum_{i=1}^{q} w_i f(\hat{x}_i).$$

It has *degree of precision* (or degree) m if the approximation is exact for polynomials of degree m or less.

Definition (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

$$\int_{\hat{K}} f(\hat{x})\ \mathrm{d}\hat{x} \approx \sum_{i=1}^{q} w_i f(\hat{x}_i).$$

It has *degree of precision* (or degree) m if the approximation is exact for polynomials of degree m or less.

In 1D, Gaussian quadrature gives us the optimal choice of weights and quadrature points to maximise the degree of the rule. For q points in an interval we get $m = 2q - 1$. In higher dimensions things are not as simple, and the best quadrature rules are collated in an encyclopaedia.

Summary

Offline calculations:

- Quadrature rule on the reference cell
- Basis functions at the quadrature points of the reference cell

Online calculations:

- Coordinate transformation & Jacobian

Summary

Offline calculations:

- Quadrature rule on the reference cell
- Basis functions at the quadrature points of the reference cell

Online calculations:

- Coordinate transformation & Jacobian

In particular we do *not* need to calculate the basis functions for each cell.

Example 2. Consider

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \, \mathrm{d}x,$$

which is what we need to calculate for solving Poisson.

Example 2. Consider

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \, \mathrm{d}x,$$

which is what we need to calculate for solving Poisson.

Transforming, we get

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \, \mathrm{d}x = \int_{\hat{K}} \nabla_x \hat{\phi}_i(\hat{x}) \cdot \nabla_x \hat{\phi}_j(\hat{x}) |J_K(\hat{x})| \, \mathrm{d}\hat{x},$$

but this is still not computable because it requires derivatives with respect to the physical coordinate.

Example 2. Consider

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \; \mathrm{d}x,$$

which is what we need to calculate for solving Poisson.

Transforming, we get

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \; \mathrm{d}x = \int_{\hat{K}} \nabla_x \hat{\phi}_i(\hat{x}) \cdot \nabla_x \hat{\phi}_j(\hat{x}) |J_K(\hat{x})| \; \mathrm{d}\hat{x},$$

but this is still not computable because it requires derivatives with respect to the physical coordinate.

To replace these, we apply the chain rule:

$$\frac{\partial \phi}{\partial x_k} = \sum_{l=1}^{n} \frac{\partial \phi}{\partial \hat{x}_l} \frac{\partial \hat{x}_l}{\partial x_k} = \sum_{l=1}^{n} \frac{\partial \hat{x}_l}{\partial x_k} \frac{\partial \phi}{\partial \hat{x}_l}, \qquad k = 1, \ldots, n.$$

Some calculation finds that

$$\nabla_x \hat{\phi}(\hat{x}) = J_K^{-\mathrm{T}}(\hat{x}) \nabla_{\hat{x}} \hat{\phi}(\hat{x}).$$

Thus, finally, we write

$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \, \mathrm{d}x = \int_{\hat{K}} \left(J_K^{-\mathrm{T}} \nabla_{\hat{x}} \hat{\phi}_i(\hat{x}) \right) \cdot \left(J_K^{-\mathrm{T}} \nabla_{\hat{x}} \hat{\phi}_j(\hat{x}) \right) |J_K(\hat{x})| \, \mathrm{d}\hat{x}.$$

Thus, finally, we write

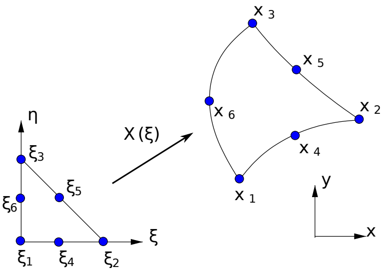
$$\int_K \nabla_x \phi_i(x) \cdot \nabla_x \phi_j(x) \,\mathrm{d}x = \int_{\hat{K}} \left(J_K^{-\mathrm{T}} \nabla_{\hat{x}} \hat{\phi}_i(\hat{x}) \right) \cdot \left(J_K^{-\mathrm{T}} \nabla_{\hat{x}} \hat{\phi}_j(\hat{x}) \right) |J_K(\hat{x})| \,\mathrm{d}\hat{x}.$$

Remarks:

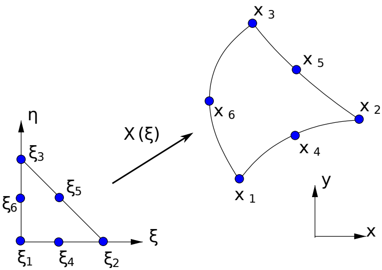
- Modern finite element software does this for you.
- We need to calculate the gradients of the basis functions at quadrature points (offline).

Section 3

Representing the element map



We represent the coordinate field with Lagrange elements of arbitrary order, allowing us to bend the mesh. This is useful if Ω is not a polygon or polyhedron.



We represent the coordinate field with Lagrange elements of arbitrary order, allowing us to bend the mesh. This is useful if Ω is not a polygon or polyhedron.

This means that for each element we can write

$$x = \sum_{i=1}^{d} x_i \hat{\psi}_i(\hat{x})$$

for (scalar-valued) coefficients x_i and (vector-valued) basis functions $\hat{\psi}_i$. This is an explicit construction for the map $x = F_K(\hat{x})$.

C6.4 Finite Element Methods for PDEs
Lecture 10: Finite elements beyond Lagrange

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

We are solving the linear variational problem

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

We are solving the linear variational problem

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

One of the great advantages of the finite element method is that one can *tailor* the approximation space V_h to the function space V. If $V \neq H^1(\Omega)$ then we usually want to use different finite elements than Lagrange.

We are solving the linear variational problem

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

One of the great advantages of the finite element method is that one can *tailor* the approximation space V_h to the function space V. If $V \neq H^1(\Omega)$ then we usually want to use different finite elements than Lagrange.

The fundamental Hilbert spaces we have met are related by the *de Rham* complex:

$$H^1(\Omega) \xrightarrow{\mathrm{grad}} H(\mathrm{curl};\Omega) \xrightarrow{\mathrm{curl}} H(\mathrm{div};\Omega) \xrightarrow{\mathrm{div}} L^2(\Omega)$$

We are solving the linear variational problem

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

One of the great advantages of the finite element method is that one can *tailor* the approximation space V_h to the function space V. If $V \neq H^1(\Omega)$ then we usually want to use different finite elements than Lagrange.

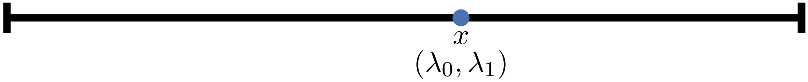
The fundamental Hilbert spaces we have met are related by the *de Rham* complex:

$$\begin{array}{ccccccc}
H^1(\Omega) & \xrightarrow{\text{grad}} & H(\text{curl};\Omega) & \xrightarrow{\text{curl}} & H(\text{div};\Omega) & \xrightarrow{\text{div}} & L^2(\Omega) \\
\downarrow & & \downarrow & & \downarrow & & \downarrow \\
\Sigma_h & \xrightarrow{\text{grad}} & V_h & \xrightarrow{\text{curl}} & W_h & \xrightarrow{\text{div}} & Q_h
\end{array}$$

We can build finite element spaces for *all* spaces in this de Rham complex in a structure-preserving way.

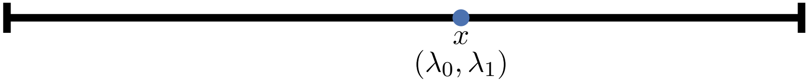
On the interval $\Omega = [0, 1]$, the obvious way to label a point is with a single coordinate x.

On the interval $\Omega = [0,1]$, the obvious way to label a point is with a single coordinate x.



It will be convenient for us to label points with *barycentric coordinates*: $n+1$ coordinates for a simplex in $\mathbb{R}^n$ with the constraint that $\sum_i \lambda_i = 1$.

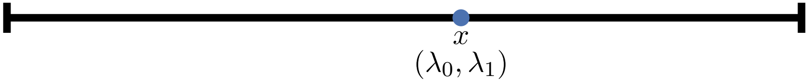
On the interval $\Omega = [0, 1]$, the obvious way to label a point is with a single coordinate x.



It will be convenient for us to label points with *barycentric coordinates*: $n+1$ coordinates for a simplex in $\mathbb{R}^n$ with the constraint that $\sum_i \lambda_i = 1$.

In 1D, $\lambda_0 = 1 - x$, $\lambda_1 = x$.

On the interval $\Omega = [0, 1]$, the obvious way to label a point is with a single coordinate x.

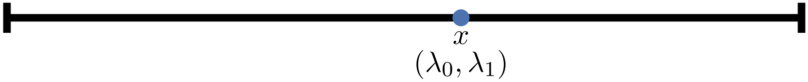


It will be convenient for us to label points with *barycentric coordinates*: $n + 1$ coordinates for a simplex in $\mathbb{R}^n$ with the constraint that $\sum_i \lambda_i = 1$.

In 1D, $\lambda_0 = 1 - x$, $\lambda_1 = x$.

The reason why this is convenient is because it gives us a nice way to describe different geometric parts of our simplex. In 1D, we have the left vertex is given by $\lambda_0 = 1$, the right vertex by $\lambda_1 = 1$.

On the interval $\Omega = [0, 1]$, the obvious way to label a point is with a single coordinate x.

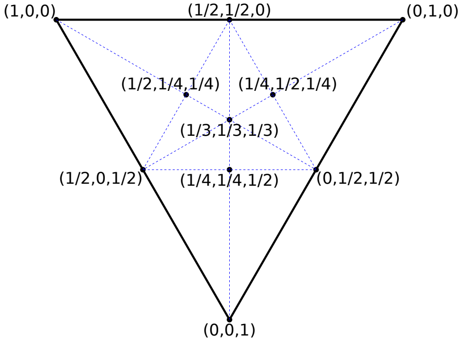


It will be convenient for us to label points with *barycentric coordinates*: $n + 1$ coordinates for a simplex in $\mathbb{R}^n$ with the constraint that $\sum_i \lambda_i = 1$.

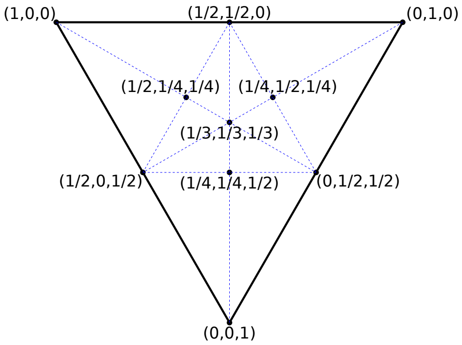
In 1D, $\lambda_0 = 1 - x$, $\lambda_1 = x$.

The reason why this is convenient is because it gives us a nice way to describe different geometric parts of our simplex. In 1D, we have the left vertex is given by $\lambda_0 = 1$, the right vertex by $\lambda_1 = 1$.

Another way to look at it: the barycentric coordinates express a point p as a convex combination of the vertices.



For a triangle, we represent points $p = (\lambda_1, \lambda_2, \lambda_3)$; three numbers, but only two degrees of freedom because of the summation constraint.



For a triangle, we represent points $p = (\lambda_1, \lambda_2, \lambda_3)$; three numbers, but only two degrees of freedom because of the summation constraint.

The edge opposite vertex $(1, 0, 0)$ is described by $\lambda_1 = 0$; the edge opposite vertex $(0, 1, 0)$ is described by $\lambda_2 = 0$; the edge opposite vertex $(0, 0, 1)$ is described by $\lambda_3 = 0$.

Lemma (First factorisation lemma)

Let P be a polynomial of degree $d \geq 1$ that vanishes on a hyperplane $\{x : L(x) = 0\}$, where $L(x)$ is a non-degenerate linear function. Then we can write $P = LQ$, where Q is a polynomial of degree $d - 1$.

Lemma (First factorisation lemma)

Let P be a polynomial of degree $d \geq 1$ that vanishes on a hyperplane $\{x : L(x) = 0\}$, where $L(x)$ is a non-degenerate linear function. Then we can write $P = LQ$, where Q is a polynomial of degree $d - 1$.

Example (One dimension)

Suppose P vanishes on the hyperplane $\{x : x - r = 0\}$. Then we can write $P = (x - r)Q$ for some polynomial Q.

Lemma (First factorisation lemma)

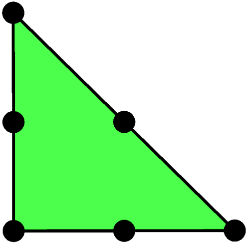
Let P be a polynomial of degree $d \geq 1$ that vanishes on a hyperplane $\{x : L(x) = 0\}$, where $L(x)$ is a non-degenerate linear function. Then we can write $P = LQ$, where Q is a polynomial of degree $d - 1$.

Example (One dimension)

Suppose P vanishes on the hyperplane $\{x : x - r = 0\}$. Then we can write $P = (x - r)Q$ for some polynomial Q.

Example (Higher dimensions)

Suppose P vanishes on each edge of a triangle. Then $P = \lambda_1 \lambda_2 \lambda_3 Q$ for some Q.



The quadratic Lagrange finite element CG_2 in two dimensions.

Unisolvence of CG_2

Suppose $v \in \mathcal{P}_2(\triangle)$ with all degrees of freedom zero. Restricted to an edge, v is a quadratic polynomial with three roots, hence $v = 0$ on each edge. By the factorisation lemma, $v = \lambda_1 \lambda_2 c$ for a constant $c \in \mathbb{R}$. Evaluating both sides on the edge $\lambda_3 = 0$ shows that $c = 0$.

Lemma (Second factorisation lemma)

Let P be a polynomial of degree $d \geq 2$ such that P and $\nabla P \cdot n$ vanish on a hyperplane $\{x : L(x) = 0\}$, where n is the normal to L. Then we can write $P = L^2 Q$, where Q is a polynomial of degree $d - 2$.

Lemma (Second factorisation lemma)

Let P be a polynomial of degree $d \geq 2$ such that P and $\nabla P \cdot n$ vanish on a hyperplane $\{x : L(x) = 0\}$, where n is the normal to L. Then we can write $P = L^2 Q$, where Q is a polynomial of degree $d - 2$.

Remark

If P vanishes on $\{x : L(x) = 0\}$, so does $\nabla P \cdot t$ for a tangent vector t.

Lemma (Second factorisation lemma)

Let P be a polynomial of degree $d \geq 2$ such that P and $\nabla P \cdot n$ vanish on a hyperplane $\{x : L(x) = 0\}$, where n is the normal to L. Then we can write $P = L^2 Q$, where Q is a polynomial of degree $d - 2$.

Remark

If P vanishes on $\{x : L(x) = 0\}$, so does $\nabla P \cdot t$ for a tangent vector t.

Proof.

Since P vanishes on $\{x : L(x) = 0\}$, we have $P = L\tilde{Q}$.

Lemma (Second factorisation lemma)

Let P be a polynomial of degree $d \geq 2$ such that P and $\nabla P \cdot n$ vanish on a hyperplane $\{x : L(x) = 0\}$, where n is the normal to L. Then we can write $P = L^2 Q$, where Q is a polynomial of degree $d - 2$.

Remark

If P vanishes on $\{x : L(x) = 0\}$, so does $\nabla P \cdot t$ for a tangent vector t.

Proof.

Since P vanishes on $\{x : L(x) = 0\}$, we have $P = L\tilde{Q}$. Calculating,

$$\nabla P \cdot n = \tilde{Q} \nabla L \cdot n + L \nabla \tilde{Q} \cdot n.$$

Lemma (Second factorisation lemma)

Let P be a polynomial of degree $d \geq 2$ such that P and $\nabla P \cdot n$ vanish on a hyperplane $\{x : L(x) = 0\}$, where n is the normal to L. Then we can write $P = L^2 Q$, where Q is a polynomial of degree $d - 2$.

Remark

If P vanishes on $\{x : L(x) = 0\}$, so does $\nabla P \cdot t$ for a tangent vector t.

Proof.

Since P vanishes on $\{x : L(x) = 0\}$, we have $P = L\tilde{Q}$. Calculating,

$$\nabla P \cdot n = \tilde{Q} \nabla L \cdot n + L \nabla \tilde{Q} \cdot n.$$

Since L vanishes on the plane, and ∇L is normal to the plane (hence colinear with n), this forces $\tilde{Q} = 0$ on $\{x : L(x) = 0\}$.

Lemma (Second factorisation lemma)

Let P be a polynomial of degree $d \geq 2$ such that P and $\nabla P \cdot n$ vanish on a hyperplane $\{x : L(x) = 0\}$, where n is the normal to L. Then we can write $P = L^2 Q$, where Q is a polynomial of degree $d - 2$.

Remark

If P vanishes on $\{x : L(x) = 0\}$, so does $\nabla P \cdot t$ for a tangent vector t.

Proof.

Since P vanishes on $\{x : L(x) = 0\}$, we have $P = L\tilde{Q}$. Calculating,

$$\nabla P \cdot n = \tilde{Q} \nabla L \cdot n + L \nabla \tilde{Q} \cdot n.$$

Since L vanishes on the plane, and ∇L is normal to the plane (hence colinear with n), this forces $\tilde{Q} = 0$ on $\{x : L(x) = 0\}$. Hence $\tilde{Q} = LQ$ for some Q. $\square$

Section 3

The biharmonic problem

The biharmonic equation arises in many areas of physics. It describes equilibrium configurations of clamped plates under transverse loading, the stresses in an elastic body, the stream function in certain flow regimes, and other things. The equation is

$$\begin{aligned} \nabla^4 u = f \quad & \text{in } \Omega, \\ u = 0 \quad & \text{on } \partial\Omega, \\ \nabla u \cdot n = 0 \quad & \text{on } \partial\Omega. \end{aligned}$$

The biharmonic equation arises in many areas of physics. It describes equilibrium configurations of clamped plates under transverse loading, the stresses in an elastic body, the stream function in certain flow regimes, and other things. The equation is

$$\begin{aligned} \nabla^4 u = f &\quad \text{in } \Omega, \\ u = 0 &\quad \text{on } \partial\Omega, \\ \nabla u \cdot n = 0 &\quad \text{on } \partial\Omega. \end{aligned}$$

Here $\nabla^4 = \nabla \cdot (\nabla(\nabla \cdot \nabla)) = \Delta^2$. More simply, in two dimensions

$$\frac{\partial^4 u}{\partial x^4} + 2\frac{\partial^4 u}{\partial x^2 \partial y^2} + \frac{\partial^4 u}{\partial y^4} = f(x, y).$$

Let us cast this into variational form formally, i.e. not yet specifying V. Multiplying by $v \in V$, for a suitable function space V to be chosen, gives

$$\int_\Omega (\nabla^4 u) v \, \mathrm{d}x = \int_\Omega (\nabla \cdot (\nabla(\nabla \cdot \nabla u))) v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x.$$

Let us cast this into variational form formally, i.e. not yet specifying V. Multiplying by $v \in V$, for a suitable function space V to be chosen, gives

$$\int_\Omega (\nabla^4 u) v \, \mathrm{d}x = \int_\Omega (\nabla \cdot (\nabla(\nabla \cdot \nabla u))) v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x.$$

We want to invoke Lax–Milgram, so we want $u \in V$ and $v \in V$. Let us integrate by parts once:

$$\int_\Omega (\nabla^4 u) v \, \mathrm{d}x = -\int_\Omega (\nabla(\nabla \cdot \nabla u)) \cdot \nabla v \, \mathrm{d}x + \int_{\partial\Omega} \nabla (\nabla \cdot \nabla u) \cdot n \, v \, \mathrm{d}s$$

Let us cast this into variational form formally, i.e. not yet specifying V. Multiplying by $v \in V$, for a suitable function space V to be chosen, gives

$$\int_\Omega (\nabla^4 u) v \,\mathrm{d}x = \int_\Omega (\nabla \cdot (\nabla(\nabla \cdot \nabla u))) v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x.$$

We want to invoke Lax–Milgram, so we want $u \in V$ and $v \in V$. Let us integrate by parts once:

$$\int_\Omega (\nabla^4 u) v \,\mathrm{d}x = -\int_\Omega (\nabla(\nabla \cdot \nabla u)) \cdot \nabla v \,\mathrm{d}x + \int_{\partial\Omega} \nabla\left(\nabla \cdot \nabla u\right) \cdot n\, v \,\mathrm{d}s$$

and again:

$$\begin{aligned}\int_\Omega (\nabla^4 u) v \,\mathrm{d}x &= \int_\Omega \left(\nabla \cdot \nabla u\right)\left(\nabla \cdot \nabla v\right) \,\mathrm{d}x \\ &\quad - \int_{\partial\Omega} \left(\nabla \cdot \nabla u\right)\left(\nabla v \cdot n\right) \mathrm{d}s \\ &\quad + \int_{\partial\Omega} \nabla\left(\nabla \cdot \nabla u\right) \cdot n\, v \,\mathrm{d}s.\end{aligned}$$

Noting that the Laplacian $\nabla^2 = \nabla \cdot \nabla = \Delta$, we rewrite

$$\int_\Omega (\nabla^4 u) v \,\mathrm{d}x = \int_\Omega \nabla^2 u \,\nabla^2 v \,\mathrm{d}x - \int_{\partial\Omega} \nabla^2 u \,\nabla v \cdot n \,\mathrm{d}s + \int_{\partial\Omega} \nabla \left(\nabla^2 u\right) \cdot n v \,\mathrm{d}s.$$

Noting that the Laplacian $\nabla^2 = \nabla \cdot \nabla = \Delta$, we rewrite

$$\int_\Omega (\nabla^4 u) v \, \mathrm{d}x = \int_\Omega \nabla^2 u \, \nabla^2 v \, \mathrm{d}x - \int_{\partial\Omega} \nabla^2 u \, \nabla v \cdot n \, \mathrm{d}s + \int_{\partial\Omega} \nabla \left(\nabla^2 u \right) \cdot n v \, \mathrm{d}s.$$

We need square-integrable second derivatives, so $V \subset H^2(\Omega)$.

Noting that the Laplacian $\nabla^2 = \nabla \cdot \nabla = \Delta$, we rewrite

$$\int_\Omega (\nabla^4 u) v \,\mathrm{d}x = \int_\Omega \nabla^2 u \, \nabla^2 v \,\mathrm{d}x - \int_{\partial\Omega} \nabla^2 u \, \nabla v \cdot n \,\mathrm{d}s + \int_{\partial\Omega} \nabla \left(\nabla^2 u\right) \cdot n v \,\mathrm{d}s.$$

We need square-integrable second derivatives, so $V \subset H^2(\Omega)$.

We have nowhere convenient to enforce the boundary conditions $u = \nabla u \cdot n = 0$. So we should take

$$V = H_0^2(\Omega) := \{v \in H^2(\Omega) : v = 0,\ \nabla v \cdot n = 0 \text{ on } \partial\Omega\}.$$

Noting that the Laplacian $\nabla^2 = \nabla \cdot \nabla = \Delta$, we rewrite

$$\int_\Omega (\nabla^4 u) v \,\mathrm{d}x = \int_\Omega \nabla^2 u \,\nabla^2 v \,\mathrm{d}x - \int_{\partial\Omega} \nabla^2 u \,\nabla v \cdot n \,\mathrm{d}s + \int_{\partial\Omega} \nabla \left(\nabla^2 u\right) \cdot n v \,\mathrm{d}s.$$

We need square-integrable second derivatives, so $V \subset H^2(\Omega)$.

We have nowhere convenient to enforce the boundary conditions $u = \nabla u \cdot n = 0$. So we should take

$$V = H_0^2(\Omega) := \{v \in H^2(\Omega) : v = 0,\ \nabla v \cdot n = 0 \text{ on } \partial\Omega\}.$$

With $v \in V$, the surface integrals vanish, leaving us with the problem: find $u \in H_0^2(\Omega)$ such that

$$\int_\Omega \nabla^2 u \,\nabla^2 v \ \mathrm{d}x = \int_\Omega f v \ \mathrm{d}x \text{ for all } v \in H_0^2(\Omega).$$

How do we discretise this problem? If we take $V_h \sim \mathrm{CG}_p$, $V_h \not\subset H^2(\Omega)$!

How do we discretise this problem? If we take $V_h \sim \mathrm{CG}_p$, $V_h \not\subset H^2(\Omega)$!

For a piecewise smooth function u, $u \in H^1(\Omega) \iff u \in C^0(\Omega)$.

Since $u \in H^2(\Omega)$ iff u and all its first derivatives are in $H^1(\Omega)$, that means for $u \in H^2(\Omega)$ we need $u \in C^1(\Omega)$.

How do we discretise this problem? If we take $V_h \sim \mathrm{CG}_p$, $V_h \not\subset H^2(\Omega)$!

For a piecewise smooth function u, $u \in H^1(\Omega) \iff u \in C^0(\Omega)$.

Since $u \in H^2(\Omega)$ iff u and all its first derivatives are in $H^1(\Omega)$, that means for $u \in H^2(\Omega)$ we need $u \in C^1(\Omega)$.

Two approaches:

- C^1-continuous finite elements;
- nonconforming discretisations.

Section 4

The Hermite element

Definition (Hermite finite element)

$K = [0, 1]$, $\mathcal{V} = \mathcal{P}_3(K)$, and

$$\begin{aligned}\mathcal{L} = \{&v \mapsto v(0),\\ &v \mapsto v'(0),\\ &v \mapsto v(1),\\ &v \mapsto v'(1)\}.\end{aligned}$$

Definition (Hermite finite element)

$K = [0,1]$, $\mathcal{V} = \mathcal{P}_3(K)$, and

$$\begin{aligned}\mathcal{L} = \{v &\mapsto v(0),\\ v &\mapsto v'(0),\\ v &\mapsto v(1),\\ v &\mapsto v'(1)\}.\end{aligned}$$

This gives a C^1 approximation, because the function value *and derivative* agree across cells (by construction).

Definition (Hermite finite element)

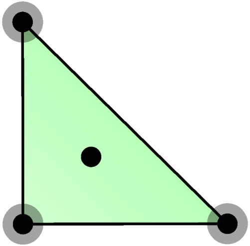
$K = [0, 1]$, $\mathcal{V} = \mathcal{P}_3(K)$, and

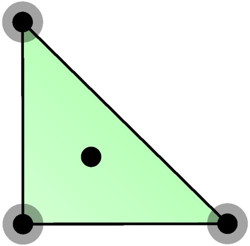
$$\begin{aligned}\mathcal{L} = \{v &\mapsto v(0),\\ v &\mapsto v'(0),\\ v &\mapsto v(1),\\ v &\mapsto v'(1)\}.\end{aligned}$$

This gives a C^1 approximation, because the function value *and derivative* agree across cells (by construction).

Unisolvence

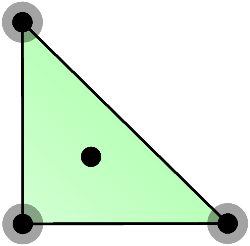
Suppose $v \in \mathcal{P}_3(K)$ with all dofs zero. Then v is a cubic polynomial with four roots (two double roots), hence zero.





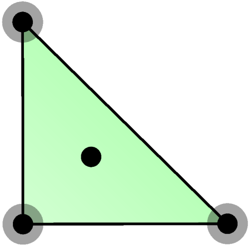
Definition (Hermite element in 2D)

$K = \triangle, \mathcal{V} = \mathcal{P}_3(\triangle), \mathcal{L}$ shown.



Lemma (Unisolvence of the triangular Hermite element)

The Hermite element in two dimensions is unisolvent.



Lemma (Unisolvence of the triangular Hermite element)

The Hermite element in two dimensions is unisolvent.

Proof.

Suppose $u \in \mathcal{P}_3(\triangle)$ with all dofs zero. Along an edge of the triangle, u is a cubic polynomial with 2 double roots, so $u = 0$ along each edge. Thus, $u(x) = c\lambda_1\lambda_2\lambda_3$ for some $c \in \mathbb{R}$. Since the value at the barycentre is also zero, and $\lambda_1 = \lambda_2 = \lambda_3 = 1/3$ at the barycentre, we must have $c = 0$. $\square$

Job done?

Job done? No!

Job done? No!

In the unisolvence proof we saw the dofs on an edge (i.e. those shared with a neighbour) determine u and thus $\nabla u \cdot t$, t the tangent vector of the edge. But they do *not* determine $\nabla u \cdot n$.

Job done? No!

In the unisolvence proof we saw the dofs on an edge (i.e. those shared with a neighbour) determine u and thus $\nabla u \cdot t$, t the tangent vector of the edge. But they do *not* determine $\nabla u \cdot n$.

For two elements, take

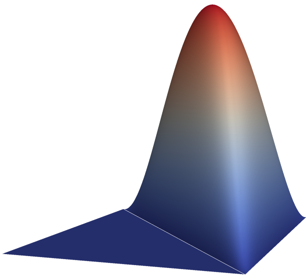
$$p(x) = \begin{cases} \lambda_1 \lambda_2 \lambda_3 & x \in K_1, \\ 0 & x \in K_2, \end{cases}$$

Job done? No!

In the unisolvence proof we saw the dofs on an edge (i.e. those shared with a neighbour) determine u and thus $\nabla u \cdot t$, t the tangent vector of the edge. But they do *not* determine $\nabla u \cdot n$.

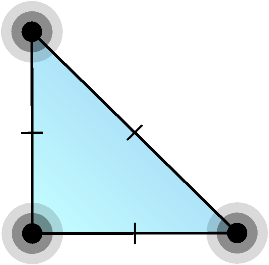
For two elements, take

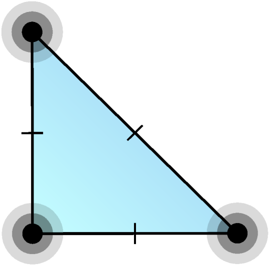
$$p(x) = \begin{cases} \lambda_1 \lambda_2 \lambda_3 & x \in K_1, \\ 0 & x \in K_2, \end{cases}$$



Section 5

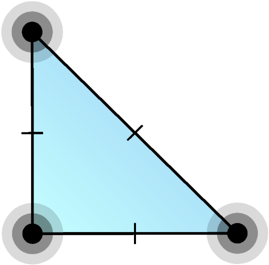
The Argyris element





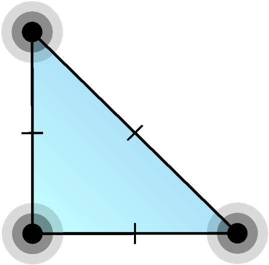
Definition (Hermite element in 2D)

$K = \triangle, \mathcal{V} = \mathcal{P}_5(\triangle), \mathcal{L}$ shown.



Lemma (Unisolvence of the triangular Argyris element)

The Argyris element in two dimensions is unisolvent.



Proof.

Suppose $u \in \mathcal{P}_5(\triangle)$ with all dofs zero. Along an edge, u is a quintic polynomial with 2 treble roots, so $u = 0$ along each edge. Moreover, $\nabla u \cdot n$ is a quartic polynomial with 2 double roots and a single root, hence zero. Thus, u is divisible by $\lambda_1^2\lambda_2^2\lambda_3^2$, which is of degree 6. Thus $u = 0$. $\square$

```python
from firedrake import *

mesh = UnitSquareMesh(20, 20)
V = FunctionSpace(mesh, "Argyris", 5)

u = Function(V)
v = TestFunction(V)

(x, y) = SpatialCoordinate(mesh)
f = sin(2*pi*x) * sin(2*pi*y)

F = (inner(div(grad(u)), div(grad(v)))*dx
    + inner(u, v)*dx
    - inner(f, v)*dx)
solve(F == 0, u)
```

dim	lowest p for $V_h \subset H^2(\Omega)$
1	3 (4 degrees of freedom)
2	5 (21 degrees of freedom)
3	9 (136 degrees of freedom)

dim	lowest p for $V_h \subset H^2(\Omega)$
1	3 (4 degrees of freedom)
2	5 (21 degrees of freedom)
3	9 (136 degrees of freedom)

An alternative: *nonconforming* discretisations. Use Lagrange elements and *weakly* enforce C^1 continuity by penalising the jump in $\nabla u \cdot n$ across edges.

dim	lowest p for $V_h \subset H^2(\Omega)$
1	3 (4 degrees of freedom)
2	5 (21 degrees of freedom)
3	9 (136 degrees of freedom)

An alternative: *nonconforming* discretisations. Use Lagrange elements and *weakly* enforce C^1 continuity by penalising the jump in $\nabla u \cdot n$ across edges.

$$a_h(u,v) = \sum_{K \in \mathcal{M}} \int_K \nabla^2 u \, \nabla^2 v \, \mathrm{d}x + \gamma \sum_{E \in \mathcal{E}_h} \frac{1}{|E|} \int_E [\![\nabla u \cdot n]\!] [\![\nabla v \cdot n]\!] \, \mathrm{d}s + \cdots$$

dim	lowest p for $V_h \subset H^2(\Omega)$
1	3 (4 degrees of freedom)
2	5 (21 degrees of freedom)
3	9 (136 degrees of freedom)

An alternative: *nonconforming* discretisations. Use Lagrange elements and *weakly* enforce C^1 continuity by penalising the jump in $\nabla u \cdot n$ across edges.

$$a_h(u,v) = \sum_{K \in \mathcal{M}} \int_K \nabla^2 u \, \nabla^2 v \, \mathrm{d}x + \gamma \sum_{E \in \mathcal{E}_h} \frac{1}{|E|} \int_E [\![\nabla u \cdot n]\!] [\![\nabla v \cdot n]\!] \, \mathrm{d}s + \cdots$$

where the jump over cells K_+, K_- is

$$[\![\nabla u \cdot n]\!] := (\nabla u)_+ \cdot n_+ + (\nabla u)_- \cdot n_-.$$

Section 7

Elements for $H(\mathrm{div})$ and $H(\mathrm{curl})$

Example 1. Find the vector-function u s.t.

$$\begin{aligned} u - \nabla(\nabla \cdot u) = f \quad &\text{in } \Omega, \\ u \cdot n = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

Example 1. Find the vector-function u s.t.

$$\begin{aligned} u - \nabla(\nabla \cdot u) = f \quad &\text{in } \Omega, \\ u \cdot n = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

Casting to variational form, we have

$$\int_\Omega u \cdot v \ \mathrm{d}x - \int_\Omega (\nabla(\nabla \cdot u)) \cdot v \ \mathrm{d}x = \int_\Omega f \cdot v \ \mathrm{d}x,$$

Example 1. Find the vector-function u s.t.

$$\begin{aligned} u - \nabla(\nabla \cdot u) &= f \quad \text{in } \Omega, \\ u \cdot n &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Casting to variational form, we have

$$\int_\Omega u \cdot v \,\mathrm{d}x - \int_\Omega (\nabla(\nabla \cdot u)) \cdot v \,\mathrm{d}x = \int_\Omega f \cdot v \,\mathrm{d}x,$$

and integrating by parts yields

$$\int_\Omega u \cdot v \,\mathrm{d}x + \int_\Omega (\nabla \cdot u)\,(\nabla \cdot v)\,\mathrm{d}x - \int_\Omega (\nabla \cdot u)\,(v \cdot n)\,\mathrm{d}s = \int_\Omega f \cdot v \,\mathrm{d}x.$$

Example 1. Find the vector-function u s.t.

$$\begin{aligned} u - \nabla(\nabla \cdot u) &= f \quad \text{in } \Omega, \\ u \cdot n &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Casting to variational form, we have

$$\int_\Omega u \cdot v \, \mathrm{d}x - \int_\Omega (\nabla(\nabla \cdot u)) \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$

and integrating by parts yields

$$\int_\Omega u \cdot v \, \mathrm{d}x + \int_\Omega (\nabla \cdot u)\,(\nabla \cdot v) \, \mathrm{d}x - \int_\Omega (\nabla \cdot u)\,(v \cdot n) \, \mathrm{d}s = \int_\Omega f \cdot v \, \mathrm{d}x.$$

The base space is $H(\mathrm{div};\Omega)$, and we need to enforce BCs:

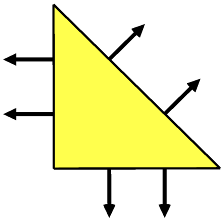
$$V = H_0(\mathrm{div};\Omega) := \{v \in H(\mathrm{div};\Omega) : v \cdot n = 0 \text{ on } \partial\Omega\},$$

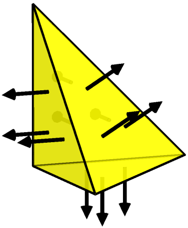
with final formulation: find $u \in V$ such that

$$\int_\Omega u \cdot v \, \mathrm{d}x + \int_\Omega (\nabla \cdot u)\,(\nabla \cdot v) \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x \quad \text{for all } v \in V.$$

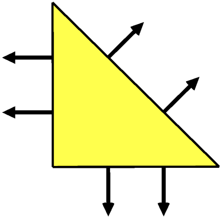
On problem sheet 1 we learn that a piecewise smooth vector-function is in $H(\mathrm{div};\Omega)$ iff its normal component is continuous.

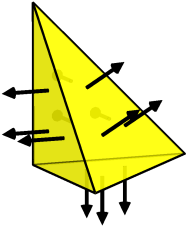
On problem sheet 1 we learn that a piecewise smooth vector-function is in $H(\mathrm{div};\Omega)$ iff its normal component is continuous.





On problem sheet 1 we learn that a piecewise smooth vector-function is in $H(\mathrm{div};\Omega)$ iff its normal component is continuous.





Definition (Lowest order Brezzi–Douglas–Marini element)

$K = \triangle$ or $K =$ (tetrahedron), $\mathcal{V} = \mathcal{P}_1(K)^n$, $\mathcal{L}$ eval normal component on facets.

Example 2. Find the vector function u such that

$$\begin{aligned} u + \nabla \times (\nabla \times u) &= f \quad \text{in } \Omega, \\ u \times n &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Example 2. Find the vector function u such that

$$\begin{aligned} u + \nabla \times (\nabla \times u) = f \quad &\text{in } \Omega, \\ u \times n = 0 \quad &\text{on } \partial\Omega. \end{aligned}$$

We end up with the variational formulation over

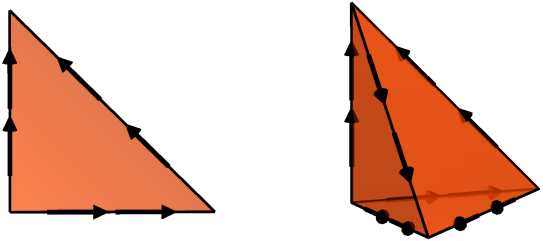
$$V := \{v \in H(\mathrm{curl}; \Omega) : v \times n = 0\}$$

to find $u \in V$ such that

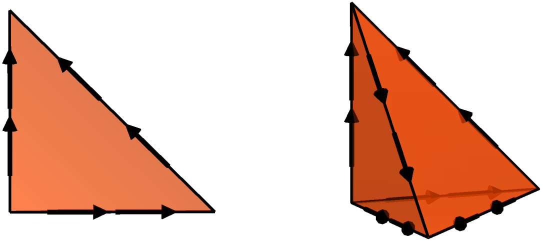
$$\int_\Omega u \cdot v \, \mathrm{d}x + \int_\Omega (\nabla \times u) \cdot (\nabla \times v) \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x \quad \text{for all } v \in V.$$

On problem sheet 1 we learn that a piecewise smooth vector-function is in $H(\mathrm{curl};\Omega)$ iff its tangential components are continuous.

On problem sheet 1 we learn that a piecewise smooth vector-function is in $H(\mathrm{curl};\Omega)$ iff its tangential components are continuous.



On problem sheet 1 we learn that a piecewise smooth vector-function is in $H(\mathrm{curl};\Omega)$ iff its tangential components are continuous.



Definition (Lowest order Nédélec element of the second kind)

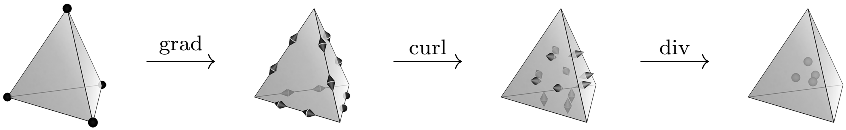
$K = \triangle$ or $K = $ (tetrahedron), $\mathcal{V} = \mathcal{P}_1(K)^n$, $\mathcal{L}$ eval tangential component on edges.

Consider again the de Rham complex:

$$H^1 \xrightarrow{\quad \mathrm{grad} \quad} H(\mathrm{curl}) \xrightarrow{\quad \mathrm{curl} \quad} H(\mathrm{div}) \xrightarrow{\quad \mathrm{div} \quad} L^2$$

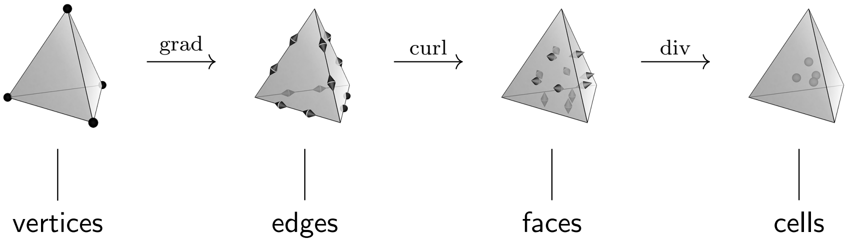
Consider again the de Rham complex:

$$H^1 \xrightarrow{\mathrm{grad}} H(\mathrm{curl}) \xrightarrow{\mathrm{curl}} H(\mathrm{div}) \xrightarrow{\mathrm{div}} L^2$$



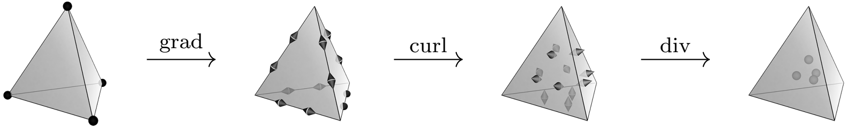
Consider again the de Rham complex:

$$H^1 \xrightarrow{\mathrm{grad}} H(\mathrm{curl}) \xrightarrow{\mathrm{curl}} H(\mathrm{div}) \xrightarrow{\mathrm{div}} L^2$$



Consider again the de Rham complex:

$$H^1 \xrightarrow{\ \text{grad}\ } H(\text{curl}) \xrightarrow{\ \text{curl}\ } H(\text{div}) \xrightarrow{\ \text{div}\ } L^2$$



vertices | edges | faces | cells

The finite element method has deep connections to algebraic topology and differential geometry: the *finite element exterior calculus*.

C6.4 Finite Element Methods for PDEs
Lecture 11: Interpolation error estimates

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

We are approximating the linear variational problem

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

with the solution of

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

We are approximating the linear variational problem

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

with the solution of

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

In Lecture 7 we saw Céa's Lemma for coercive, bounded a:

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

We are approximating the linear variational problem

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

with the solution of

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

In Lecture 7 we saw Céa's Lemma for coercive, bounded a:

$$\begin{aligned}\|u - u_h\|_V &\leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_V \\ &\leq \frac{C}{\alpha} \|u - \mathcal{I}_h u\|_V\end{aligned}$$

where $\mathcal{I}_h : V \to V_h$ is the finite element interpolation operator.

We are approximating the linear variational problem

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

with the solution of

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

In Lecture 7 we saw Céa's Lemma for coercive, bounded a:

$$\begin{aligned}
\|u - u_h\|_V &\leq \frac{C}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_V \\
&\leq \frac{C}{\alpha} \|u - \mathcal{I}_h u\|_V \\
&\leq \frac{C}{\alpha} f(h, p)
\end{aligned}$$

where $\mathcal{I}_h : V \to V_h$ is the finite element interpolation operator.

We seek a bound in terms of parameters we control: h, p.

In Lecture 3, we saw that the Sobolev space $W_p^k(\Omega)$ for $p < \infty$ is equipped with the norm

$$\|u\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| \le k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p}.$$

In Lecture 3, we saw that the Sobolev space $W_p^k(\Omega)$ for $p < \infty$ is equipped with the norm

$$\|u\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p}.$$

At this point it is convenient to introduce the Sobolev *seminorm*

$$|u|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| = k} \|D^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p}.$$

In Lecture 3, we saw that the Sobolev space $W_p^k(\Omega)$ for $p < \infty$ is equipped with the norm

$$\|u\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{1/p} .$$

At this point it is convenient to introduce the Sobolev *seminorm*

$$|u|_{W_p^k(\Omega)} = \left(\sum_{|\alpha| = k} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{1/p} .$$

Example

$$\|u\|^2_{H^1(\Omega)} = \|u\|^2_{L^2(\Omega)} + |u|^2_{H^1(\Omega)}.$$

We want to bound the interpolation error in terms of the mesh size $h \to 0$ and polynomial degree $p \to \infty$. How should we characterise the size of the mesh?

We want to bound the interpolation error in terms of the mesh size $h \to 0$ and polynomial degree $p \to \infty$. How should we characterise the size of the mesh?

The quantity we will use on each cell is its *diameter*.

Definition (diameter of a cell)

$$h_K = \operatorname{diam}(K) = \max\{\|x - y\| : x, y \in K\}.$$

For a triangle or tetrahedron, this resolves to the length of its longest edge.

We want to bound the interpolation error in terms of the mesh size $h \to 0$ and polynomial degree $p \to \infty$. How should we characterise the size of the mesh?

The quantity we will use on each cell is its *diameter*.

Definition (diameter of a cell)

$$h_K = \mathrm{diam}(K) = \max\{\|x - y\| : x, y \in K\}.$$

For a triangle or tetrahedron, this resolves to the length of its longest edge.

Over the whole mesh we take a pessimistic view.

Definition (mesh size)

Given a mesh $\mathcal{M}$, its mesh size h is given by

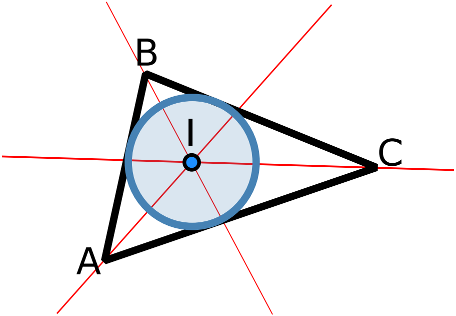
$$h = \max_{K \in \mathcal{M}} \mathrm{diam}(K).$$

We consider a sequence of meshes $(\mathcal{M}_h)_h$ indexed by the mesh size $h \to 0$. We need a technical condition on the sequence of meshes.

We consider a sequence of meshes $(\mathcal{M}_h)_h$ indexed by the mesh size $h \to 0$. We need a technical condition on the sequence of meshes.

Definition (incircle diameter of a cell)

The incircle diameter ρ_K of a cell K is the diameter of the largest closed ball (i.e. closed disc in two dimensions) that is completely contained in K.



We consider a sequence of meshes $(\mathcal{M}_h)_h$ indexed by the mesh size $h \to 0$. We need a technical condition on the sequence of meshes.

Definition (incircle diameter of a cell)

The incircle diameter ρ_K of a cell K is the diameter of the largest closed ball (i.e. closed disc in two dimensions) that is completely contained in K.

Definition (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

$$\sup_{h>0} \max_{K \in \mathcal{M}_h} \frac{h_K}{\rho_K} \leq \sigma.$$

Section 3

Interpolation error for Lagrange elements

First, consider CG_p elements applied to problems in $H^1(\Omega)$.

Theorem

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 the mesh size h. Let $u \in H^{p+1}(\Omega)$ with $p+1 > d/2$, and suppose that $\mathcal{I}_h : H^{p+1}(\Omega) \to V_h$ is the interpolation operator associated with V_h. Then there exists a constant $D < \infty$ independent of u such that

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq D h^p |u|_{H^{p+1}(\Omega)}.$$

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \le D h^p |u|_{H^{p+1}(\Omega)}.$$

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \le D h^p |u|_{H^{p+1}(\Omega)}.$$

Remark

For $p = 1$, the interpolation error depends on the second derivatives (curvature). If $|u|_{H^2(\Omega)}$ is zero, i.e. for a linear function u, the interpolant $\mathcal{I}_h u$ is exact.

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq D h^p |u|_{H^{p+1}(\Omega)}.$$

Remark

For $p = 1$, the interpolation error depends on the second derivatives (curvature). If $|u|_{H^2(\Omega)}$ is zero, i.e. for a linear function u, the interpolant $\mathcal{I}_h u$ is exact.

Remark

Error scales like h^p. If solutions are smooth, increasing p is better. If not, decreasing h is better. These can be combined in hp-adaptive schemes.

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq D h^p |u|_{H^{p+1}(\Omega)}.$$

Remark

For $p = 1$, the interpolation error depends on the second derivatives (curvature). If $|u|_{H^2(\Omega)}$ is zero, i.e. for a linear function u, the interpolant $\mathcal{I}_h u$ is exact.

Remark

Error scales like h^p. If solutions are smooth, increasing p is better. If not, decreasing h is better. These can be combined in hp-adaptive schemes.

Remark

Notice that we require $u \in H^2(\Omega)$ for CG_1 and $d = 1, 2, 3$, as before.

How do we know if $u \in H^2(\Omega)$? An elliptic regularity result.

How do we know if $u \in H^2(\Omega)$? An elliptic regularity result.

Theorem (Example elliptic regularity result)

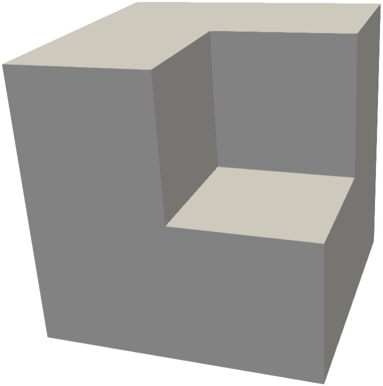
Let Ω be a convex polygon ($d = 2$), a convex polyhedron ($d = 3$), or a C^2-smooth domain (for $d \geq 2$) i.e. $\partial\Omega$ possesses a local parametrisation by C^2 functions. Suppose that $f \in L^2(\Omega)$. Then the solution $u \in H_0^1(\Omega)$ to the Poisson equation is an element of $H^2(\Omega) \cap H_0^1(\Omega)$ and satisfies

$$|u|_{H^2(\Omega)} \leq c\|f\|_{L^2(\Omega)}$$

for some constant c, independent of u and f.

The requirement that Ω has some smoothness is indispensable.

The requirement that Ω has some smoothness is indispensable.



With a re-entrant corner, $u \in H^1(\Omega) \setminus H^2(\Omega)$.

Section 4

Changing norms: Aubin–Nitsche duality

The interpolation error bound also depends on the norm used.

The interpolation error bound also depends on the norm used.

The $L^2(\Omega)$ norm is *weaker* than the $H^1(\Omega)$ norm: it only measures how good the approximation of the function is, while the $H^1(\Omega)$ norm also measures the function and its derivative.

The interpolation error bound also depends on the norm used.

The $L^2(\Omega)$ norm is *weaker* than the $H^1(\Omega)$ norm: it only measures how good the approximation of the function is, while the $H^1(\Omega)$ norm also measures the function and its derivative.

When the interpolation error is measured in a weaker norm, the convergence rate improves:

Theorem

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 the mesh size h. Let $u \in H^{p+1}(\Omega)$ with $p+1 > d/2$, and suppose that $\mathcal{I}_h : H^{p+1}(\Omega) \to V_h$ is the interpolation operator associated with each V_h. Then there exists a constant $D < \infty$ independent of u such that

$$\|u - \mathcal{I}_h u\|_{L^2(\Omega)} \leq D h^{p+1} |u|_{H^{p+1}(\Omega)}.$$

Let us consider solving the Poisson equation. We are not primarily interested in $\|u - \mathcal{I}_h u\|_{L^2(\Omega)}$. We are interested in $\|u - u_h\|_{L^2(\Omega)}$.

Let us consider solving the Poisson equation. We are not primarily interested in $\|u - \mathcal{I}_h u\|_{L^2(\Omega)}$. We are interested in $\|u - u_h\|_{L^2(\Omega)}$.

Céa's Lemma only tells us about $\|u - u_h\|_{H^1(\Omega)}$ in terms of $\|u - \mathcal{I}_h u\|_{H^1(\Omega)}$. How can we get a hold of estimates of $\|u - u_h\|_{L^2(\Omega)}$?

Let us consider solving the Poisson equation. We are not primarily interested in $\|u - \mathcal{I}_h u\|_{L^2(\Omega)}$. We are interested in $\|u - u_h\|_{L^2(\Omega)}$.

Céa's Lemma only tells us about $\|u - u_h\|_{H^1(\Omega)}$ in terms of $\|u - \mathcal{I}_h u\|_{H^1(\Omega)}$. How can we get a hold of estimates of $\|u - u_h\|_{L^2(\Omega)}$?

The Aubin–Nitsche duality argument.

Consider the variational problem

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = (f,v)_{L^2(\Omega)} \text{ for all } v \in H_0^1(\Omega),$$

where Ω satisfies one of the assumptions stated on $p.9.$ and

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

Consider the variational problem

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = (f,v)_{L^2(\Omega)} \text{ for all } v \in H_0^1(\Omega),$$

where Ω satisfies one of the assumptions stated on $p.9.$ and

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

We know that this has a unique solution $u \in H_0^1(\Omega)$ by the Lax–Milgram theorem, and that $u \in H^2(\Omega) \cap H_0^1(\Omega)$ by elliptic regularity.

Consider the variational problem

$$\text{find } u \in H_0^1(\Omega) \text{ such that } a(u,v) = (f,v)_{L^2(\Omega)} \text{ for all } v \in H_0^1(\Omega),$$

where Ω satisfies one of the assumptions stated on $p.9.$ and

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

We know that this has a unique solution $u \in H_0^1(\Omega)$ by the Lax–Milgram theorem, and that $u \in H^2(\Omega) \cap H_0^1(\Omega)$ by elliptic regularity.

We thus know that its Galerkin approximation with CG_1 elements satisfies

$$\begin{aligned}
\|u - u_h\|_{H^1(\Omega)} &\le CD\alpha^{-1} h |u|_{H^2(\Omega)} \\
&\le cCD\alpha^{-1} h \|f\|_{L^2(\Omega)}.
\end{aligned}$$

Consider the error $e = u - u_h \in H_0^1(\Omega)$. Given any element $v \in H_0^1(\Omega)$, let

$$e^*(v) := (u - u_h, v)_{L^2(\Omega)}.$$

This makes sense as the source term for an auxiliary problem ('adjoint' or 'dual'):

$$\text{find } w \in H_0^1(\Omega) \text{ such that } a(v, w) = e^*(v) \text{ for all } v \in H_0^1(\Omega).$$

Consider the error $e = u - u_h \in H_0^1(\Omega)$. Given any element $v \in H_0^1(\Omega)$, let

$$e^*(v) := (u - u_h, v)_{L^2(\Omega)}.$$

This makes sense as the source term for an auxiliary problem ('adjoint' or 'dual'):

find $w \in H_0^1(\Omega)$ such that $a(v, w) = e^*(v)$ for all $v \in H_0^1(\Omega)$.

By the Lax–Milgram theorem we know that this has a unique solution $w \in H_0^1(\Omega)$, and by elliptic regularity there exists a c such that

$$|w|_{H^2(\Omega)} \leq c\|e\|_{L^2(\Omega)}.$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$\|u - u_h\|^2_{L^2(\Omega)} = (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h)$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$\begin{aligned}\|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\ &= a(u - u_h, w)\end{aligned}$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$\begin{aligned} \|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\ &= a(u - u_h, w) \\ &= a(u - u_h, w - \mathcal{I}_h w) \qquad \text{(GO)} \end{aligned}$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$\begin{aligned}
\|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\
&= a(u - u_h, w) \\
&= a(u - u_h, w - \mathcal{I}_h w) && \text{(GO)} \\
&\leq C\|u - u_h\|_{H^1(\Omega)}\|w - \mathcal{I}_h w\|_{H^1(\Omega)} && (a(\cdot,\cdot) \text{ continuous})
\end{aligned}$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$
\begin{aligned}
\|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\
&= a(u - u_h, w) \\
&= a(u - u_h, w - \mathcal{I}_h w) && \text{(GO)} \\
&\le C\|u - u_h\|_{H^1(\Omega)}\|w - \mathcal{I}_h w\|_{H^1(\Omega)} && (a(\cdot,\cdot) \text{ continuous}) \\
&\le CDh\|u - u_h\|_{H^1(\Omega)}|w|_{H^2(\Omega)} && \text{(interp error bnd)}
\end{aligned}
$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$\begin{aligned}
\|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\
&= a(u - u_h, w) \\
&= a(u - u_h, w - \mathcal{I}_h w) && \text{(GO)} \\
&\leq C\|u - u_h\|_{H^1(\Omega)}\|w - \mathcal{I}_h w\|_{H^1(\Omega)} && (a(\cdot,\cdot)\ \text{continuous}) \\
&\leq CDh\|u - u_h\|_{H^1(\Omega)}|w|_{H^2(\Omega)} && \text{(interp error bnd)} \\
&\leq C^2 D^2 \alpha^{-1} h^2 |u|_{H^2(\Omega)}|w|_{H^2(\Omega)} && (H^1\ \text{error est})
\end{aligned}$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$\begin{aligned}
\|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\
&= a(u - u_h, w) \\
&= a(u - u_h, w - \mathcal{I}_h w) && \text{(GO)} \\
&\leq C\|u - u_h\|_{H^1(\Omega)}\|w - \mathcal{I}_h w\|_{H^1(\Omega)} && (a(\cdot,\cdot)\text{ continuous}) \\
&\leq CDh\|u - u_h\|_{H^1(\Omega)}|w|_{H^2(\Omega)} && \text{(interp error bnd)} \\
&\leq C^2D^2\alpha^{-1}h^2|u|_{H^2(\Omega)}|w|_{H^2(\Omega)} && (H^1\text{ error est)} \\
&\leq C^2D^2\alpha^{-1}ch^2|u|_{H^2(\Omega)}\|u - u_h\|_{L^2(\Omega)} && \text{(elliptic reg)}
\end{aligned}$$

Now consider $\|u - u_h\|^2_{L^2(\Omega)}$, the quantity we wish to bound. We have

$$
\begin{aligned}
\|u - u_h\|^2_{L^2(\Omega)} &= (u - u_h, u - u_h)_{L^2(\Omega)} = e^*(u - u_h) \\
&= a(u - u_h, w) \\
&= a(u - u_h, w - \mathcal{I}_h w) && \text{(GO)} \\
&\le C\|u - u_h\|_{H^1(\Omega)}\|w - \mathcal{I}_h w\|_{H^1(\Omega)} && (a(\cdot,\cdot) \text{ continuous}) \\
&\le CDh\|u - u_h\|_{H^1(\Omega)}|w|_{H^2(\Omega)} && \text{(interp error bnd)} \\
&\le C^2D^2\alpha^{-1}h^2|u|_{H^2(\Omega)}|w|_{H^2(\Omega)} && (H^1 \text{ error est}) \\
&\le C^2D^2\alpha^{-1}ch^2|u|_{H^2(\Omega)}\|u - u_h\|_{L^2(\Omega)} && \text{(elliptic reg)}
\end{aligned}
$$

and hence there exists a constant C' such that

$$\|u - u_h\|_{L^2(\Omega)} \le C'h^2|u|_{H^2(\Omega)}$$

as required.

Section 5

Interpolation error estimates for other elements

For the Argyris element we have:

$$\|u - \mathcal{I}_h u\|_{H^2(\Omega)} \leq Dh^4 |u|_{H^6(\Omega)},$$
$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leq Dh^5 |u|_{H^6(\Omega)},$$
$$\|u - \mathcal{I}_h u\|_{L^2(\Omega)} \leq Dh^6 |u|_{H^6(\Omega)}.$$

It can also be of interest to consider what interpolation does to other quantities. For example, suppose we are solving the time-dependent Maxwell's equations, which involve Gauss' law:

$$\nabla \cdot B = 0$$

for the magnetic field $B \in H(\mathrm{div}; \Omega)$.

It can also be of interest to consider what interpolation does to other quantities. For example, suppose we are solving the time-dependent Maxwell's equations, which involve Gauss' law:

$$\nabla \cdot B = 0$$

for the magnetic field $B \in H(\mathrm{div}; \Omega)$.

If we start with initial data B_0 for the magnetic field that satisfies this, we need to make sure that

$$\nabla \cdot (\mathcal{I}_h B_0) = 0$$

also.

It can also be of interest to consider what interpolation does to other quantities. For example, suppose we are solving the time-dependent Maxwell's equations, which involve Gauss' law:

$$\nabla \cdot B = 0$$

for the magnetic field $B \in H(\mathrm{div}; \Omega)$.

If we start with initial data B_0 for the magnetic field that satisfies this, we need to make sure that

$$\nabla \cdot (\mathcal{I}_h B_0) = 0$$

also.

For $H(\mathrm{div}; \Omega)$-conforming elements such as the BDM family, we have such a result:

$$\|\nabla \cdot (u - \mathcal{I}_h u)\|_{L^2(\Omega)} \leq D h^s |\nabla \cdot u|_{H^s(\Omega)}.$$

If $\nabla \cdot u = 0$, then this forces $\nabla \cdot \mathcal{I}_h u = 0$ also.

C6.4 Finite Element Methods for PDEs
Lecture 12: Nonlinear problems

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

Section 1

Variational formulation of nonlinear problems

Our basic abstraction for linear problems was:

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

Our basic abstraction for linear problems was:

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

Our abstraction for nonlinear problems will be:

$$\text{find } u \in V \text{ such that } G(u; v) = 0 \text{ for all } v \in V,$$

where $G : V \times V \to \mathbb{R}$.

Our basic abstraction for linear problems was:

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

Our abstraction for nonlinear problems will be:

$$\text{find } u \in V \text{ such that } G(u; v) = 0 \text{ for all } v \in V,$$

where $G : V \times V \to \mathbb{R}$.

We use $G(u; v)$ to record that G is nonlinear in u but always linear in v.

It is often useful to reformulate the variational statement as an equation, as we did in the linear case.

It is often useful to reformulate the variational statement as an equation, as we did in the linear case.

Define $H : V \to V^*$ via

$$(H(u))\,(v) = G(u; v).$$

It is often useful to reformulate the variational statement as an equation, as we did in the linear case.

Define $H : V \to V^*$ via

$$(H(u))\,(v) = G(u; v).$$

Solutions of our nonlinear variational problem are exactly roots of H, i.e. we seek $u \in V$ such that

$$H(u) = 0.$$

Algorithms in numerical analysis break down a problem into a sequence of simpler ones. For example, for a time-dependent ODE, we have

$$\begin{aligned}
\text{transient DE} \quad &\xrightarrow{\text{implicit scheme}} \quad \text{system of nonlinear equations}\\
&\xrightarrow{\text{Newton's method}} \quad \text{system of linear equations}\\
&\xrightarrow{\text{Linear solver}} \quad \text{numerical solution}
\end{aligned}$$

Algorithms in numerical analysis break down a problem into a sequence of simpler ones. For example, for a time-dependent ODE, we have

$$\begin{aligned} \text{transient DE} \quad &\xrightarrow{\text{implicit scheme}} \quad \text{system of nonlinear equations} \\ &\xrightarrow{\text{Newton's method}} \quad \text{system of linear equations} \\ &\xrightarrow{\text{Linear solver}} \quad \text{numerical solution} \end{aligned}$$

Now consider solving a time-dependent PDE. We must *also* discretise in space. We could spatially discretise at any level:

Algorithms in numerical analysis break down a problem into a sequence of simpler ones. For example, for a time-dependent ODE, we have

$$
\begin{array}{lcl}
\text{transient DE} & \xrightarrow{\text{implicit scheme}} & \text{system of nonlinear equations} \\
 & \xrightarrow{\text{Newton's method}} & \text{system of linear equations} \\
 & \xrightarrow{\text{Linear solver}} & \text{numerical solution}
\end{array}
$$

Now consider solving a time-dependent PDE. We must *also* discretise in space. We could spatially discretise at any level:

- the time-dependent PDE, yielding a very large ODE system;

Algorithms in numerical analysis break down a problem into a sequence of simpler ones. For example, for a time-dependent ODE, we have

$$\text{transient DE} \xrightarrow{\text{implicit scheme}} \text{system of nonlinear equations}$$
$$\xrightarrow{\text{Newton's method}} \text{system of linear equations}$$
$$\xrightarrow{\text{Linear solver}} \text{numerical solution}$$

Now consider solving a time-dependent PDE. We must *also* discretise in space. We could spatially discretise at any level:

- the time-dependent PDE, yielding a very large ODE system;
- the nonlinear PDEs arising from time discretisation;

Algorithms in numerical analysis break down a problem into a sequence of simpler ones. For example, for a time-dependent ODE, we have

$$\begin{aligned} \text{transient DE} \quad & \xrightarrow{\text{implicit scheme}} \quad \text{system of nonlinear equations} \\ & \xrightarrow{\text{Newton's method}} \quad \text{system of linear equations} \\ & \xrightarrow{\text{Linear solver}} \quad \text{numerical solution} \end{aligned}$$

Now consider solving a time-dependent PDE. We must *also* discretise in space. We could spatially discretise at any level:

- the time-dependent PDE, yielding a very large ODE system;
- the nonlinear PDEs arising from time discretisation;
- the linear PDEs arising from the nonlinear solver.

Algorithms in numerical analysis break down a problem into a sequence of simpler ones. For example, for a time-dependent ODE, we have

$$\begin{aligned} \text{transient DE} \quad &\xrightarrow{\text{implicit scheme}} \quad \text{system of nonlinear equations} \\ &\xrightarrow{\text{Newton's method}} \quad \text{system of linear equations} \\ &\xrightarrow{\text{Linear solver}} \quad \text{numerical solution} \end{aligned}$$

Now consider solving a time-dependent PDE. We must *also* discretise in space. We could spatially discretise at any level:

- the time-dependent PDE, yielding a very large ODE system;
- the nonlinear PDEs arising from time discretisation;
- the linear PDEs arising from the nonlinear solver.

These operations sometimes commute, and sometimes do not.

For a nonlinear PDE, we must

- devise a scheme to solve the nonlinear problem;
- discretise in space to make the problem finite-dimensional.

For a nonlinear PDE, we must

- devise a scheme to solve the nonlinear problem;
- discretise in space to make the problem finite-dimensional.

We can do these in either order.

Section 2

Discretise, then solve

We consider the Galerkin approximation:

$$\text{find } u_h \in V_h \text{ such that } G(u_h; v_h) = 0 \text{ for all } v_h \in V_h.$$

We consider the Galerkin approximation:

$$\text{find } u_h \in V_h \text{ such that } G(u_h; v_h) = 0 \text{ for all } v_h \in V_h.$$

A primary goal of an analysis is bounding the error $\|u - u_h\|_V$.

We consider the Galerkin approximation:

$$\text{find } u_h \in V_h \text{ such that } G(u_h; v_h) = 0 \text{ for all } v_h \in V_h.$$

A primary goal of an analysis is bounding the error $\|u - u_h\|_V$.

But *which* u, and which u_h? How do we pair the continuous and discrete solutions across mesh levels? How do we know that the discrete problem supports the right number of solutions? How do we know there are no spurious solutions?

We consider the Galerkin approximation:

$$\text{find } u_h \in V_h \text{ such that } G(u_h; v_h) = 0 \text{ for all } v_h \in V_h.$$

A primary goal of an analysis is bounding the error $\|u - u_h\|_V$.

But *which* u, and which u_h? How do we pair the continuous and discrete solutions across mesh levels? How do we know that the discrete problem supports the right number of solutions? How do we know there are no spurious solutions?

These are difficult questions; possible to address, but we will sidestep them.

Section 3

Solve, then discretise

We will, later on, develop an algorithm for dealing with the nonlinearity *on the infinite-dimensional level*, the *Newton–Kantorovich* iteration.

We will, later on, develop an algorithm for dealing with the nonlinearity *on the infinite-dimensional level*, the *Newton–Kantorovich* iteration.

This will achieve the solution of the nonlinear problem by solving a sequence of linear problems, each of which we then discretise with the finite element method.

We will, later on, develop an algorithm for dealing with the nonlinearity *on the infinite-dimensional level*, the *Newton–Kantorovich* iteration.

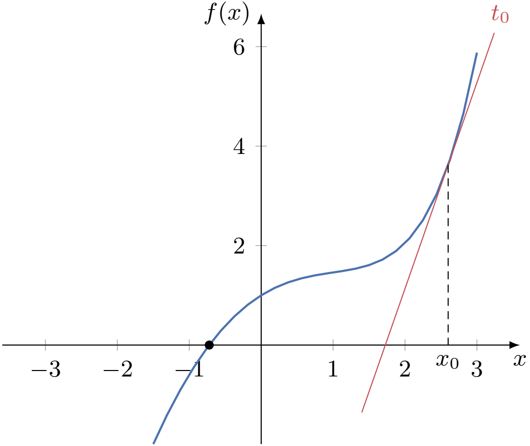
This will achieve the solution of the nonlinear problem by solving a sequence of linear problems, each of which we then discretise with the finite element method.

First, however, let us recall Newton's method in $\mathbb{R}$ and $\mathbb{R}^N$.

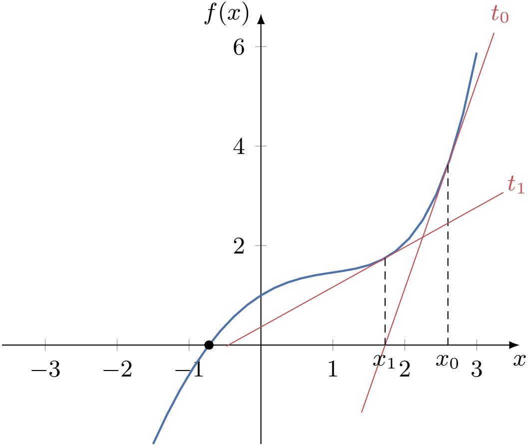
Section 4

Newton's method in $\mathbb{R}^N$

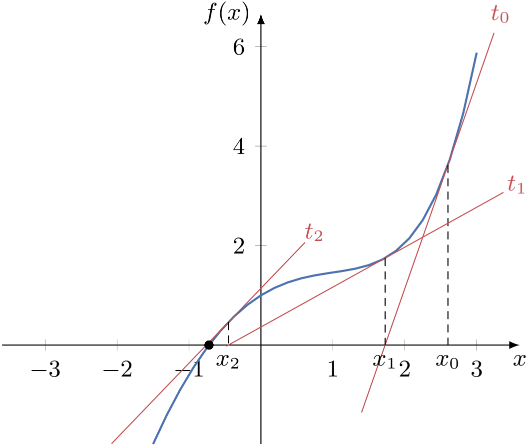
Consider $f(x) = e^{9x/10} - x^2$ with $x_0 = 2.6$.



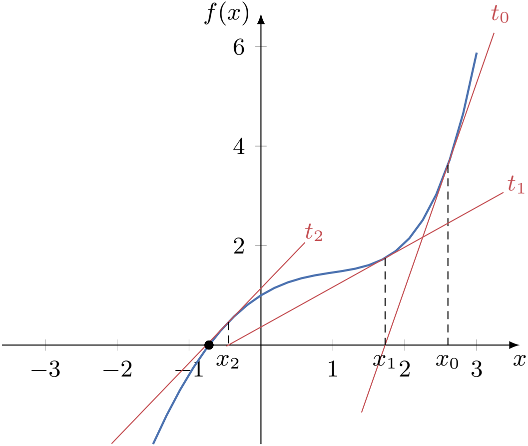
Consider $f(x) = e^{9x/10} - x^2$ with $x_0 = 2.6$.



Consider $f(x) = e^{9x/10} - x^2$ with $x_0 = 2.6$.



Consider $f(x) = e^{9x/10} - x^2$ with $x_0 = 2.6$.



solve $f'(x_k)\delta x_k = -f(x_k)$; update $x_{k+1} = x_k + \delta x_k$.

Termination

The algorithm terminates if $f(x_k) = 0$, as desired.

Termination

The algorithm terminates if $f(x_k) = 0$, as desired.

Invertibility

We require $f'(x_k)$ to be invertible at every iteration.

Termination

The algorithm terminates if $f(x_k) = 0$, as desired.

Invertibility

We require $f'(x_k)$ to be invertible at every iteration.

Poor global convergence

The initial guess matters. With poor initial guesses, Newton's method may diverge to infinity, or get stuck in a cycle.

Termination

The algorithm terminates if $f(x_k) = 0$, as desired.

Invertibility

We require $f'(x_k)$ to be invertible at every iteration.

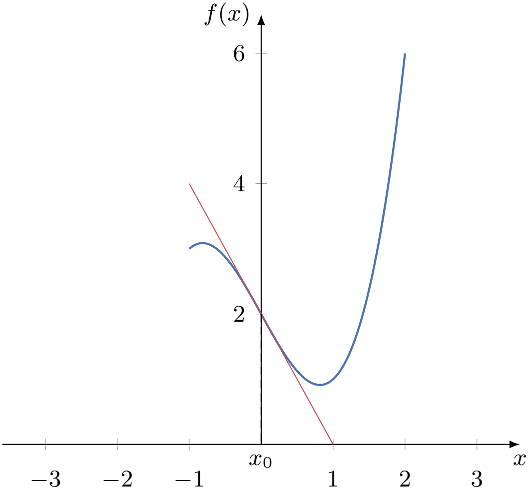
Poor global convergence

The initial guess matters. With poor initial guesses, Newton's method may diverge to infinity, or get stuck in a cycle.

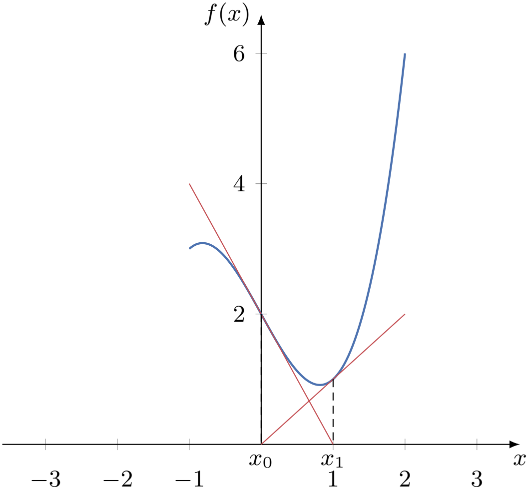
Good local convergence

If f is smooth, the solution is isolated, and the guess close, Newton converges quadratically.

Consider $f(x) = x^3 - 2x + 2$ with $x_0 = 0$.



Consider $f(x) = x^3 - 2x + 2$ with $x_0 = 0$.



This geometric reasoning is hard to generalise to higher dimensions. Let us look at a derivation that *does* extend.

This geometric reasoning is hard to generalise to higher dimensions. Let us look at a derivation that *does* extend.

Consider the Taylor expansion of f around x_k:

$$f(x_k + \delta x) = f(x_k) + f'(x_k)\delta x + \mathcal{O}(\delta x^2).$$

This geometric reasoning is hard to generalise to higher dimensions. Let us look at a derivation that *does* extend.

Consider the Taylor expansion of f around x_k:

$$f(x_k + \delta x) = f(x_k) + f'(x_k)\delta x + \mathcal{O}(\delta x^2).$$

Linearise the model by ignoring higher-order terms:

$$f(x_k + \delta x) \approx f(x_k) + f'(x_k)\delta x$$

and find δx (denoted by δx_k) such that

$$0 = f(x_k) + f'(x_k)\delta x_k;$$

that is, for $x_0 \in \mathbb{R}$ given, and $k = 0, 1, \ldots$, find $\delta x_k \in \mathbb{R}$ such that

$$f'(x_k)\delta x_k = -f(x_k); \qquad \text{update } x_{k+1} := x_k + \delta x_k.$$

This does extend to an $F \in C^1(\mathbb{R}^N; \mathbb{R}^N)$.

Newton's method is: for $x_0 \in \mathbb{R}^N$ given, and $k = 0, 1, \ldots$, find $\delta x_k \in \mathbb{R}^N$ such that

$$DF(x_k)\delta x_k = -F(x_k); \qquad \text{update } x_{k+1} := x_k + \delta x_k,$$

where $DF(x_k) \in \mathbb{R}^{N \times N}$ is the Jacobian matrix (Fréchet derivative) of F evaluated at $x_k \in \mathbb{R}^N$.

Section 5

Newton's method in Banach spaces

The generalisation of Newton's method to Banach spaces is called the *Newton–Kantorovich* algorithm. It *does not assume the existence of a solution* to $H(u) = 0$: given certain conditions on H and initial guess, it *proves* the existence of a solution u to $H(u) = 0$.

Theorem (Kantorovich (1948))

Let X and Y be two Banach spaces. Let Ω be an open subset of X, let $H \in C^1(\Omega, Y)$, and let $x_0 \in \Omega$ be an initial guess such that the Fréchet derivative $H'(x_0)$ is boundedly 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 $\tilde{x}, x \in B(x_0, r)$,

$$\|H'(x_0)^{-1} \left(H'(\tilde{x}) - H'(x)\right) \|_{L(X;X)} \leq \frac{1}{r} \|\tilde{x} - x\|_X .$$

Theorem (Kantorovich (1948))

Then

(1) $H'(x) \in L(X;Y)$ *is invertible at each* $x \in B(x_0, r)$.

(2) The Newton sequence $(x_k)_{k=0}^{\infty}$ *defined by*

$$x_{k+1} = x_k - H'(x_k)^{-1} H(x_k)$$

is such that $x_k \in B(x_0, r)$ *for all* $k \geq 0$ *and converges to a root* $x^\star \in \overline{B(x_0, r)}$ *of* H.

(3) For each $k \geq 0$,

$$\|x^\star - x_k\|_X \leq \frac{r}{2^k}.$$

(4) The root $x^\star$ *is locally unique, i.e.* $x^\star$ *is the only root of* H *in the ball* $\overline{B(x_0, r)}$.

The Bratu–Gelfand equation

$$u''(x) + \lambda e^u = 0, \quad u(0) = 0 = u(1)$$

has variational formulation: find $u \in H_0^1(0,1)$ such that

$$(H(u))(v) = G(u; v) := -\int_0^1 u'(x)v'(x)\,\mathrm{d}x + \int_0^1 \lambda e^u v\,\mathrm{d}x = 0$$

for all $v \in H_0^1(0,1)$. Note that we then have that

$$(H'(u))(w)(v) = G_u(u; w, v) = -\int_0^1 w'(x)v'(x)\,\mathrm{d}x + \int_0^1 \lambda e^u wv\,\mathrm{d}x.$$

Unwinding the statement of the Newton step, the update δu_k solves

$$G_u(u_k; \delta u_k, v) = -G(u_k; v) \qquad \text{for all } v \in V.$$

So the Newton equation becomes: with $u_0 \in H_0^1(0,1)$ given, for $k = 0, 1, \ldots$, find $\delta u_k \in H_0^1(0,1)$ such that

$$-\int_0^1 \delta u_k'(x) v'(x) \, \mathrm{d}x + \int_0^1 \lambda e^{u_k} \delta u_k v \, \mathrm{d}x = \int_0^1 u_k'(x) v'(x) \, \mathrm{d}x - \int_0^1 \lambda e^{u_k} v \, \mathrm{d}x$$

for all $v \in H_0^1(0,1)$; then update $u_{k+1} := u_k + \delta u_k$.

Unwinding the statement of the Newton step, the update δu_k solves

$$G_u(u_k; \delta u_k, v) = -G(u_k; v) \qquad \text{for all } v \in V.$$

So the Newton equation becomes: with $u_0 \in H_0^1(0,1)$ given, for $k = 0, 1, \ldots$, find $\delta u_k \in H_0^1(0,1)$ such that

$$-\int_0^1 \delta u_k'(x) v'(x) \, \mathrm{d}x + \int_0^1 \lambda e^{u_k} \delta u_k v \, \mathrm{d}x = \int_0^1 u_k'(x) v'(x) \, \mathrm{d}x - \int_0^1 \lambda e^{u_k} v \, \mathrm{d}x$$

for all $v \in H_0^1(0,1)$; then update $u_{k+1} := u_k + \delta u_k$.

This is, for $u_k \in H_0^1(0,1)$ given, a linear problem for $\delta u_k \in H_0^1(0,1)$, which we discretise with the finite element method.

Many questions remain.

Are the resulting linear variational problems for δu_k well-posed? In general they are not coercive. We need a more general theory of well-posedness.

Many questions remain.

Are the resulting linear variational problems for δu_k well-posed? In general they are not coercive. We need a more general theory of well-posedness.

We will never compute the exact δu_k; we can only compute a finite-dimensional approximation. One therefore develops *inexact* Newton–Kantorovich methods.

Many questions remain.

Are the resulting linear variational problems for δu_k well-posed? In general they are not coercive. We need a more general theory of well-posedness.

We will never compute the exact δu_k; we can only compute a finite-dimensional approximation. One therefore develops *inexact* Newton–Kantorovich methods.

How can we control the approximation error of δu_k?

Many questions remain.

Are the resulting linear variational problems for δu_k well-posed? In general they are not coercive. We need a more general theory of well-posedness.

We will never compute the exact δu_k; we can only compute a finite-dimensional approximation. One therefore develops *inexact* Newton–Kantorovich methods.

How can we control the approximation error of δu_k?

How can the algorithm be globalised?

C6.4 Finite Element Methods for PDEs
Lecture 13: Noncoercive problems

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

So far we have treated *coercive* problems. This means that the bilinear form $a(u, v)$ in the linear variational problem we are trying to solve

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

satisfies

So far we have treated *coercive* problems. This means that the bilinear form $a(u, v)$ in the linear variational problem we are trying to solve

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

satisfies

$$a(u, u) \geq \alpha \|u\|_V^2$$

for some $\alpha > 0$.

So far we have treated *coercive* problems. This means that the bilinear form $a(u, v)$ in the linear variational problem we are trying to solve

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

satisfies

$$a(u, u) \geq \alpha \|u\|_V^2$$

for some $\alpha > 0$.

Recall that the best constant α satisfying the definition is given by

$$\alpha := \inf_{\substack{u \in V \\ u \neq 0}} \frac{a(u, u)}{\|u\|_V^2}.$$

We now consider *noncoercive* problems, one for which no such $\alpha > 0$ exists. We will develop more general criteria for well-posedness of the linear variational problem, the so-called inf-sup or Babuška conditions.

We now consider *noncoercive* problems, one for which no such $\alpha > 0$ exists. We will develop more general criteria for well-posedness of the linear variational problem, the so-called inf-sup or Babuška conditions.

For coercive problems, well-posedness is inherited for any $V_h \subset V$.

We now consider *noncoercive* problems, one for which no such $\alpha > 0$ exists. We will develop more general criteria for well-posedness of the linear variational problem, the so-called inf-sup or Babuška conditions.

For coercive problems, well-posedness is inherited for any $V_h \subset V$.

This is not true for noncoercive problems. Well-posedness is **not inherited for arbitrary** $V_h \subset V$.

Theorem (Babuška's theorem)

Let V_1 and V_2 be two Hilbert spaces. 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

Theorem (Babuška's theorem)

Let V_1 and V_2 be two Hilbert spaces. 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* $u \in V_1, v \in V_2$;

Theorem (Babuška's theorem)

Let V_1 *and* V_2 *be two Hilbert spaces. 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) $\gamma \leq \inf\limits_{\substack{u\in V_1\\ u\neq 0}} \sup\limits_{\substack{v\in V_2\\ v\neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$*;*

Theorem (Babuška's theorem)

Let V_1 and V_2 be two Hilbert spaces. 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) $\gamma \leq \inf\limits_{\substack{u\in V_1\\u\neq 0}} \sup\limits_{\substack{v\in V_2\\v\neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$*;*

(3) $\gamma' \leq \inf\limits_{\substack{v\in V_2\\v\neq 0}} \sup\limits_{\substack{u\in V_1\\u\neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$.

Theorem (Babuška's theorem)

Let V_1 and V_2 be two Hilbert spaces. 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) $|a(u,v)| \le C\|u\|_{V_1}\|v\|_{V_2}$ *for all* $u \in V_1, v \in V_2$*;*

(2) $\gamma \le \inf\limits_{\substack{u\in V_1\\ u\neq 0}} \sup\limits_{\substack{v\in V_2\\ v\neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$*;*

(3) $\gamma' \le \inf\limits_{\substack{v\in V_2\\ v\neq 0}} \sup\limits_{\substack{u\in V_1\\ u\neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$.

Then for all $F \in V_2^*$ there exists exactly one element $u \in V_1$ such that

$$a(u,v) = F(v) \text{ for all } v \in V_2.$$

Theorem (Babuška's theorem)

Let V_1 and V_2 be two Hilbert spaces. 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) $|a(u,v)| \le C\|u\|_{V_1}\|v\|_{V_2}$ *for all* $u \in V_1, v \in V_2$*;*

(2) $\gamma \le \inf\limits_{\substack{u\in V_1\\ u\neq 0}} \sup\limits_{\substack{v\in V_2\\ v\neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$*;*

(3) $\gamma' \le \inf\limits_{\substack{v\in V_2\\ v\neq 0}} \sup\limits_{\substack{u\in V_1\\ u\neq 0}} \frac{a(u,v)}{\|u\|_{V_1}\|v\|_{V_2}}$*.*

Then for all $F \in V_2^*$ there exists exactly one element $u \in V_1$ such that

$$a(u,v) = F(v) \text{ for all } v \in V_2.$$

Furthermore the problem is stable in that

$$\|u\|_{V_1} \le \frac{\|F\|_{V_2^*}}{\gamma}.$$

Example. A coercive bilinear form satisfies the inf-sup conditions.

Example. A coercive bilinear form satisfies the inf-sup conditions. Indeed, suppose $a(\cdot,\cdot)$ satisfies

$$(\exists\, \alpha > 0)\ (\forall\, u \in V) \qquad \alpha \|u\|_V^2 \le a(u,u).$$

Dividing both sides of the inequality by $\|u\|_V$ for $u \neq 0$, we have

$$\alpha \|u\|_V \le \frac{a(u,u)}{\|u\|_V}$$

Example. A coercive bilinear form satisfies the inf-sup conditions. Indeed, suppose $a(\cdot,\cdot)$ satisfies

$$(\exists\, \alpha > 0)\ (\forall\, u \in V) \qquad \alpha \|u\|_V^2 \leq a(u,u).$$

Dividing both sides of the inequality by $\|u\|_V$ for $u \neq 0$, we have

$$\begin{aligned} \alpha \|u\|_V &\leq \frac{a(u,u)}{\|u\|_V} \\ &\leq \sup_{\substack{v \in V \\ v \neq 0}} \frac{a(u,v)}{\|v\|_V}. \end{aligned}$$

Example. A coercive bilinear form satisfies the inf-sup conditions. Indeed, suppose $a(\cdot,\cdot)$ satisfies

$$(\exists\,\alpha > 0)\;(\forall\,u \in V) \qquad \alpha\|u\|_V^2 \le a(u,u).$$

Dividing both sides of the inequality by $\|u\|_V$ for $u \neq 0$, we have

$$\begin{aligned}\alpha\|u\|_V &\le \frac{a(u,u)}{\|u\|_V}\\ &\le \sup_{\substack{v\in V\\ v\neq 0}} \frac{a(u,v)}{\|v\|_V}.\end{aligned}$$

Infimising over $u \neq 0$, we have

$$0 < \alpha \le \inf_{\substack{u\in V\\ u\neq 0}} \sup_{\substack{v\in V\\ v\neq 0}} \frac{a(u,v)}{\|u\|_V\|v\|_V}.$$

So the coercivity constant α works for γ (and γ').

Section 2

The proof of Babuška's theorem

Remember that we can rewrite

$$\text{find } u \in V_1 \text{ such that } a(u,v) = F(v) \text{ for all } v \in V_2$$

as

$$\text{find } u \in V_1 \text{ such that } Au = F,$$

where

$$A : V_1 \to V_2^*, \quad (Au)(v) := a(u,v) \quad \text{for } u \in V_1 \text{ and } v \in V_2$$

and $F \in V_2^*$. Let us further define $A' : V_2 \to V_1^*$ (the *transpose* of A) by

$$A' : V_2 \to V_1^*, \quad (A'v)(u) = a(u,v) \quad \text{for } v \in V_2 \text{ and } u \in V_1.$$

Remember that we can rewrite

$$\text{find } u \in V_1 \text{ such that } a(u,v) = F(v) \text{ for all } v \in V_2$$

as

$$\text{find } u \in V_1 \text{ such that } Au = F,$$

where

$$A : V_1 \to V_2^*, \quad (Au)(v) := a(u,v) \quad \text{for } u \in V_1 \text{ and } v \in V_2$$

and $F \in V_2^*$. Let us further define $A' : V_2 \to V_1^*$ (the *transpose* of A) by

$$A' : V_2 \to V_1^*, \quad (A'v)(u) = a(u,v) \quad \text{for } v \in V_2 \text{ and } u \in V_1.$$

The inf-sup conditions are just a variational way of stating facts about A.

Remember that we can rewrite

$$\text{find } u \in V_1 \text{ such that } a(u,v) = F(v) \text{ for all } v \in V_2$$

as

$$\text{find } u \in V_1 \text{ such that } Au = F,$$

where

$$A : V_1 \to V_2^*, \quad (Au)(v) := a(u,v) \quad \text{for } u \in V_1 \text{ and } v \in V_2$$

and $F \in V_2^*$. Let us further define $A' : V_2 \to V_1^*$ (the *transpose* of A) by

$$A' : V_2 \to V_1^*, \quad (A'v)(u) = a(u,v) \quad \text{for } v \in V_2 \text{ and } u \in V_1.$$

The inf-sup conditions are just a variational way of stating facts about A.

The key to understanding the inf-sup conditions is to recall the norm on the dual of a Hilbert space:

$$\|j\|_{V^*} = \sup_{\|u\|_V = 1} |j(u)| = \sup_{\substack{u \in V \\ u \neq 0}} \frac{|j(u)|}{\|u\|_V}.$$

Condition (1) in Babuška's theorem implies that

$$\|Au\|_{V_2^*} \leq C\|u\|_{V_1} \quad \text{and} \quad \|A'v\|_{V_1^*} \leq C\|v\|_{V_2}.$$

In other words $A \in L(V_1, V_2^*)$ and $A' \in L(V_2, V_1^*)$. [Recall that $L(X, Y)$ denotes the set of all bounded linear operators from X to Y.]

Condition (1) in Babuška's theorem implies that

$$\|Au\|_{V_2^*} \le C\|u\|_{V_1} \quad \text{and} \quad \|A'v\|_{V_1^*} \le C\|v\|_{V_2}.$$

In other words $A \in L(V_1, V_2^*)$ and $A' \in L(V_2, V_1^*)$. [Recall that $L(X, Y)$ denotes the set of all bounded linear operators from X to Y.]

Condition (2) in Babuška's theorem is equivalent to requiring the existence of a $\gamma > 0$ such that

$$\gamma\|u\|_{V_1} \le \|Au\|_{V_2^*} \quad \text{for all } u \in V_1.$$

Condition (1) in Babuška's theorem implies that

$$\|Au\|_{V_2^*} \leq C\|u\|_{V_1} \quad \text{and} \quad \|A'v\|_{V_1^*} \leq C\|v\|_{V_2}.$$

In other words $A \in L(V_1, V_2^*)$ and $A' \in L(V_2, V_1^*)$. [Recall that $L(X, Y)$ denotes the set of all bounded linear operators from X to Y.]

Condition (2) in Babuška's theorem is equivalent to requiring the existence of a $\gamma > 0$ such that

$$\gamma\|u\|_{V_1} \leq \|Au\|_{V_2^*} \quad \text{for all } u \in V_1.$$

Condition (3) in Babuška's theorem is equivalent to requiring the existence of a $\gamma' > 0$ such that

$$\gamma'\|v\|_{V_2} \leq \|A'v\|_{V_1^*} \quad \text{for all } v \in V_2.$$

Thus $A \in L(V_1, V_2^*)$ and $A' \in L(V_2, V_1^*)$ are injective, because $Au = 0$ implies that $u = 0$ (by Condition (2)) and $A'v = 0$ implies that $v = 0$ (by Condition (3).

Further properties of A

(i) A is a closed linear operator: for each sequence $(u_n)_{n\geq 1} \subset V_1$, if $u_n \to u$ in V_1 and $Au_n \to w$ in V_2^* then $w = Au$.

Proof: Suppose that $(u_n)_{n\geq 1}$ converges to u in V_1. Thanks to Condition (1),

$$\|Au_n - Au\|_{V_2^*} = \|A(u_n - u)\|_{V_2^*} \leq C\|u_n - u\|_{V_1},$$

and therefore $Au_n \to Au = w$ in V_2^* as $n \to \infty$.

(ii) The range $R(A)$ of A is a closed set: if $(Au_n)_{n\geq 1} \subset R(A)$ and $Au_n \to w$ in V_2^* as $n \to \infty$, then $w = Au \in R(A)$.

Proof: As $(Au_n)_{n\geq 1}$ is convergent in V_2^* it is Cauchy in V_2^*. Thanks to Condition (2),

$$\gamma \|u_m - u_n\|_{V_1} \leq \|A(u_m - u_n)\|_{V_2^*} = \|Au_m - Au_n\|_{V_2^*},$$

and therefore $(u_n)_{n\geq 1}$ is a Cauchy sequence in V_1. However, as V_1 is a Hilbert space (and therefore complete) the sequence $(u_n)_{n\geq 1}$ is convergent in V_1. Let $\lim_{n\to\infty} u_n = u \in V_1$. Then, by Condition (1),

$$\|Au_n - Au\|_{V_2^*} = \|A(u_n - u)\|_{V_2^*} \leq C\|u_n - u\|_{V_1}.$$

Hence $Au_n \to Au = w \in R(A)$.

Theorem (Banach's Closed Range Theorem)

Let X and Y be Banach spaces, let $T : D(T) \to Y$ be a closed linear operator whose domain $D(T)$ is dense in X, and let $T' : Y^* \to X^*$ be the transpose of T. Let $N(T)$ and $N(T')$ denote the null spaces of T and T', respectively. Then, the following conditions are equivalent:

(a) $R(T)$, *the range of* T, *is closed in* Y.

(b) $R(T')$, *the range of* T', *is closed in* X^* *the dual of* X.

(c) $R(T) = N(T')^\perp = \{y \in Y : \langle x^*, y\rangle = 0 \text{ for all } x^* \in N(T')\}$.

(d) $R(T') = N(T)^\perp = \{x^* \in X^* : \langle x^*, y\rangle = 0 \text{ for all } y \in N(T)\}$.

Now, take $T = A$, $X = V_1$, $Y = V_2^*$, $D(T) = V_1$, $T' = A'$ in this theorem.

Theorem (Banach's Closed Range Theorem)

Let X and Y be Banach spaces, let $T : D(T) \to Y$ be a closed linear operator whose domain $D(T)$ is dense in X, and let $T' : Y^* \to X^*$ be the transpose of T. Let $N(T)$ and $N(T')$ denote the null spaces of T and T', respectively. Then, the following conditions are equivalent:

(a) $R(T)$, *the range of* T, *is closed in* Y.

(b) $R(T')$, *the range of* T', *is closed in* X^* *the dual of* X.

(c) $R(T) = N(T')^{\perp} = \{y \in Y : \langle x^*, y\rangle = 0 \text{ for all } x^* \in N(T')\}$.

(d) $R(T') = N(T)^{\perp} = \{x^* \in X^* : \langle x^*, y\rangle = 0 \text{ for all } y \in N(T)\}$.

Now, take $T = A$, $X = V_1$, $Y = V_2^*$, $D(T) = V_1$, $T' = A'$ in this theorem. As $N(A') = \{0\}$, "(a) $\Rightarrow$ (c)" gives that $R(A) = N(A')^{\perp} = \{0\}^{\perp} = V_2^*$.

Theorem (Banach's Closed Range Theorem)

Let X and Y be Banach spaces, let $T : D(T) \to Y$ be a closed linear operator whose domain $D(T)$ is dense in X, and let $T' : Y^* \to X^*$ be the transpose of T. Let $N(T)$ and $N(T')$ denote the null spaces of T and T', respectively. Then, the following conditions are equivalent:

(a) $R(T)$*, the range of* T*, is closed in* Y*.*

(b) $R(T')$*, the range of* T'*, is closed in* X^* *the dual of* X*.*

(c) $R(T) = N(T')^{\perp} = \{y \in Y : \langle x^*, y\rangle = 0 \text{ for all } x^* \in N(T')\}$*.*

(d) $R(T') = N(T)^{\perp} = \{x^* \in X^* : \langle x^*, y\rangle = 0 \text{ for all } y \in N(T)\}$*.*

Now, take $T = A$, $X = V_1$, $Y = V_2^*$, $D(T) = V_1$, $T' = A'$ in this theorem. As $N(A') = \{0\}$, "(a) $\Rightarrow$ (c)" gives that $R(A) = N(A')^{\perp} = \{0\}^{\perp} = V_2^*$. Thus we have shown that $A : V_1 \to V_2^*$ is both injective and surjective; hence it is bijective. Therefore for each $F \in V_2^*$ the equation $Au = F$ has a unique solution $u \in V_1$. Finally, $\gamma\|u\|_{V_1} \leq \|Au\|_{V_2^*} = \|F\|_{V_2^*}$. $\square$

Section 3

Discretisation and quasioptimality

Start with

$$\text{find } u \in V_1 \text{ such that } a(u, v) = F(v) \text{ for all } v \in V_2,$$

and take the Galerkin approximation over a finite-dimensional (and thus closed) subspace $V_h \subset V_1$ with test functions from a finite-dimensional (and thus closed) subspace $W_h \subset V_2$ of the same dimension as V_h:

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

Start with

$$\text{find } u \in V_1 \text{ such that } a(u, v) = F(v) \text{ for all } v \in V_2,$$

and take the Galerkin approximation over a finite-dimensional (and thus closed) subspace $V_h \subset V_1$ with test functions from a finite-dimensional (and thus closed) subspace $W_h \subset V_2$ of the same dimension as V_h:

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

Note that Galerkin orthogonality still holds because $W_h \subset V_2$.

Start with

$$\text{find } u \in V_1 \text{ such that } a(u,v) = F(v) \text{ for all } v \in V_2,$$

and take the Galerkin approximation over a finite-dimensional (and thus closed) subspace $V_h \subset V_1$ with test functions from a finite-dimensional (and thus closed) subspace $W_h \subset V_2$ of the same dimension as V_h:

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

Note that Galerkin orthogonality still holds because $W_h \subset V_2$.

Is the discrete problem well-posed?

Let us check the Babuška conditions.

Satisfaction of (1) is inherited. What about (2) and (3)?

$$\inf_{\substack{u_h \in V_h \\ u_h \neq 0}} \sup_{\substack{v_h \in W_h \\ v_h \neq 0}} \frac{a(u_h, v_h)}{\|u_h\|_{V_1} \|v_h\|_{V_2}} \geq \tilde{\gamma} > 0, \qquad (2)_h$$

$$\inf_{\substack{v_h \in W_h \\ v_h \neq 0}} \sup_{\substack{u_h \in V_h \\ u_h \neq 0}} \frac{a(u_h, v_h)}{\|u_h\|_{V_1} \|v_h\|_{V_2}} \geq \tilde{\gamma}' > 0, \qquad (3)_h$$

with $\tilde{\gamma}$ and $\tilde{\gamma}'$ independent of the mesh size h?

Let us check the Babuška conditions.

Satisfaction of (1) is inherited. What about (2) and (3)?

$$\inf_{\substack{u_h \in V_h \\ u_h \neq 0}} \sup_{\substack{v_h \in W_h \\ v_h \neq 0}} \frac{a(u_h, v_h)}{\|u_h\|_{V_1} \|v_h\|_{V_2}} \geq \tilde{\gamma} > 0, \qquad (2)_h$$

$$\inf_{\substack{v_h \in W_h \\ v_h \neq 0}} \sup_{\substack{u_h \in V_h \\ u_h \neq 0}} \frac{a(u_h, v_h)}{\|u_h\|_{V_1} \|v_h\|_{V_2}} \geq \tilde{\gamma}' > 0, \qquad (3)_h$$

with $\tilde{\gamma}$ and $\tilde{\gamma}'$ independent of the mesh size h?

No! (2) does not imply $(2)_h$ because W_h is a strict subset of V_2;
(3) does not imply $(3)_h$ because V_h is a strict subset of V_1.

Let us check the Babuška conditions.

Satisfaction of (1) is inherited. What about (2) and (3)?

$$\inf_{\substack{u_h \in V_h \\ u_h \neq 0}} \sup_{\substack{v_h \in W_h \\ v_h \neq 0}} \frac{a(u_h, v_h)}{\|u_h\|_{V_1} \|v_h\|_{V_2}} \geq \tilde{\gamma} > 0, \qquad (2)_h$$

$$\inf_{\substack{v_h \in W_h \\ v_h \neq 0}} \sup_{\substack{u_h \in V_h \\ u_h \neq 0}} \frac{a(u_h, v_h)}{\|u_h\|_{V_1} \|v_h\|_{V_2}} \geq \tilde{\gamma}' > 0, \qquad (3)_h$$

with $\tilde{\gamma}$ and $\tilde{\gamma}'$ independent of the mesh size h?

No! (2) does not imply $(2)_h$ because W_h is a strict subset of V_2;
(3) does not imply $(3)_h$ because V_h is a strict subset of V_1.

Good news: We do not need to assume both $(2)_h$ and $(3)_h$: assuming $(2)_h$ suffices when $\dim(V_h) = \dim(W_h) < \infty$ to ensure existence a unique solution $u_h \in V_h$ to the problem $a(u_h, v_h) = F(v_h)$ for all $v_h \in W_h$, thanks to the *Rank-Nullity Theorem* from Linear Algebra.

Theorem

Assume that (1), (2), (3), (2)$_h$ and all hold. Then we have a well-posed discretisation of a well-posed problem. Furthermore

$$\|u - u_h\|_{V_1} \leq \left(1 + \frac{C}{\tilde{\gamma}}\right) \inf_{v_h \in V_h} \|u - v_h\|_{V_1}.$$

Proof.

For every $v_h \in V_h$, we have

$$\tilde{\gamma}\|v_h - u_h\|_{V_1} \leq \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u_h, w_h)}{\|w_h\|_{V_2}} \qquad \text{(discrete inf-sup } (2)_h)$$

Proof.

For every $v_h \in V_h$, we have

$$\begin{aligned}
\tilde{\gamma}\|v_h - u_h\|_{V_1} &\leq \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(discrete inf-sup } (2)_h\text{)} \\
&= \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h) + a(u - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(bilinearity of } a\text{)}
\end{aligned}$$

Proof.

For every $v_h \in V_h$, we have

$$
\begin{aligned}
\tilde{\gamma}\|v_h - u_h\|_{V_1} &\le \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(discrete inf-sup } (2)_h\text{)} \\
&= \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h) + a(u - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(bilinearity of } a\text{)} \\
&= \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h)}{\|w_h\|_{V_2}} && \text{(Galerkin orth.)}
\end{aligned}
$$

Proof.

For every $v_h \in V_h$, we have

$$
\begin{aligned}
\tilde{\gamma}\|v_h - u_h\|_{V_1} &\leq \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(discrete inf-sup } (2)_h\text{)} \\
&= \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h) + a(u - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(bilinearity of } a\text{)} \\
&= \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h)}{\|w_h\|_{V_2}} && \text{(Galerkin orth.)} \\
&\leq \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{C\|v_h - u\|_{V_1}\|w_h\|_{V_2}}{\|w_h\|_{V_2}} && \text{(bddness of } a\text{)}
\end{aligned}
$$

Proof.

For every $v_h \in V_h$, we have

$$
\begin{aligned}
\tilde{\gamma}\|v_h - u_h\|_{V_1} &\leq \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(discrete inf-sup } (2)_h\text{)} \\
&= \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h) + a(u - u_h, w_h)}{\|w_h\|_{V_2}} && \text{(bilinearity of } a\text{)} \\
&= \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{a(v_h - u, w_h)}{\|w_h\|_{V_2}} && \text{(Galerkin orth.)} \\
&\leq \sup_{\substack{w_h \in W_h \\ w_h \neq 0}} \frac{C\|v_h - u\|_{V_1}\|w_h\|_{V_2}}{\|w_h\|_{V_2}} && \text{(bddness of } a\text{)} \\
&= C\|v_h - u\|_{V_1}.
\end{aligned}
$$

Proof.

Now apply the triangle inequality to $\|u - u_h\|_{V_1}$:

$$\|u - u_h\|_{V_1} \leq \|u - v_h\|_{V_1} + \|v_h - u_h\|_{V_1}$$

$\square$

Proof.

Now apply the triangle inequality to $\|u - u_h\|_{V_1}$:

$$\begin{aligned}
\|u - u_h\|_{V_1} &\le \|u - v_h\|_{V_1} + \|v_h - u_h\|_{V_1} \\
&\le \|u - v_h\|_{V_1} + \frac{C}{\tilde{\gamma}} \|u - v_h\|_{V_1}
\end{aligned}$$

$\square$

Proof.

Now apply the triangle inequality to $\|u - u_h\|_{V_1}$:

$$\begin{aligned}\|u - u_h\|_{V_1} &\leq \|u - v_h\|_{V_1} + \|v_h - u_h\|_{V_1} \\ &\leq \|u - v_h\|_{V_1} + \frac{C}{\tilde{\gamma}}\|u - v_h\|_{V_1} \\ &= \left(1 + \frac{C}{\tilde{\gamma}}\right)\|u - v_h\|_{V_1}.\end{aligned}$$

□

Proof.

Now apply the triangle inequality to $\|u - u_h\|_{V_1}$:

$$\begin{aligned}
\|u - u_h\|_{V_1} &\leq \|u - v_h\|_{V_1} + \|v_h - u_h\|_{V_1} \\
&\leq \|u - v_h\|_{V_1} + \frac{C}{\tilde{\gamma}}\|u - v_h\|_{V_1} \\
&= \left(1 + \frac{C}{\tilde{\gamma}}\right)\|u - v_h\|_{V_1}.
\end{aligned}$$

Consequently,

$$\|u - u_h\|_{V_1} \leq \left(1 + \frac{C}{\tilde{\gamma}}\right) \inf_{v_h \in V_h} \|u - v_h\|_{V_1}.$$

□

Proof.

Now apply the triangle inequality to $\|u - u_h\|_{V_1}$:

$$\begin{aligned}
\|u - u_h\|_{V_1} &\leq \|u - v_h\|_{V_1} + \|v_h - u_h\|_{V_1} \\
&\leq \|u - v_h\|_{V_1} + \frac{C}{\tilde{\gamma}}\|u - v_h\|_{V_1} \\
&= \left(1 + \frac{C}{\tilde{\gamma}}\right)\|u - v_h\|_{V_1}.
\end{aligned}$$

Consequently,

$$\|u - u_h\|_{V_1} \leq \left(1 + \frac{C}{\tilde{\gamma}}\right)\inf_{v_h \in V_h}\|u - v_h\|_{V_1}.$$

□

As before, we can combine this with an approximation result and a regularity result to derive error estimates for finite element discretisations.

C6.4 Finite Element Methods for PDEs
Lecture 14: Saddle point problems

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

We have seen Babuška's conditions for the well-posedness of

$$\text{find } u \in V_1 \text{ such that } a(u, v) = F(v) \text{ for all } v \in V_2.$$

We have seen Babuška's conditions for the well-posedness of

$$\text{find } u \in V_1 \text{ such that } a(u,v) = F(v) \text{ for all } v \in V_2.$$

Many noncoercive problems arise via mixed formulations (solving for more than one variable), and in this lecture we will discuss the well-posedness conditions for saddle point problems: find $(u,p) \in V \times Q$ such that

$$\begin{aligned} a(u,v) + b(v,p) &= F(v) \\ b(u,q) \qquad\qquad &= G(q) \end{aligned}$$

for all $(v,q) \in V \times Q$.

We have seen Babuška's conditions for the well-posedness of

$$\text{find } u \in V_1 \text{ such that } a(u,v) = F(v) \text{ for all } v \in V_2.$$

Many noncoercive problems arise via mixed formulations (solving for more than one variable), and in this lecture we will discuss the well-posedness conditions for saddle point problems: find $(u,p) \in V \times Q$ such that

$$\begin{aligned} a(u,v) + b(v,p) &= F(v) \\ b(u,q) \qquad\qquad &= G(q) \end{aligned}$$

for all $(v,q) \in V \times Q$.

These are the *Brezzi* conditions.

Note that the problem: find $(u, p) \in V \times Q$ such that

$$\begin{aligned} a(u,v) + b(v,p) &= F(v) \\ b(u,q) \qquad\qquad &= G(q) \end{aligned}$$

for all $(v, q) \in V \times Q$

Note that the problem: find $(u, p) \in V \times Q$ such that

$$\begin{aligned} a(u, v) + b(v, p) &= F(v) \\ b(u, q) &= G(q) \end{aligned}$$

for all $(v, q) \in V \times Q$

is equivalent to the following:

Note that the problem: find $(u,p) \in V \times Q$ such that

$$\begin{aligned} a(u,v) + b(v,p) &= F(v) \\ b(u,q) \qquad\qquad &= G(q) \end{aligned}$$

for all $(v,q) \in V \times Q$

is equivalent to the following:

find $(u,p) \in V \times Q$ such that

$$a(u,v) + b(v,p) + b(u,q) = F(v) + G(q)$$

for all $(v,q) \in V \times Q$.

(Set $v = 0$ and vary $q \in Q$, set $q = 0$ and vary $v \in V$.)

We have already seen one example:

Mixed Poisson (lecture 5)

Find $(\sigma, u) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$\int_\Omega \sigma \cdot v \, \mathrm{d}x - \int_\Omega (\nabla \cdot v) u - \int_\Omega (\nabla \cdot \sigma) w \, \mathrm{d}x = - \int_\Omega f w \, \mathrm{d}x$$

for all $(v, w) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$.

We have already seen one example:

Mixed Poisson (lecture 5)

Find $(\sigma, u) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$\int_\Omega \sigma \cdot v \; \mathrm{d}x - \int_\Omega (\nabla \cdot v) u - \int_\Omega (\nabla \cdot \sigma) w \; \mathrm{d}x = -\int_\Omega f w \; \mathrm{d}x$$

for all $(v, w) \in H(\mathrm{div}, \Omega) \times L^2(\Omega)$.

Here

$$a(\sigma, v) = \int_\Omega \sigma \cdot v \; \mathrm{d}x, \quad b(v, u) = -\int_\Omega (\nabla \cdot v) u \; \mathrm{d}x.$$

We have already seen one system that is *not* an example:

Mixed linear elasticity (lecture 7)

Find $(u, p) \in H_0^1(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ such that

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x + \int_\Omega p(\nabla\cdot v)\,\mathrm{d}x + \int_\Omega q(\nabla\cdot u)\,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x = \int_\Omega f\cdot v\,\mathrm{d}x$$

for all $(v, q) \in H_0^1(\Omega; \mathbb{R}^2) \times L^2(\Omega)$.

We have already seen one system that is *not* an example:

Mixed linear elasticity (lecture 7)

Find $(u, p) \in H_0^1(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ such that

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x + \int_\Omega p(\nabla\cdot v)\,\mathrm{d}x + \int_\Omega q(\nabla\cdot u)\,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x = \int_\Omega f\cdot v\,\mathrm{d}x$$

for all $(v, q) \in H_0^1(\Omega; \mathbb{R}^2) \times L^2(\Omega)$.

We can restructure this as

$$\begin{aligned} a(u, v) + b(v, p) &= F(v) \\ b(u, q) + c(p, q) &= G(q) \end{aligned}$$

$$a(u, v) = \int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x,\ b(v, p) = \int_\Omega (\nabla\cdot v)p\,\mathrm{d}x,\ c(p, q) = -\frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x.$$

We have already seen one system that is *not* an example:

Mixed linear elasticity (lecture 7)

Find $(u, p) \in H^1_0(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ such that

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x + \int_\Omega p(\nabla\cdot v)\,\mathrm{d}x + \int_\Omega q(\nabla\cdot u)\,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x = \int_\Omega f\cdot v\,\mathrm{d}x$$

for all $(v, q) \in H^1_0(\Omega; \mathbb{R}^2) \times L^2(\Omega)$.

We can restructure this as

$$\begin{aligned} a(u, v) + b(v, p) &= F(v) \\ b(u, q) + c(p, q) &= G(q) \end{aligned}$$

$$a(u, v) = \int_\Omega 2\mu\varepsilon(u) : \varepsilon(v)\,\mathrm{d}x,\ b(v, p) = \int_\Omega (\nabla\cdot v)p\,\mathrm{d}x,\ c(p, q) = -\frac{1}{\lambda}\int_\Omega pq\,\mathrm{d}x.$$

We have already seen one system that is *not* an example:

Mixed linear elasticity (lecture 7)

Find $(u, p) \in H_0^1(\Omega; \mathbb{R}^n) \times L^2(\Omega)$ such that

$$\int_\Omega 2\mu\varepsilon(u) : \varepsilon(v) \,\mathrm{d}x + \int_\Omega p(\nabla\cdot v) \,\mathrm{d}x + \int_\Omega q(\nabla\cdot u) \,\mathrm{d}x - \frac{1}{\lambda}\int_\Omega pq \,\mathrm{d}x = \int_\Omega f\cdot v \,\mathrm{d}x$$

for all $(v, q) \in H_0^1(\Omega; \mathbb{R}^2) \times L^2(\Omega)$.

We can restructure this as

$$\begin{aligned} a(u, v) + b(v, p) &= F(v) \\ b(u, q) + c(p, q) &= G(q) \end{aligned}$$

$$a(u, v) = \int_\Omega 2\mu\varepsilon(u) : \varepsilon(v) \,\mathrm{d}x, \; b(v, p) = \int_\Omega (\nabla\cdot v)p \,\mathrm{d}x, \; c(p, q) = -\frac{1}{\lambda}\int_\Omega pq \,\mathrm{d}x.$$

This *is* a saddle point problem for $\lambda \to \infty$.

Let us consider one more example.

The Stokes equations are an elementary model in fluid mechanics. They describe the motion of a steady, incompressible, viscous, Newtonian, isothermal, slow-moving fluid.

$$
\begin{aligned}
-\nabla^2 u + \nabla p &= f \quad \text{in } \Omega, \\
\nabla \cdot u &= 0 \quad \text{in } \Omega, \\
u &= 0 \quad \text{on } \partial\Omega.
\end{aligned}
$$

Let us consider one more example.

The Stokes equations are an elementary model in fluid mechanics. They describe the motion of a steady, incompressible, viscous, Newtonian, isothermal, slow-moving fluid.

$$\begin{aligned} -\nabla^2 u + \nabla p &= f \quad \text{in } \Omega, \\ \nabla \cdot u &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Here $u : \Omega \to \mathbb{R}^n$ is the flow velocity and $p : \Omega \to \mathbb{R}$ is the pressure.

Multiply the momentum equation by a vector-valued test function $v \in V$, and the continuity equation by a scalar-valued test function $q \in Q$:

$$\int_\Omega (-\nabla \cdot \nabla u) \cdot v \,\mathrm{d}x + \int_\Omega \nabla p \cdot v \,\mathrm{d}x = \int_\Omega f \cdot v \,\mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \,\mathrm{d}x = 0.$$

Multiply the momentum equation by a vector-valued test function $v \in V$, and the continuity equation by a scalar-valued test function $q \in Q$:

$$\int_\Omega (-\nabla \cdot \nabla u) \cdot v \,\mathrm{d}x + \int_\Omega \nabla p \cdot v \,\mathrm{d}x = \int_\Omega f \cdot v \,\mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \,\mathrm{d}x = 0.$$

Integrate the vector Laplacian by parts:

$$\int_\Omega \nabla u : \nabla v \,\mathrm{d}x - \int_{\partial\Omega} (n \cdot \nabla u) \cdot v \,\mathrm{d}s + \int_\Omega \nabla p \cdot v \,\mathrm{d}x = \int_\Omega f \cdot v \,\mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \,\mathrm{d}x = 0.$$

Multiply the momentum equation by a vector-valued test function $v \in V$, and the continuity equation by a scalar-valued test function $q \in Q$:

$$\int_\Omega (-\nabla \cdot \nabla u) \cdot v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q (\nabla \cdot u) \, \mathrm{d}x = 0.$$

Integrate the vector Laplacian by parts:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_{\partial\Omega} (n \cdot \nabla u) \cdot v \, \mathrm{d}s + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q (\nabla \cdot u) \, \mathrm{d}x = 0.$$

We have nowhere to weakly enforce $u = 0$, so take $V = H_0^1(\Omega; \mathbb{R}^n)$.

Multiply the momentum equation by a vector-valued test function $v \in V$, and the continuity equation by a scalar-valued test function $q \in Q$:

$$\int_\Omega (-\nabla \cdot \nabla u) \cdot v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0.$$

Integrate the vector Laplacian by parts:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_{\partial\Omega} (n \cdot \nabla u) \cdot v \, \mathrm{d}s + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0.$$

We have nowhere to weakly enforce $u = 0$, so take $V = H_0^1(\Omega; \mathbb{R}^n)$.

Multiply the momentum equation by a vector-valued test function $v \in V$, and the continuity equation by a scalar-valued test function $q \in Q$:

$$\int_\Omega (-\nabla \cdot \nabla u) \cdot v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0.$$

Integrate the vector Laplacian by parts:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x \qquad + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0.$$

We have nowhere to weakly enforce $u = 0$, so take $V = H_0^1(\Omega; \mathbb{R}^n)$.

The formulation

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0,$$

requires $u \in H^1_0(\Omega; \mathbb{R}^n)$ and $p \in H^1(\Omega)$.

The formulation

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0,$$

requires $u \in H_0^1(\Omega; \mathbb{R}^n)$ and $p \in H^1(\Omega)$.

We can weaken the regularity requirement to $p \in L^2(\Omega)$ by integrating by parts, and then negating the second equation for symmetry:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_\Omega p(\nabla \cdot v) \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0,$$

The formulation

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0,$$

requires $u \in H_0^1(\Omega; \mathbb{R}^n)$ and $p \in H^1(\Omega)$.

We can weaken the regularity requirement to $p \in L^2(\Omega)$ by integrating by parts, and then negating the second equation for symmetry:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_\Omega p(\nabla \cdot v) \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$- \int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0,$$

The formulation

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x + \int_\Omega \nabla p \cdot v \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0,$$

requires $u \in H_0^1(\Omega; \mathbb{R}^n)$ and $p \in H^1(\Omega)$.

We can weaken the regularity requirement to $p \in L^2(\Omega)$ by integrating by parts, and then negating the second equation for symmetry:

$$\int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_\Omega p(\nabla \cdot v) \, \mathrm{d}x = \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$-\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0,$$

Here

$$a(u, v) = \int_\Omega \nabla u : \nabla v \, \mathrm{d}x, \quad b(v, p) = -\int_\Omega p(\nabla \cdot v) \, \mathrm{d}x.$$

In the strong form of the problem, p only appears via ∇p.

In the strong form of the problem, p only appears via ∇p.

So if (u, p) is a solution, so is $(u, p + c)$ for $c \in \mathbb{R}$. We can see this variationally:

$$\int_\Omega (p + c)(\nabla \cdot v) \; \mathrm{d}x = \int_\Omega p(\nabla \cdot v) \; \mathrm{d}x + c \int_\Omega \nabla \cdot v \; \mathrm{d}x$$

In the strong form of the problem, p only appears via ∇p.

So if (u, p) is a solution, so is $(u, p + c)$ for $c \in \mathbb{R}$. We can see this variationally:

$$\begin{aligned}\int_\Omega (p + c)(\nabla \cdot v) \, \mathrm{d}x &= \int_\Omega p(\nabla \cdot v) \, \mathrm{d}x + c \int_\Omega \nabla \cdot v \, \mathrm{d}x \\ &= \int_\Omega p(\nabla \cdot v) \, \mathrm{d}x + c \int_{\partial\Omega} v \cdot n \, \mathrm{d}s\end{aligned}$$

In the strong form of the problem, p only appears via ∇p.

So if (u, p) is a solution, so is $(u, p + c)$ for $c \in \mathbb{R}$. We can see this variationally:

$$\begin{aligned}\int_\Omega (p + c)(\nabla \cdot v)\ \mathrm{d}x &= \int_\Omega p(\nabla \cdot v)\ \mathrm{d}x + c \int_\Omega \nabla \cdot v\ \mathrm{d}x \\ &= \int_\Omega p(\nabla \cdot v)\ \mathrm{d}x + c \int_{\partial\Omega} v \cdot n\ \mathrm{d}s \\ &= \int_\Omega p(\nabla \cdot v)\ \mathrm{d}x.\end{aligned}$$

In the strong form of the problem, p only appears via ∇p.

So if (u, p) is a solution, so is $(u, p + c)$ for $c \in \mathbb{R}$. We can see this variationally:

$$\begin{aligned}\int_\Omega (p+c)(\nabla \cdot v)\ \mathrm{d}x &= \int_\Omega p(\nabla \cdot v)\ \mathrm{d}x + c\int_\Omega \nabla \cdot v\ \mathrm{d}x \\ &= \int_\Omega p(\nabla \cdot v)\ \mathrm{d}x + c\int_{\partial\Omega} v \cdot n\ \mathrm{d}s \\ &= \int_\Omega p(\nabla \cdot v)\ \mathrm{d}x.\end{aligned}$$

To fix a unique pressure we choose

$$Q = L^2_0(\Omega) = \{q \in L^2(\Omega) : \int_\Omega q\ \mathrm{d}x = 0\}.$$

Section 2

Energy minimisation

Many weak formulations arise from energy minimisation.

Many weak formulations arise from energy minimisation.

Many problems of saddle-point form arise from *constrained* minimisation.

Consider

$$u = \operatorname*{argmin}_{v \in H_0^1(\Omega;\mathbb{R}^n)} \ \frac{1}{2}\int_\Omega \nabla v : \nabla v \, \mathrm{d}x - \int_\Omega f \cdot v \, \mathrm{d}x,$$

$$\text{subject to} \qquad \nabla \cdot v = 0 \quad \text{in } \Omega.$$

Consider

$$u = \underset{v \in H_0^1(\Omega;\mathbb{R}^n)}{\operatorname{argmin}} \ \frac{1}{2}\int_\Omega \nabla v : \nabla v \, \mathrm{d}x - \int_\Omega f \cdot v \, \mathrm{d}x,$$
$$\text{subject to} \qquad \nabla \cdot v = 0 \quad \text{in } \Omega.$$

We introduce a Lagrange multiplier p and write the Lagrangian $L : H_0^1(\Omega;\mathbb{R}^n) \times L_0^2(\Omega) \to \mathbb{R}$:

$$L(u,p) = \frac{1}{2}\int_\Omega \nabla u : \nabla u \, \mathrm{d}x - \int_\Omega f \cdot u \, \mathrm{d}x - \int_\Omega p(\nabla \cdot u) \, \mathrm{d}x.$$

$$L(u,p) = \frac{1}{2}\int_\Omega \nabla u : \nabla u \, \mathrm{d}x - \int_\Omega p(\nabla \cdot u) \, \mathrm{d}x - \int_\Omega f \cdot u \, \mathrm{d}x.$$

$$L(u,p) = \frac{1}{2}\int_\Omega \nabla u : \nabla u \; \mathrm{d}x - \int_\Omega p(\nabla \cdot u) \; \mathrm{d}x - \int_\Omega f \cdot u \; \mathrm{d}x.$$

Calculating the Euler–Lagrange equations, we have

$$L_u(u,p;v) = \int_\Omega \nabla u : \nabla v \; \mathrm{d}x - \int_\Omega p(\nabla \cdot v) \; \mathrm{d}x - \int_\Omega f \cdot v \; \mathrm{d}x = 0,$$

$$L(u, p) = \frac{1}{2}\int_\Omega \nabla u : \nabla u \, \mathrm{d}x - \int_\Omega p(\nabla \cdot u) \, \mathrm{d}x - \int_\Omega f \cdot u \, \mathrm{d}x.$$

Calculating the Euler–Lagrange equations, we have

$$\begin{aligned} L_u(u, p; v) &= \int_\Omega \nabla u : \nabla v \, \mathrm{d}x - \int_\Omega p(\nabla \cdot v) \, \mathrm{d}x - \int_\Omega f \cdot v \, \mathrm{d}x = 0, \\ L_p(u, p; q) &= -\int_\Omega q(\nabla \cdot u) \, \mathrm{d}x = 0, \end{aligned}$$

the Stokes equations in weak form.

$$L(u, p) = \frac{1}{2}\int_\Omega \nabla u : \nabla u \,\mathrm{d}x - \int_\Omega p(\nabla \cdot u)\,\mathrm{d}x - \int_\Omega f \cdot u \,\mathrm{d}x.$$

Calculating the Euler–Lagrange equations, we have

$$\begin{aligned} L_u(u, p; v) &= \int_\Omega \nabla u : \nabla v \,\mathrm{d}x - \int_\Omega p(\nabla \cdot v)\,\mathrm{d}x - \int_\Omega f \cdot v \,\mathrm{d}x = 0, \\ L_p(u, p; q) &= -\int_\Omega q(\nabla \cdot u)\,\mathrm{d}x = 0, \end{aligned}$$

the Stokes equations in weak form.

In general constrained optimisation problems result in saddle point problems.

Section 3

Prelude: Orthogonal decompositions in Hilbert spaces

A useful fact: Hilbert spaces can be separated into any closed subspace and its orthogonal complement.

A useful fact: Hilbert spaces can be separated into any closed subspace and its orthogonal complement.

Theorem (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

$$K^{\perp} := \{v \in H : v \perp k \text{ for all } k \in K\}$$

is also a closed subspace, and

$$H = K \oplus K^{\perp},$$

which means that every $v \in H$ can be uniquely written as

$$v = v^{K} + v^{\perp},$$

with $v^{K} \in K$ and $v^{\perp} \in K^{\perp}$.

Section 4

Saddle point systems in finite dimensions: homogeneous case

Consider the following $N \times N$ linear system:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix},$$

where $A \in \mathbb{R}^{NA \times NA}$, $B \in \mathbb{R}^{NB \times NA}$, $NA + NB = N$, $NA > NB$.

Consider the following $N \times N$ linear system:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix},$$

where $A \in \mathbb{R}^{NA \times NA}$, $B \in \mathbb{R}^{NB \times NA}$, $NA + NB = N$, $NA > NB$.

For Stokes, A is the vector Laplacian, B^T is the pressure gradient, B is the divergence.

Consider the following $N \times N$ linear system:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix},$$

where $A \in \mathbb{R}^{NA \times NA}$, $B \in \mathbb{R}^{NB \times NA}$, $NA + NB = N$, $NA > NB$.

For Stokes, A is the vector Laplacian, B^T is the pressure gradient, B is the divergence.

Since the second equation says that $Bu = 0$, we know that

$$u \in K := \mathrm{kernel}(B) = \{v \in V : Bv = 0\}.$$

Consider the following $N \times N$ linear system:

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ 0 \end{pmatrix},$$

where $A \in \mathbb{R}^{NA \times NA}$, $B \in \mathbb{R}^{NB \times NA}$, $NA + NB = N$, $NA > NB$.

For Stokes, A is the vector Laplacian, B^T is the pressure gradient, B is the divergence.

Since the second equation says that $Bu = 0$, we know that

$$u \in K := \mathrm{kernel}(B) = \{v \in V : Bv = 0\}.$$

If $K = \{0\}$, then $u = 0$, and the system of equations reduces to $B^T p = f$. Let us therefore suppose thet $\dim(K) > 0$.

Let us test the first equation with $v \in K$:

$$v^T A u + v^T B^T p = v^T f,$$

Let us test the first equation with $v \in K$:

$$v^T A u + v^T B^T p = v^T f,$$

and since $v^T B^T p = p^T (Bv) = 0$, we derive the problem

$$\text{find } u \in K \text{ such that } v^T A u = v^T f \quad \text{for all } v \in K.$$

Let us test the first equation with $v \in K$:

$$v^T A u + v^T B^T p = v^T f,$$

and since $v^T B^T p = p^T (Bv) = 0$, we derive the problem

$$\text{find } u \in K \text{ such that } v^T A u = v^T f \quad \text{for all } v \in K.$$

This is a variational problem posed on a closed subspace K of V, and we know how to ensure its well-posedness:

Let us test the first equation with $v \in K$:

$$v^T A u + v^T B^T p = v^T f,$$

and since $v^T B^T p = p^T (Bv) = 0$, we derive the problem

$$\text{find } u \in K \text{ such that } v^T A u = v^T f \quad \text{for all } v \in K.$$

This is a variational problem posed on a closed subspace K of V, and we know how to ensure its well-posedness:

In the simpler case, assume the form is coercive on the kernel, i.e.

$$v^T A v \geq \alpha \|v\|^2 \text{ for all } v \in K.$$

In general, we assume the Babuška conditions hold.

Suppose we have solved the problem on the kernel for u, and let us see what we have achieved.

Suppose we have solved the problem on the kernel for u, and let us see what we have achieved.

Using the orthogonal decomposition, we know that we can write

$$f = f^K + f^\perp,$$

where $f^K \in K$ and $f^\perp \perp K$.

Suppose we have solved the problem on the kernel for u, and let us see what we have achieved.

Using the orthogonal decomposition, we know that we can write

$$f = f^K + f^\perp,$$

where $f^K \in K$ and $f^\perp \perp K$.

Testing with $v \in K$ yields

$$v^T f = v^T f^K + v^T f^\perp = v^T f^K,$$

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

$$B^T p = f - Au = f^{\perp}.$$

Having solved for the variable u, we must complete the solution of the problem by computing the unique $p \in Q$ such that

$$B^T p = f - Au = f^{\perp}.$$

By the fundamental theorem of linear algebra,

$$\mathrm{range}(B^T) = \mathrm{kernel}(B)^{\perp} = K^{\perp},$$

and since $f^{\perp} \in \mathrm{range}(B^T)$, there exists at least one $p \in Q$ such that $B^T p = f - Au$.

Having solved for the variable u, we must complete the solution of the problem by computing the unique $p \in Q$ such that

$$B^T p = f - Au = f^{\perp}.$$

By the fundamental theorem of linear algebra,

$$\mathrm{range}(B^T) = \mathrm{kernel}(B)^{\perp} = K^{\perp},$$

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 $B^T p = 0 \implies p = 0$.

To ensure injectivity, we wish to ensure that $B^T p = 0 \implies p = 0$.

We saw in the previous lecture that the inf-sup condition expresses this variationally: there exists a $\gamma \in \mathbb{R}$ such that

$$0 < \gamma \leq \inf_{\substack{q \in Q \\ q \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{v^T B^T q}{\|q\| \|v\|}.$$

To ensure injectivity, we wish to ensure that $B^T p = 0 \implies p = 0$.

We saw in the previous lecture that the inf-sup condition expresses this variationally: there exists a $\gamma \in \mathbb{R}$ such that

$$0 < \gamma \leq \inf_{\substack{q \in Q \\ q \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{v^T B^T q}{\|q\| \|v\|}.$$

If we assume this holds, then the operator $B^T : Q \to K^\perp$ is a bijection, and we can solve for $p \in Q$ uniquely.

Section 5

Saddle point systems in finite dimensions: the inhomogeneous case

Now consider the modified problem

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix},$$

the inhomogeneous case.

Now consider the modified problem

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix},$$

the inhomogeneous case.

Again, define K to be the kernel of B

$$K = \mathrm{kernel}(B)$$

and write

$$u = u^K + u^\perp.$$

Suppose that we change our basis so that we may write

$$u = \begin{pmatrix} u^K \\ u^\perp \end{pmatrix}, \quad A = \begin{pmatrix} A^{KK} & A^{K\perp} \\ A^{\perp K} & A^{\perp\perp} \end{pmatrix}.$$

Such a change of basis is always possible.

Suppose that we change our basis so that we may write

$$u = \begin{pmatrix} u^K \\ u^\perp \end{pmatrix}, \quad A = \begin{pmatrix} A^{KK} & A^{K\perp} \\ A^{\perp K} & A^{\perp\perp} \end{pmatrix}.$$

Such a change of basis is always possible.

We can therefore write

$$\begin{aligned} A^{KK} u^K + A^{K\perp} u^\perp &= f^K, \\ A^{\perp K} u^K + A^{\perp\perp} u^\perp + B^T p &= f^\perp, \\ B u^\perp &= g. \end{aligned}$$

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.

Can we solve $Bu^\perp = g$ for $u^\perp$? Yes! The inf-sup condition guarantees $B^T : Q \to K^\perp$ is a bijection, and so $B : K^\perp \to Q$ is also a bijection. We solve this for $u^\perp$.

Can we solve $Bu^\perp = g$ for $u^\perp$? Yes! The inf-sup condition guarantees $B^T : Q \to K^\perp$ is a bijection, and so $B : K^\perp \to Q$ is also a bijection. We solve this for $u^\perp$.

Testing the equation

$$A^{KK}u^K = f^K - A^{K\perp}u^\perp$$

with $v \in K$ yields a linear variational problem over K

find $u^K \in K$ such that $v^T A^{KK} u^K = v^T f^K - v^T A^{K\perp} u^\perp$ for all $v \in K$

as before. We solve the problem on the kernel for u^K.

Can we solve $Bu^{\perp} = g$ for $u^{\perp}$? Yes! The inf-sup condition guarantees $B^T : Q \to K^{\perp}$ is a bijection, and so $B : K^{\perp} \to Q$ is also a bijection. We solve this for $u^{\perp}$.

Testing the equation

$$A^{KK} u^K = f^K - A^{K\perp} u^{\perp}$$

with $v \in K$ yields a linear variational problem over K

find $u^K \in K$ such that $v^T A^{KK} u^K = v^T f^K - v^T A^{K\perp} u^{\perp}$ for all $v \in K$

as before. We solve the problem on the kernel for u^K.

We can then solve

$$B^T p = f^{\perp} - A^{\perp K} u^K - A^{\perp\perp} u^{\perp}$$

for p as before.

Can we solve $Bu^\perp = g$ for $u^\perp$? Yes! The inf-sup condition guarantees $B^T : Q \to K^\perp$ is a bijection, and so $B : K^\perp \to Q$ is also a bijection. We solve this for $u^\perp$.

Testing the equation

$$A^{KK} u^K = f^K - A^{K\perp} u^\perp$$

with $v \in K$ yields a linear variational problem over K

find $u^K \in K$ such that $v^T A^{KK} u^K = v^T f^K - v^T A^{K\perp} u^\perp$ for all $v \in K$

as before. We solve the problem on the kernel for u^K.

We can then solve

$$B^T p = f^\perp - A^{\perp K} u^K - A^{\perp\perp} u^\perp$$

for p as before.

So no further assumptions are required for the inhomogeneous case.

C6.4 Finite Element Methods for PDEs
Lecture 15: Brezzi's theorem and the mixed Poisson equation

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

Section 1

Brezzi's theorem

We now state the Brezzi conditions for the well-posedness of the abstract saddle point problem.

Theorem (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

$$\begin{aligned} a(u,v) + b(v,p) &= F(v), \\ b(u,q) &= G(q), \end{aligned}$$

for all $(v, q) \in V \times Q$. Let

$$K = \{v \in V : b(v,q) = 0 \textit{ for all } q \in Q\}.$$

Theorem (Well-posedness of saddle point problems)

Suppose that

(1) $a : V \times V \to \mathbb{R}$ *and* $b : V \times Q \to \mathbb{R}$ *are bounded bilinear forms;*

(2) The variational problem

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

is well-posed;

(3) b *satisfies the following inf-sup condition: there exists* $\gamma \in \mathbb{R}$ *such that*

$$0 < \gamma \leq \inf_{\substack{q \in Q \\ q \neq 0}} \sup_{\substack{v \in V \\ v \neq 0}} \frac{b(v, q)}{\|v\|_V \|q\|_Q}.$$

Then there exists a unique pair $(u, p) \in V \times Q$ *that solves the variational problem, and the solution is stable with respect to the data* F *and* G*.*

Section 2

Finite element discretisations of mixed problems

Take $V_h \times Q_h \subset V \times Q$, and consider: find $(u_h, p_h) \in V_h \times Q_h$ such that

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= F(v_h), \\ b(u_h, q_h) &= G(q_h), \end{aligned}$$

for all $(v_h, q_h) \in V_h \times Q_h$.

Take $V_h \times Q_h \subset V \times Q$, and consider: find $(u_h, p_h) \in V_h \times Q_h$ such that

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= F(v_h), \\ b(u_h, q_h) &= G(q_h), \end{aligned}$$

for all $(v_h, q_h) \in V_h \times Q_h$.

For this to be well-posed, Brezzi's conditions require that the linear variational problem: "*find* $u_h \in K_h$ *such that* $a(u_h, v_h) = F(v_h)$ *for all* $v_h \in K_h$" is well-posed on the *discrete* kernel

$$K_h = \{v_h \in V_h : b(v_h, q_h) = 0 \text{ for all } q_h \in Q_h\}.$$

Take $V_h \times Q_h \subset V \times Q$, and consider: find $(u_h, p_h) \in V_h \times Q_h$ such that

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= F(v_h), \\ b(u_h, q_h) &= G(q_h), \end{aligned}$$

for all $(v_h, q_h) \in V_h \times Q_h$.

For this to be well-posed, Brezzi's conditions require that the linear variational problem: "*find* $u_h \in K_h$ *such that* $a(u_h, v_h) = F(v_h)$ *for all* $v_h \in K_h$" is well-posed on the *discrete* kernel

$$K_h = \{v_h \in V_h : b(v_h, q_h) = 0 \text{ for all } q_h \in Q_h\}.$$

Compare K_h with

$$K \cap V_h = \{v_h \in V_h : b(v_h, q) = 0 \text{ for all } q \in Q\}.$$

In general, for $v_h \in V_h$, the property

$$b(v_h, q_h) = 0 \text{ for all } q_h \in Q_h$$

will not imply

$$b(v_h, q) = 0 \text{ for all } q \in Q.$$

In general, for $v_h \in V_h$, the property

$$b(v_h, q_h) = 0 \text{ for all } q_h \in Q_h$$

will not imply

$$b(v_h, q) = 0 \text{ for all } q \in Q.$$

(It will sometimes, but not always.)

In general, for $v_h \in V_h$, the property

$$b(v_h, q_h) = 0 \text{ for all } q_h \in Q_h$$

will not imply

$$b(v_h, q) = 0 \text{ for all } q \in Q.$$

(It will sometimes, but not always.)

So in general $K_h \not\subset K$. This means that *well-posedness of a on the discrete kernel K_h does not necessarily follow automatically from well-posedness of a on the full kernel K.*

That is one way a discretisation might fail. Any others?

That is one way a discretisation might fail. Any others?

Given that b satisfies the inf-sup condition over V and Q, it does *not* follow that b satisfies the inf-sup condition over V_h and Q_h, i.e. that: there exists $\tilde{\gamma} \in \mathbb{R}$ such that

$$0 < \tilde{\gamma} \leq \inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{b(v_h, q_h)}{\|v_h\| \|q_h\|}.$$

That is one way a discretisation might fail. Any others?

Given that b satisfies the inf-sup condition over V and Q, it does *not* follow that b satisfies the inf-sup condition over V_h and Q_h, i.e. that: there exists $\tilde{\gamma} \in \mathbb{R}$ such that

$$0 < \tilde{\gamma} \leq \inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{b(v_h, q_h)}{\|v_h\| \|q_h\|}.$$

We will see this by counterexample (later).

That is one way a discretisation might fail. Any others?

Given that b satisfies the inf-sup condition over V and Q, it does *not* follow that b satisfies the inf-sup condition over V_h and Q_h, i.e. that: there exists $\tilde{\gamma} \in \mathbb{R}$ such that

$$0 < \tilde{\gamma} \leq \inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{b(v_h, q_h)}{\|v_h\| \|q_h\|}.$$

We will see this by counterexample (later).

We must *additionally* assume that the Brezzi conditions hold for our discrete problem. This is a *compatibility* condition on the elements we choose for V_h and Q_h: they must work together.

Section 3

Quasi-optimality

Theorem

Consider the Galerkin approximation of our saddle point problem over $V_h \times Q_h$, a closed subspace of $V \times Q$:

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= F(v_h), \\ b(u_h, q_h) &= G(q_h). \end{aligned}$$

Let

$$K_h = \{v_h \in V_h : b(v_h, q_h) = 0 \textit{ for all } q_h \in Q_h\}.$$

Theorem

In addition to the assumptions of Brezzi's theorem that guarantee well-posedness of the continuous problem, suppose that

(1) The variational problem

$$\text{find } u_h \in K_h \text{ such that } a(u_h, v_h) = F(v_h) \text{ for all } v_h \in K_h$$

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

$$0 < \tilde{\gamma} \leq \inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{v_h \in V_h \\ v_h \neq 0}} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_Q}.$$

Then the Galerkin approximation is well-posed.

Theorem

Furthermore, the approximate solutions are quasi-optimal: there exists a $c < \infty$ *such that*

$$\|u - u_h\|_V + \|p - p_h\|_Q \le c \left(\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right).$$

Theorem

Furthermore, the approximate solutions are quasi-optimal: there exists a $c < \infty$ *such that*

$$\|u - u_h\|_V + \|p - p_h\|_Q \leq c \left(\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right).$$

This means that generally we have to think about the quality of the approximation for u and p together: there is probably no point having a very high-order discretisation for u and a very low-order one for p!

Theorem

Furthermore, the approximate solutions are quasi-optimal: there exists a $c < \infty$ *such that*

$$\|u - u_h\|_V + \|p - p_h\|_Q \leq c \left(\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q \right).$$

This means that generally we have to think about the quality of the approximation for u and p together: there is probably no point having a very high-order discretisation for u and a very low-order one for p!

We will apply this theory to the mixed Poisson equation.

Section 4

Mixed Poisson equation in 1D

Let us consider the mixed Poisson equation in one dimension. Start with

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

Let us consider the mixed Poisson equation in one dimension. Start with

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

and introduce $\sigma = -u'$ to get the system

$$\begin{aligned} \sigma + u' &= 0, \\ \sigma' &= f. \end{aligned}$$

Let us consider the mixed Poisson equation in one dimension. Start with

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

and introduce $\sigma = -u'$ to get the system

$$\begin{aligned} \sigma + u' &= 0, \\ \sigma' &= f. \end{aligned}$$

Testing the equations with $(\tau, v) \in V \times Q$, we get

$$\begin{aligned} \int_\Omega \sigma\tau \,\mathrm{d}x + \int_\Omega u'\tau \,\mathrm{d}x &= 0, \\ \int_\Omega \sigma' v \,\mathrm{d}x &= \int_\Omega f v \,\mathrm{d}x. \end{aligned}$$

$$
\begin{aligned}
\int_\Omega \sigma\tau \;\mathrm{d}x + \int_\Omega u'\tau \;\mathrm{d}x &= 0,\\
\int_\Omega \sigma' v \;\mathrm{d}x &= \int_\Omega f v \;\mathrm{d}x.
\end{aligned}
$$

$$\begin{aligned}\int_\Omega \sigma\tau \,\mathrm{d}x + \int_\Omega u'\tau \,\mathrm{d}x &= 0,\\ \int_\Omega \sigma' v \,\mathrm{d}x &= \int_\Omega fv \,\mathrm{d}x.\end{aligned}$$

As it stands we need both $\sigma, \tau \in H^1(\Omega)$ and $u, v \in H^1(\Omega)$. Let us integrate by parts to remove the derivative from u onto τ, and negate:

$$\begin{aligned}\int_\Omega \sigma\tau \,\mathrm{d}x - \int_\Omega u\tau' \,\mathrm{d}x + \int_{\partial\Omega} u\tau \,\mathrm{d}s &= 0,\\ -\int_\Omega \sigma' v \,\mathrm{d}x &= -\int_\Omega fv \,\mathrm{d}x.\end{aligned}$$

$$\begin{aligned}\int_\Omega \sigma\tau \, \mathrm{d}x + \int_\Omega u'\tau \, \mathrm{d}x &= 0,\\ \int_\Omega \sigma' v \, \mathrm{d}x &= \int_\Omega f v \, \mathrm{d}x.\end{aligned}$$

As it stands we need both $\sigma, \tau \in H^1(\Omega)$ and $u, v \in H^1(\Omega)$. Let us integrate by parts to remove the derivative from u onto τ, and negate:

$$\begin{aligned}\int_\Omega \sigma\tau \, \mathrm{d}x - \int_\Omega u\tau' \, \mathrm{d}x + \int_{\partial\Omega} u\tau \, \mathrm{d}s &= 0,\\ -\int_\Omega \sigma' v \, \mathrm{d}x &= -\int_\Omega f v \, \mathrm{d}x.\end{aligned}$$

We impose the Dirichlet BCs on u naturally by dropping the boundary term.

We thus have: find $(\sigma, u) \in V \times Q := H^1(\Omega) \times L^2(\Omega)$ such that

$$\int_\Omega \sigma\tau \, \mathrm{d}x - \int_\Omega u\tau' \, \mathrm{d}x - \int_\Omega \sigma' v \, \mathrm{d}x = -\int_\Omega fv \, \mathrm{d}x$$

for all $(\tau, v) \in V \times Q$.

We thus have: find $(\sigma, u) \in V \times Q := H^1(\Omega) \times L^2(\Omega)$ such that

$$\int_\Omega \sigma\tau \,\mathrm{d}x - \int_\Omega u\tau' \,\mathrm{d}x - \int_\Omega \sigma' v \,\mathrm{d}x = -\int_\Omega fv \,\mathrm{d}x$$

for all $(\tau, v) \in V \times Q$.

Let us think about well-posedness. Is

$$a(\sigma, \tau) = \int_\Omega \sigma\tau \,\mathrm{d}x = (\sigma, \tau)_{L^2(\Omega)}$$

coercive over the kernel

$$K := \{\tau \in H^1(\Omega) : b(\tau, v) := \int_\Omega \tau' v \,\mathrm{d}x = 0 \text{ for all } v \in L^2(\Omega)\}?$$

We thus have: find $(\sigma, u) \in V \times Q := H^1(\Omega) \times L^2(\Omega)$ such that

$$\int_\Omega \sigma\tau \,\mathrm{d}x - \int_\Omega u\tau' \,\mathrm{d}x - \int_\Omega \sigma' v \,\mathrm{d}x = -\int_\Omega f v \,\mathrm{d}x$$

for all $(\tau, v) \in V \times Q$.

Let us think about well-posedness. Is

$$a(\sigma, \tau) = \int_\Omega \sigma\tau \,\mathrm{d}x = (\sigma, \tau)_{L^2(\Omega)}$$

coercive over the kernel

$$K := \{\tau \in H^1(\Omega) : b(\tau, v) := \int_\Omega \tau' v \,\mathrm{d}x = 0 \text{ for all } v \in L^2(\Omega)\}?$$

Since $\tau' \in L^2(\Omega)$, choosing $v = \tau'$ as test function yields that

$$\tau \in K \iff \tau' = 0.$$

For coercivity on the kernel, note that

$$a(\tau,\tau) = \|\tau\|^2_{L^2(\Omega)} = \|\tau\|^2_{L^2(\Omega)} + \|\tau'\|^2_{L^2(\Omega)} = \|\tau\|^2_{H^1(\Omega)},$$

so it is coercive with constant $\alpha = 1$. This is *only* true on the kernel, where $\tau' = 0$. The bilinear form $a(\cdot,\cdot)$ is *not* coercive on the whole of $H^1(\Omega)$!

For the inf-sup condition, we require that there exists a γ such that

$$0 < \gamma \leq \inf_{\substack{v \in L^2(\Omega) \\ v \neq 0}} \sup_{\substack{\tau \in H^1(\Omega) \\ \tau \neq 0}} \frac{\int_\Omega \tau' v \, \mathrm{d}x}{\|\tau\|_{H^1(\Omega)} \|v\|_{L^2(\Omega)}}.$$

For the inf-sup condition, we require that there exists a γ such that

$$0 < \gamma \leq \inf_{\substack{v \in L^2(\Omega) \\ v \neq 0}} \sup_{\substack{\tau \in H^1(\Omega) \\ \tau \neq 0}} \frac{\int_\Omega \tau' v \, \mathrm{d}x}{\|\tau\|_{H^1(\Omega)} \|v\|_{L^2(\Omega)}}.$$

For a given $v \in L^2(\Omega)$, choose

$$\tau(x) = \int_0^x v(s) \, \mathrm{d}s$$

so that $\tau' = v$ and $\tau(0) = 0$. We saw in Lecture 5, slides 7–8, that for such a function $\|\tau\|_{L^2(\Omega)} \leq c\|\tau'\|_{L^2(\Omega)} = c\|v\|_{L^2(\Omega)}$ and hence

$$\|\tau\|_{H^1(\Omega)} \leq c\|v\|_{L^2(\Omega)}$$

for some (different) c.

With this choice, for any $v \in L^2(\Omega)$,

$$\sup_{\substack{\tau \in V \\ \tau \neq 0}} \frac{\int_\Omega \tau' v \, \mathrm{d}x}{\|\tau\|_{H^1(\Omega)} \|v\|_{L^2(\Omega)}} \geq \frac{\|v\|^2_{L^2(\Omega)}}{c\|v\|^2_{L^2(\Omega)}} = \frac{1}{c} > 0$$

so the inf-sup condition holds.

With this choice, for any $v \in L^2(\Omega)$,

$$\sup_{\substack{\tau \in V \\ \tau \neq 0}} \frac{\int_\Omega \tau' v \, \mathrm{d}x}{\|\tau\|_{H^1(\Omega)} \|v\|_{L^2(\Omega)}} \geq \frac{\|v\|^2_{L^2(\Omega)}}{c\|v\|^2_{L^2(\Omega)}} = \frac{1}{c} > 0$$

so the inf-sup condition holds.

Applying Brezzi's theorem, we conclude that the mixed formulation is well-posed.

C6.4 Finite Element Methods for PDEs
Lecture 16: The mixed Poisson equation (continued)

Endre Süli
(slides by courtesy of Patrick E. Farrell)

University of Oxford

Section 1

Discretising the mixed Poisson equation in 1D

Let us consider three different discretisations for $V_h \times Q_h$:

$$
\begin{aligned}
&(\mathrm{A}) \quad \mathrm{CG}_1 \times \mathrm{CG}_1 \\
&(\mathrm{B}) \quad \mathrm{CG}_1 \times \mathrm{DG}_0 \\
&(\mathrm{C}) \quad \mathrm{CG}_2 \times \mathrm{DG}_0
\end{aligned}
$$

Let us consider three different discretisations for $V_h \times Q_h$:

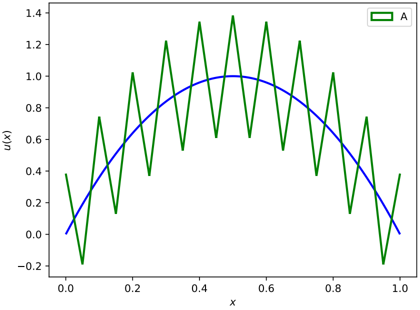
$$\begin{aligned}&(\mathrm{A})\ \ \mathrm{CG}_1 \times \mathrm{CG}_1\\&(\mathrm{B})\ \ \mathrm{CG}_1 \times \mathrm{DG}_0\\&(\mathrm{C})\ \ \mathrm{CG}_2 \times \mathrm{DG}_0\end{aligned}$$

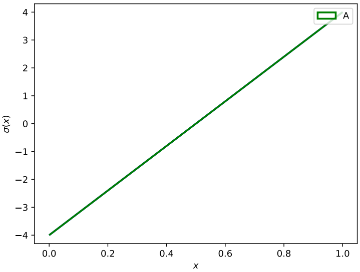
Try them with $f = 8$, so that the exact solution is $u = -4x(x-1)$:

Let us consider three different discretisations for $V_h \times Q_h$:

$$\begin{aligned} &(\mathrm{A}) \quad \mathrm{CG}_1 \times \mathrm{CG}_1 \\ &(\mathrm{B}) \quad \mathrm{CG}_1 \times \mathrm{DG}_0 \\ &(\mathrm{C}) \quad \mathrm{CG}_2 \times \mathrm{DG}_0 \end{aligned}$$

Try them with $f = 8$, so that the exact solution is $u = -4x(x-1)$:

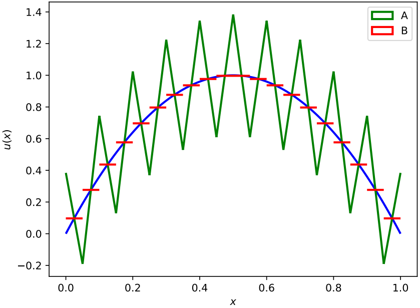


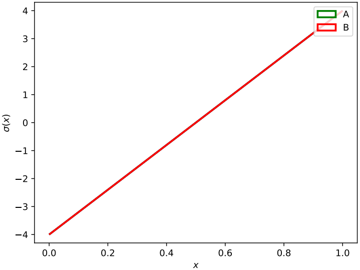


Let us consider three different discretisations for $V_h \times Q_h$:

$$
\begin{aligned}
&\text{(A)} \quad \mathrm{CG}_1 \times \mathrm{CG}_1\\
&\text{(B)} \quad \mathrm{CG}_1 \times \mathrm{DG}_0\\
&\text{(C)} \quad \mathrm{CG}_2 \times \mathrm{DG}_0
\end{aligned}
$$

Try them with $f = 8$, so that the exact solution is $u = -4x(x-1)$:

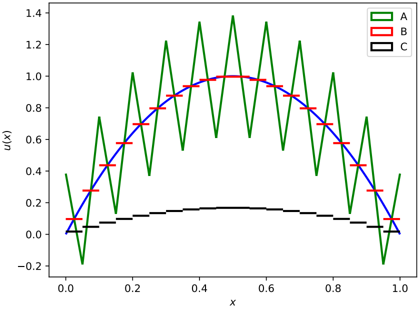


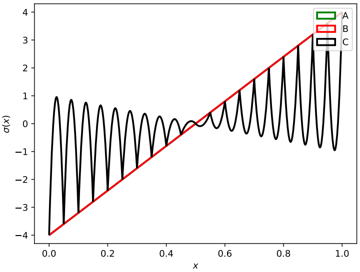


Let us consider three different discretisations for $V_h \times Q_h$:

$$\begin{aligned} &(\mathrm{A}) \quad \mathrm{CG}_1 \times \mathrm{CG}_1 \\ &(\mathrm{B}) \quad \mathrm{CG}_1 \times \mathrm{DG}_0 \\ &(\mathrm{C}) \quad \mathrm{CG}_2 \times \mathrm{DG}_0 \end{aligned}$$

Try them with $f = 8$, so that the exact solution is $u = -4x(x-1)$:





Subsection 1

Discretisation (A): $\mathrm{CG}_1 \times \mathrm{CG}_1$

The discrete inf-sup condition is that there exists a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \le \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}.$$

The discrete inf-sup condition is that there exists a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}.$$

Let us choose $v_h \in Q_h$ in the worst possible way. On each cell v_h is a linear polynomial and τ_h' is a constant.

The discrete inf-sup condition is that there exists a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}.$$

Let us choose $v_h \in Q_h$ in the worst possible way. On each cell v_h is a linear polynomial and τ_h' is a constant.

For integrands of polynomial degree one, midpoint quadrature is exact.

The discrete inf-sup condition is that there exists a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}.$$

Let us choose $v_h \in Q_h$ in the worst possible way. On each cell v_h is a linear polynomial and τ_h' is a constant.

For integrands of polynomial degree one, midpoint quadrature is exact.

If we choose v_h to have value zero at each midpoint, then for all $\tau_h \in V_h$:

$$\int_\Omega \tau_h' v_h \, \mathrm{d}x = \sum_{I \in \mathcal{M}} \int_I \tau_h' v_h \, \mathrm{d}x$$

The discrete inf-sup condition is that there exists a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \le \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}.$$

Let us choose $v_h \in Q_h$ in the worst possible way. On each cell v_h is a linear polynomial and τ_h' is a constant.

For integrands of polynomial degree one, midpoint quadrature is exact.

If we choose v_h to have value zero at each midpoint, then for all $\tau_h \in V_h$:

$$\begin{aligned}
\int_\Omega \tau_h' v_h \, \mathrm{d}x &= \sum_{I \in \mathcal{M}} \int_I \tau_h' v_h \, \mathrm{d}x \\
&= \sum_{I \in \mathcal{M}} \tau_h' v_h \, (\mathrm{midpoint}(I))
\end{aligned}$$

The discrete inf-sup condition is that there exists a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}.$$

Let us choose $v_h \in Q_h$ in the worst possible way. On each cell v_h is a linear polynomial and τ_h' is a constant.

For integrands of polynomial degree one, midpoint quadrature is exact.

If we choose v_h to have value zero at each midpoint, then for all $\tau_h \in V_h$:

$$\begin{aligned} \int_\Omega \tau_h' v_h \, \mathrm{d}x &= \sum_{I \in \mathcal{M}} \int_I \tau_h' v_h \, \mathrm{d}x \\ &= \sum_{I \in \mathcal{M}} \tau_h' v_h \, (\mathrm{midpoint}(I)) \\ &= 0 \end{aligned}$$

for all $\tau_h \in V_h$.

Thus, for this discretisation, for each $\tilde{\gamma} > 0$ there exists a $v_h \in Q_h$ such that, for all $\tau_h \in V_h$:

$$0 = \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} \leq \tilde{\gamma}.$$

Thus, for this discretisation, for each $\tilde{\gamma} > 0$ there exists a $v_h \in Q_h$ such that, for all $\tau_h \in V_h$:

$$0 = \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} \leq \tilde{\gamma}.$$

Conclusion

The $\mathrm{CG}_1 \times \mathrm{CG}_1$ discretisation is unstable: it fails the inf-sup condition.

Subsection 2

Discretisation (B): $\mathrm{CG}_1 \times \mathrm{DG}_0$

Let us check the two discrete Brezzi conditions.

Let us check the two discrete Brezzi conditions.

The discrete kernel is

$$K_h := \{\tau_h \in V_h : b(\tau_h, v_h) := \int_\Omega \tau_h' v_h \,\mathrm{d}x = 0 \text{ for all } v_h \in Q_h\}.$$

Let us check the two discrete Brezzi conditions.

The discrete kernel is

$$K_h := \{\tau_h \in V_h : b(\tau_h, v_h) := \int_\Omega \tau_h' v_h \, \mathrm{d}x = 0 \text{ for all } v_h \in Q_h\}.$$

The derivative of a CG_1 function is a DG_0 function.

Let us check the two discrete Brezzi conditions.

The discrete kernel is

$$K_h := \{\tau_h \in V_h : b(\tau_h, v_h) := \int_\Omega \tau_h' v_h \, \mathrm{d}x = 0 \text{ for all } v_h \in Q_h\}.$$

The derivative of a CG_1 function is a DG_0 function.

Thus, if $\tau_h \in K_h$, choosing $v_h = \tau_h'$ yields that $\tau_h' = 0$, so $K_h = K \cap V_h$.

Let us check the two discrete Brezzi conditions.

The discrete kernel is

$$K_h := \{\tau_h \in V_h : b(\tau_h, v_h) := \int_\Omega \tau_h' v_h \,\mathrm{d}x = 0 \text{ for all } v_h \in Q_h\}.$$

The derivative of a CG_1 function is a DG_0 function.

Thus, if $\tau_h \in K_h$, choosing $v_h = \tau_h'$ yields that $\tau_h' = 0$, so $K_h = K \cap V_h$.

Hence, for $\tau_h \in K_h$,

$$a(\tau_h, \tau_h) = \|\tau_h\|^2_{L^2(\Omega)} = \|\tau_h\|^2_{L^2(\Omega)} + \|\tau_h'\|^2_{L^2(\Omega)} = \|\tau_h\|^2_{H^1(\Omega)},$$

so the $a(\tau_h, \tau_h)$ form is coercive over K_h.

What about the discrete inf-sup condition? Does there exist a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} ?$$

What about the discrete inf-sup condition? Does there exist a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}?$$

In infinite dimensions we proved this by constructing a $\tau \in V = H^1(\Omega)$ with $\tau(0) = 0$ for any $v \in Q = L^2(\Omega)$ by choosing

$$\tau(x) = \int_0^x v(s) \, \mathrm{d}s.$$

What about the discrete inf-sup condition? Does there exist a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} ?$$

In infinite dimensions we proved this by constructing a $\tau \in V = H^1(\Omega)$ with $\tau(0) = 0$ for any $v \in Q = L^2(\Omega)$ by choosing

$$\tau(x) = \int_0^x v(s) \, \mathrm{d}s.$$

If v_h is piecewise constant, the associated $\tau_h(x) = \int_0^x v_h(s)\mathrm{d}s$ is piecewise linear and continuous, so $\tau_h \in V_h$! So the inf-sup argument works in the same way.

What about the discrete inf-sup condition? Does there exist a $\tilde{\gamma} > 0$ such that

$$\tilde{\gamma} \leq \inf_{\substack{v_h \in Q_h \\ v_h \neq 0}} \sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \, \mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}?$$

In infinite dimensions we proved this by constructing a $\tau \in V = H^1(\Omega)$ with $\tau(0) = 0$ for any $v \in Q = L^2(\Omega)$ by choosing

$$\tau(x) = \int_0^x v(s) \, \mathrm{d}s.$$

If v_h is piecewise constant, the associated $\tau_h(x) = \int_0^x v_h(s) \mathrm{d}s$ is piecewise linear and continuous, so $\tau_h \in V_h$! So the inf-sup argument works in the same way.

Conclusion

This discretisation is stable, by Brezzi's theorem.

These are the same arguments that worked in infinite dimensions. Why did they work again?

These are the same arguments that worked in infinite dimensions. Why did they work again?

$$
\begin{array}{ccc}
H^1 & \xrightarrow{\frac{\mathrm{d}}{\mathrm{d}x}} & L^2 \\
\big\downarrow P_h & & \big\downarrow \Pi_h \\
V_h & \xrightarrow{\frac{\mathrm{d}}{\mathrm{d}x}} & Q_h
\end{array}
$$

Here P_h is the orthogonal projector onto $V_h \subset H^1(\Omega)$ with vanishing value at $x = 0$ in the inner product (τ', σ'); and Π_h is the orthogonal projector onto $Q_h \subset L^2(\Omega)$ in the inner product (v, w).

These are the same arguments that worked in infinite dimensions. Why did they work again?

$$\begin{tikzcd} H^1 \arrow[r, "\frac{\mathrm{d}}{\mathrm{d}x}"] \arrow[d, "P_h"] & L^2 \arrow[d, "\Pi_h"] \\ V_h \arrow[r, "\frac{\mathrm{d}}{\mathrm{d}x}"] & Q_h \end{tikzcd}$$

Here P_h is the orthogonal projector onto $V_h \subset H^1(\Omega)$ with vanishing value at $x = 0$ in the inner product (τ', σ'); and Π_h is the orthogonal projector onto $Q_h \subset L^2(\Omega)$ in the inner product (v, w).

Structure preservation

The discrete Brezzi condition holds because our choice of the finite element spaces V_h and Q_h mimics the structure of the infinite-dimensional problem: the diagram commutes.

Subsection 3

Discretisation (C): $\mathrm{CG}_2 \times \mathrm{DG}_0$

First let us consider the inf-sup condition. Let V_h be constructed with CG_2 elements, and let $\tilde{V}_h \subsetneq V_h$ be constructed with CG_1 elements. Then, for any $v_h \in Q_h$,

$$\sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} \geq \sup_{\substack{\tau_h \in \tilde{V}_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}$$

First let us consider the inf-sup condition. Let V_h be constructed with CG_2 elements, and let $\tilde{V}_h \subsetneq V_h$ be constructed with CG_1 elements. Then, for any $v_h \in Q_h$,

$$\begin{aligned}\sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} &\geq \sup_{\substack{\tau_h \in \tilde{V}_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} \\ &\geq \tilde{\gamma} > 0,\end{aligned}$$

since the discretisation (B) *is* stable.

First let us consider the inf-sup condition. Let V_h be constructed with CG_2 elements, and let $\tilde{V}_h \subsetneq V_h$ be constructed with CG_1 elements. Then, for any $v_h \in Q_h$,

$$\begin{aligned}\sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} &\geq \sup_{\substack{\tau_h \in \tilde{V}_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} \\ &\geq \tilde{\gamma} > 0,\end{aligned}$$

since the discretisation (B) *is* stable.

In other words, enriching the discretisation of the V-space can only *improve* the inf-sup condition.

First let us consider the inf-sup condition. Let V_h be constructed with CG_2 elements, and let $\tilde{V}_h \subsetneq V_h$ be constructed with CG_1 elements. Then, for any $v_h \in Q_h$,

$$\sup_{\substack{\tau_h \in V_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}} \geq \sup_{\substack{\tau_h \in \tilde{V}_h \\ \tau_h \neq 0}} \frac{\int_\Omega \tau_h' v_h \,\mathrm{d}x}{\|\tau_h\|_{H^1(\Omega)} \|v_h\|_{L^2(\Omega)}}$$
$$\geq \tilde{\gamma} > 0,$$

since the discretisation (B) *is* stable.

In other words, enriching the discretisation of the V-space can only *improve* the inf-sup condition.

What about the well-posedness of the LVP involving a on the kernel?

Increasing the size of V_h makes the discrete kernel K_h larger, so it is harder to satisfy coercivity on K_h.

Increasing the size of V_h makes the discrete kernel K_h larger, so it is harder to satisfy coercivity on K_h.

Here the discrete kernel is:

$$K_h = \{\tau_h \in V_h : b(\tau_h, v_h) := \int_\Omega \tau_h' v_h \,\mathrm{d}x = 0 \text{ for all } v_h \in Q_h\}.$$

Increasing the size of V_h makes the discrete kernel K_h larger, so it is harder to satisfy coercivity on K_h.

Here the discrete kernel is:

$$K_h = \{\tau_h \in V_h : b(\tau_h, v_h) := \int_\Omega \tau_h' v_h \,\mathrm{d}x = 0 \text{ for all } v_h \in Q_h\}.$$

Consider a single mesh cell $I = [\bar{x}, \bar{x} + h]$. Define

$$\tau_h(x) = (x - \bar{x})(x - (\bar{x} + h))$$

on I, and zero elsewhere. We have $\tau_h \in V_h$.

Claim: $\tau_h \in K_h$.

Calculating, on I,

$$\tau_h(x) = x^2 - x(\bar{x} + h) - x\bar{x} + \bar{x}(\bar{x} + h),$$

and so

$$\tau_h'(x) = 2x - 2\bar{x} - h.$$

Calculating, on I,

$$\tau_h(x) = x^2 - x(\bar{x} + h) - x\bar{x} + \bar{x}(\bar{x} + h),$$

and so

$$\tau_h'(x) = 2x - 2\bar{x} - h.$$

We know that for integrands of degree 1, midpoint quadrature is exact, so

$$\int_{\bar{x}}^{\bar{x}+h} \tau_h' v_h \, \mathrm{d}x = v_h \int_{\bar{x}}^{\bar{x}+h} \tau_h' \, \mathrm{d}x = h v_h \tau_h'(\bar{x} + h/2)$$

and thus we evaluate $\tau_h'(\bar{x} + h/2)$:

Calculating, on I,

$$\tau_h(x) = x^2 - x(\bar{x}+h) - x\bar{x} + \bar{x}(\bar{x}+h),$$

and so

$$\tau_h'(x) = 2x - 2\bar{x} - h.$$

We know that for integrands of degree 1, midpoint quadrature is exact, so

$$\int_{\bar{x}}^{\bar{x}+h} \tau_h' v_h \,\mathrm{d}x = v_h \int_{\bar{x}}^{\bar{x}+h} \tau_h' \,\mathrm{d}x = h v_h \tau_h'(\bar{x}+h/2)$$

and thus we evaluate $\tau_h'(\bar{x}+h/2)$:

$$\tau_h'\left(\bar{x}+\frac{h}{2}\right) = 2\bar{x} + h - 2\bar{x} - h = 0.$$

Calculating, on I,

$$\tau_h(x) = x^2 - x(\bar{x} + h) - x\bar{x} + \bar{x}(\bar{x} + h),$$

and so

$$\tau_h'(x) = 2x - 2\bar{x} - h.$$

We know that for integrands of degree 1, midpoint quadrature is exact, so

$$\int_{\bar{x}}^{\bar{x}+h} \tau_h' v_h \, \mathrm{d}x = v_h \int_{\bar{x}}^{\bar{x}+h} \tau_h' \, \mathrm{d}x = h v_h \tau_h'(\bar{x} + h/2)$$

and thus we evaluate $\tau_h'(\bar{x} + h/2)$:

$$\tau_h'\left(\bar{x} + \frac{h}{2}\right) = 2\bar{x} + h - 2\bar{x} - h = 0.$$

So $\tau_h \in K_h$.

So now we must decide whether $a(\sigma, \tau) = (\sigma, \tau)_{L^2(\Omega)}$ is coercive on K_h.

So now we must decide whether $a(\sigma, \tau) = (\sigma, \tau)_{L^2(\Omega)}$ is coercive on K_h.

Taking τ_h as previously constructed,

$$a(\tau_h, \tau_h) = \mathcal{O}(h^5), \quad \|\tau_h\|^2_{H^1(\Omega)} = \mathcal{O}(h^3),$$

so

$$a(\tau_h, \tau_h)/\|\tau_h\|^2_{H^1(\Omega)} = \mathcal{O}(h^2) \to 0 \quad \text{as } h \to 0.$$

So now we must decide whether $a(\sigma, \tau) = (\sigma, \tau)_{L^2(\Omega)}$ is coercive on K_h.

Taking τ_h as previously constructed,

$$a(\tau_h, \tau_h) = \mathcal{O}(h^5), \quad \|\tau_h\|^2_{H^1(\Omega)} = \mathcal{O}(h^3),$$

so

$$a(\tau_h, \tau_h)/\|\tau_h\|^2_{H^1(\Omega)} = \mathcal{O}(h^2) \to 0 \quad \text{as } h \to 0.$$

So there is a loss of coercivity on the kernel as the mesh is refined.

Assembling on any given mesh yields a linear system with an invertible square matrix, but the resulting sequence of approximations does not converge to the exact solution as the mesh is refined.

So now we must decide whether $a(\sigma, \tau) = (\sigma, \tau)_{L^2(\Omega)}$ is coercive on K_h.

Taking τ_h as previously constructed,

$$a(\tau_h, \tau_h) = \mathcal{O}(h^5), \quad \|\tau_h\|^2_{H^1(\Omega)} = \mathcal{O}(h^3),$$

so

$$a(\tau_h, \tau_h)/\|\tau_h\|^2_{H^1(\Omega)} = \mathcal{O}(h^2) \to 0 \quad \text{as } h \to 0.$$

So there is a loss of coercivity on the kernel as the mesh is refined.

Assembling on any given mesh yields a linear system with an invertible square matrix, but the resulting sequence of approximations does not converge to the exact solution as the mesh is refined.

Conclusion

The $\mathrm{CG}_2 \times \mathrm{DG}_0$ discretisation is unstable: loss of coercivity on the kernel.

(A) $\mathrm{CG}_1 \times \mathrm{CG}_1$ ✗ (does not satisfy inf-sup)

(B) $\mathrm{CG}_1 \times \mathrm{DG}_0$ ✓

(C) $\mathrm{CG}_2 \times \mathrm{DG}_0$ ✗ (coercivity on the kernel)