Advanced Numerical Analysis

Introduction to FEM

Author: Ziyi Wang

Institute: UESTC SMS

Date: February 8, 2026

Version: 1.0

Bio: Information

电子科技大学 数学科学学院
School of Mathematical Sciences UESTC

Stay hungry, stay foolish.

Contents

Chapter 1 Ritz-Galerkin Methods for Second-Order ODEs in 1D

1.1 Model Problem and Weak Formulation

1.1.1 The Boundary Value Problem

A two-point Boundary Value Problem (BVP) is often used to describe various physical models, such as:

- The deformation of an elastic bar under load.
- Heat conduction in a rod.
- The displacement of an elastic cord.

Let us consider the following model problem for an elastic cord:

$$\begin{cases} -u''(x) = f(x), \quad x \in (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{1.1}$$

We refer to (1.1) as the **strong form** of the problem.

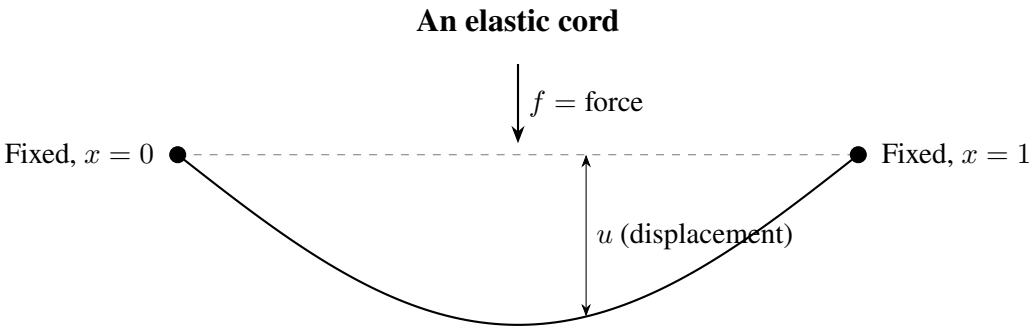


Figure 1.1: Deformation of a Fixed-Fixed Elastic Cord

1.1.2 Derivation of the Weak Formulation

Suppose u is the classical solution and v is any smooth "test function" satisfying the boundary conditions $v(0) = v(1) = 0$. Multiplying the equation by v and integrating over $(0,1)$, integration by parts yields:

$$\begin{aligned} (f, v) &:= \int_0^1 f(x)v(x)\,\mathrm{d}x \\ &= \int_0^1 -u''(x)v(x)\,\mathrm{d}x \\ &= \left[-u'(x)v(x)\right]_0^1 + \int_0^1 u'(x)v'(x)\,\mathrm{d}x \\ &= \int_0^1 u'(x)v'(x)\,\mathrm{d}x =: a(u,v), \end{aligned}$$

where the boundary term vanishes because $v(0) = v(1) = 0$.

Let us formally define the "energy space" (mathematically known as the Sobolev space $H_0^1(0,1)$):

$$V = \left\{ v \in L^2(0,1) : a(v,v) < +\infty \text{ and } v(0) = v(1) = 0 \right\}.$$

Here, $v \in L^2(0,1)$ means v is Lebesgue measurable and square-integrable, i.e., $\int_0^1 |v(x)|^2\,\mathrm{d}x < +\infty$.

Remark We shall formally introduce notations like Sobolev spaces in future lectures. For now, V can be thought of as the space of functions with square-integrable first derivatives satisfying the boundary conditions.

The **weak formulation** (or variational formulation) is stated as follows:

$$\text{Find } u \in V \text{ such that} \quad a(u, v) = (f, v) \quad \text{for all } v \in V. \tag{1.2}$$

A solution u to (1.2) is called a *weak solution*.

Remark If u is a strong solution to (1.1), it implies that u is a solution to (1.2). However, the converse is not necessarily true (pointwise). A weak solution requires less regularity.

Example: Consider the problem:

$$\begin{cases} -u''(x) = f(x) = \dfrac{1}{\sqrt{|x|}}, & x \in (-1, 1), \\ u(-1) = u(1) = 0. \end{cases}$$

We can verify that $u(x) = \frac{4}{3}(1 - |x|^{3/2})$ is a weak solution. However, it is **not** a strong solution in the classical sense because $u''(x)$ blows up at $x = 0$.

1.1.3 Regularity and Variational Principle

If the data f is sufficiently smooth, the weak solution recovers the strong solution.

Theorem 1.1 (Regularity)

Suppose $f \in C^0([0, 1])$ and $u \in C^2([0, 1])$ satisfy Problem (1.2). Then u solves the strong Problem (1.1). ♡

Proof Choose any $v \in V \cap C^0([0, 1])$. Integration by parts on the weak form gives:

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

Thus, rearranging terms:

$$\int_0^1 (f + u'')v \, \mathrm{d}x = 0, \quad \forall v \in V \cap C^0([0, 1]). \tag{1.3}$$

Since $f + u''$ is continuous on $[0, 1]$, suppose for the sake of contradiction that there exists x_0 such that $(f + u'')(x_0) \neq 0$. By continuity, there exists an interval (x_1, x_2) where $f + u''$ maintains the same sign.

Construct a bump function $v \in V$:

$$v(x) = \begin{cases} (x - x_1)^2(x - x_2)^2, & \text{if } x \in (x_1, x_2), \\ 0, & \text{otherwise.} \end{cases}$$

Then, the integral $\int_{x_1}^{x_2} (f + u'')v \, \mathrm{d}x$ is strictly non-zero, contradicting (1.3). Hence, $f(x) + u''(x) = 0$ for all $x \in (0, 1)$.

Why prefer the weak formulation?

1. It requires less regularity (derivatives) of the solution.
2. It is the basis for the Finite Element Method.
3. It is equivalent to a minimization problem (principle of minimum potential energy).

Theorem 1.2 (Variational Principle)

Define the energy functional $F : V \to \mathbb{R}$ as

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

Consider the minimization problem:

$$\text{Find } u \in V \quad \text{s.t.} \quad F(u) \leq F(v) \quad \text{for all } v \in V. \tag{1.4}$$

Then, Problem (1.2) is equivalent to Problem (1.4). ♡

Proof ($\Rightarrow$) Let u solve (1.2). Let $w = v - u \in V$. Then:

$$\begin{aligned} F(v) = F(u+w) &= \frac{1}{2}a(u+w,u+w) - (f,u+w) \\ &= \frac{1}{2}a(u,u) + a(u,w) + \frac{1}{2}a(w,w) - (f,u) - (f,w) \\ &= \underbrace{\left[\frac{1}{2}a(u,u) - (f,u)\right]}_{F(u)} + \underbrace{[a(u,w) - (f,w)]}_{=0} + \underbrace{\frac{1}{2}a(w,w)}_{\geq 0} \\ &\geq F(u). \end{aligned}$$

($\Leftarrow$) Let u be the minimizer of (1.4). For any $v \in V$ and $\varepsilon \in \mathbb{R}$, define $g(\varepsilon) = F(u + \varepsilon v)$. Since u is a minimizer, $g(\varepsilon)$ has a minimum at $\varepsilon = 0$, implying $g'(0) = 0$.

$$\begin{aligned} g(\varepsilon) &= \frac{1}{2}a(u+\varepsilon v, u+\varepsilon v) - (f, u+\varepsilon v) \\ &= \frac{1}{2}a(u,u) + \varepsilon a(u,v) + \frac{1}{2}\varepsilon^2 a(v,v) - (f,u) - \varepsilon(f,v). \end{aligned}$$

The derivative is $g'(\varepsilon) = a(u,v) + \varepsilon a(v,v) - (f,v)$. Setting $g'(0) = 0$ yields:

$$a(u,v) - (f,v) = 0 \quad \Longrightarrow \quad a(u,v) = (f,v).$$

Remark Problem (1.4) corresponds to the "Principle of Minimum Potential Energy in Mechanics."

✍ **Exercise 1.1** Prove that Problem (1.2) has a **unique** solution. (Hint: Use the Lax-Milgram Theorem or strictly convexity of F).

1.2 The Ritz-Galerkin Approximation

Let us replace the infinite-dimensional energy space V with a finite-dimensional subspace $S \subseteq V$. The discrete problem reads:

$$\text{Find } u_S \in S \text{ such that} \quad a(u_S, v) = (f,v) \quad \text{for all } v \in S. \tag{1.5}$$

Here, u_S is the **Ritz-Galerkin approximation** of u.

Let $\mathcal{B} = \{\phi_j\}_{j=1}^n$ be a basis of S. We can expand the solution and test function as:

$$u_S = \sum_{j=1}^{n} u_j \phi_j, \quad v = \sum_{i=1}^{n} v_i \phi_i,$$

where $\{u_j\}$ are the unknown coefficients to be determined. Substituting these into (1.5) and testing with each basis function ϕ_i yields the linear system:

$$\sum_{j=1}^{n} u_j \, a(\phi_j, \phi_i) = (f, \phi_i), \quad i = 1, \ldots, n.$$

In matrix form:

$$AU = F,$$

where

- $U = (u_1, \ldots, u_n)^\top$ is the coefficient vector.
- $A \in \mathbb{R}^{n\times n}$ is the **Stiffness Matrix** with entries $A_{ij} = a(\phi_j, \phi_i)$.
- $F \in \mathbb{R}^n$ is the load vector with entries $F_i = (f, \phi_i)$.

Proposition 1.1

The linear system $AU = F$ admits a unique solution. ♠

Proof The matrix A is symmetric by the symmetry of $a(\cdot,\cdot)$. To show uniqueness, assume $F = 0$ (homogeneous system). This implies:

$$a(u_S, v) = 0, \quad \forall v \in S.$$

Choosing $v = u_S$, we get $a(u_S, u_S) = \int_0^1 (u_S')^2 \,\mathrm{d}x = 0$. This implies $u_S' = 0$ almost everywhere, so u_S is constant. Given the boundary conditions $u_S(0) = u_S(1) = 0$, we must have $u_S \equiv 0$. Thus, $U = 0$, and A is non-singular.

1.3 The Finite Element Method in 1D

In this course, we focus on a specific class of subspaces S constructed using subdivisions of the domain. This approach is called the **Finite Element Method (FEM)**.

Let $0 = x_0 < x_1 < \cdots < x_n = 1$ be a partition (mesh) of $[0, 1]$. Define the mesh size $h_i = x_i - x_{i-1}$ and $h = \max_i h_i$. Let S_h be the space of continuous, piecewise linear functions subordinate to this partition:

$$S_h = \{v \in C^0([0,1]) : v|_{[x_{i-1},x_i]} \in \mathbb{P}^1, \forall i, \text{ and } v(0) = v(1) = 0\}.$$

Exercise: Verify that S_h is a subspace of V.

We choose the **nodal basis** (hat functions) $\{\phi_i\}_{i=1}^{n-1}$ defined by:

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

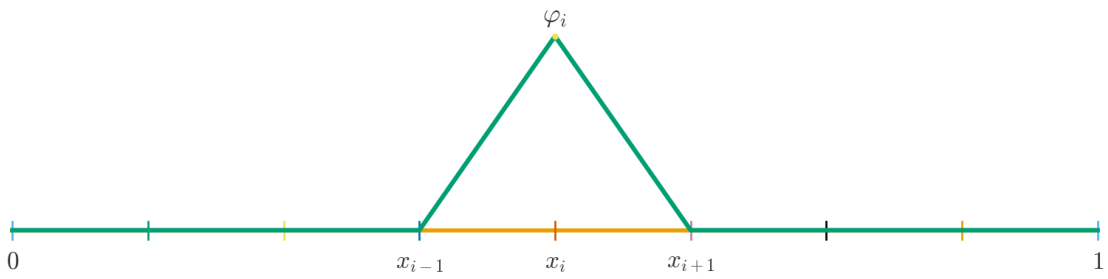


Figure 1.2: Illustration of a 1D linear basis function (Hat function).

Definition 1.1

Define the ***Interpolation Operator*** $\mathcal{I} : C^0([0,1]) \to S$ *as:*

$$\mathcal{I}v(x) = \sum_{i=1}^{n-1} v(x_i)\phi_i(x).$$

Note that $\mathcal{I}v(x_i) = v(x_i)$ *at the nodes.*

1.4 A Priori Error Analysis

We now analyze the error between the exact solution u and the discrete solution u_S.

1.4.1 Interpolation Error

Assume $v \in C^2([0,1])$. The error of the linear interpolant is bounded as follows. On an interval $[x_{i-1}, x_i]$, the derivative error is:

$$(\mathcal{I}_h v)'(x) - v'(x) = v''(\eta)(x - \xi),$$

where $\xi \in (x_{i-1}, x_i)$ is a point where the derivatives match (by Mean Value Theorem), and η is an intermediate point. Integrating the square of the error:

$$\int_{x_{i-1}}^{x_i} |(\mathcal{I}_h v)' - v'|^2 \, \mathrm{d}x \leq h^2 \int_{x_{i-1}}^{x_i} |v''|^2 \, \mathrm{d}x.$$

Summing over all elements, we obtain the estimate in the energy norm $\|v\|_E := (\int_0^1 (v')^2 \, \mathrm{d}x)^{1/2}$:

$$\|v - \mathcal{I}_h v\|_E \leq Ch \left(\int_0^1 |v''|^2 \, \mathrm{d}x \right)^{1/2}.$$

1.4.2 Cea's Lemma (Energy Norm Estimate)

Now we are ready to show the approximation error between u and u_s. Let us write the two problem again:

$$a(u, v) = (f, v) \quad \forall v \in V.$$
$$a(u_S, v) = (f, v) \quad \forall v \in S_h.$$

Since $S_h \subset V$, we set $v \in S_h \subset V$ and substruct the above two equations, we have the **Galerkin Orthogonality**:

$$a(u - u_S, v) = 0, \quad \forall v \in S_h.$$

Using this, we can derive the best approximation property (Cea's Lemma):

$$\begin{aligned} \|u - u_S\|_E^2 &= a(u - u_S, u - u_S) \\ &= a(u - u_S, u - v) + \underbrace{a(u - u_S, v - u_S)}_{=0} \\ &\leq \|u - u_S\|_E \|u - v\|_E . \end{aligned}$$

Thus, $\|u - u_S\|_E \leq \inf_{v \in S_h} \|u - v\|_E$. Choosing $v = \mathcal{I}_h u$:

$$\|u - u_S\|_E \leq Ch \left\|u''\right\|_{L^2}.$$

1.4.3 The Duality Argument (L^2 Norm Estimate)

To estimate $\|u - u_S\|_{L^2}$, we use Nitsche's duality trick. Consider the dual problem: find $z \in V$ such that

$$-z'' = u - u_S \quad \text{in } (0,1), \quad z(0) = z(1) = 0.$$

This implies $a(v, z) = (u - u_S, v)$ for all $v \in V$. Regularity theory gives $\|z''\|_{L^2} \le \|u - u_S\|_{L^2}$.

Testing with $v = u - u_S$:

$$\begin{aligned}
\|u - u_S\|_{L^2}^2 &= (u - u_S, u - u_S) = a(u - u_S, z) \\
&= a(u - u_S, z - \mathcal{I}_h z) \quad \text{(by orthogonality)} \\
&\le \|u - u_S\|_E \|z - \mathcal{I}_h z\|_E \\
&\le (Ch \|u''\|_{L^2}) \cdot (Ch \|z''\|_{L^2}) \\
&\le Ch^2 \|u''\|_{L^2} \|u - u_S\|_{L^2}.
\end{aligned}$$

Dividing by $\|u - u_S\|_{L^2}$ gives the quadratic convergence rate:

Theorem 1.3

If $u \in H^2(0,1)$, then:

$$\|u - u_S\|_{L^2} + h \|u - u_S\|_E \le Ch^2 \|u''\|_{L^2}.$$

♡

Remark (other finite dimensional subspace S_h)

1. $\phi_j(x) = \sin(j\pi x), j = 1, \cdots, n. \quad x \in (0,1).$
2. $\phi_j(x) = x^j(1-x), \quad j = 1, \cdots, n \quad x \in (0,1).$

1.5 Implementation and Comparison

1.5.1 Stiffness Matrix Assembly

Let us try to compute the discrete system using the finite element space. Note that

$$A_{ij} = a \int_0^1 \phi_j' \phi_i' \, \mathrm{d}x \quad F_i = \int_0^1 f \phi_i \, \mathrm{d}x.$$

for $i, j = 1, \cdots, n-1$.

The shape functions are

$$\phi_j(x) = \begin{cases} \frac{x - x_{j-1}}{x_j - x_{j-1}} & x \in [x_{j-1}, x_j] \\ \frac{x_{j+1} - x}{x_{j+1} - x_j} & x \in [x_j, x_{j+1}] \\ 0, & \text{otherwise.} \end{cases}$$

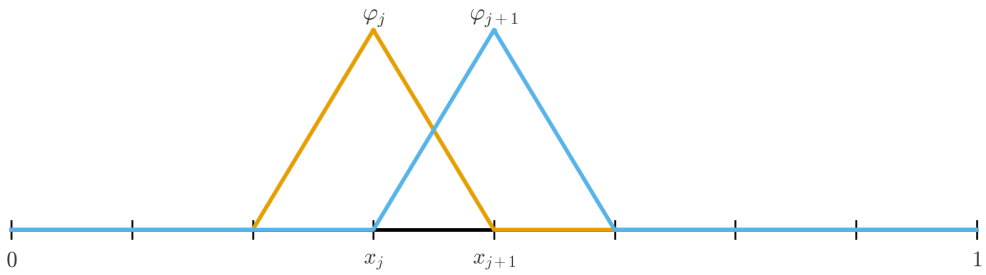


Figure 1.3: Basis functions for the linear hat functions.

For the linear hat functions, the derivative is piecewise constant:

$$\phi_j'(x) = \begin{cases} \frac{1}{h_j}, & x \in [x_{j-1}, x_j] \\ -\frac{1}{h_{j+1}} & x \in [x_j, x_{j+1}] \end{cases}.$$

What's more, if $|i-j| > 1$, then $\int_0^1 \phi_i' \phi_j' dx = 0$. The stiffness matrix entries are calculated as:

$$A_{jj} = \int_0^1 (\phi_j')^2 \, \mathrm{d}x = \frac{1}{h_j} + \frac{1}{h_{j+1}},$$

$$A_{j,j+1} = A_{j+1,j} = \int_0^1 \phi_j' \phi_{j+1}' \, \mathrm{d}x = -\frac{1}{h_{j+1}}.$$

The Matrix form is as follows:

$$A = \begin{pmatrix} \frac{1}{h_1} + \frac{1}{h_2} & -\frac{1}{h_2} & & \\ -\frac{1}{h_2} & \frac{1}{h_2} + \frac{1}{h_3} & -\frac{1}{h_3} & \\ & \ddots & \ddots & \\ & & -\frac{1}{h_n} & \frac{1}{h_{n-1}} + \frac{1}{h_n} \end{pmatrix}$$

The following properties hold:

1. A is symmetric.
2. A is a tri-diagonal matrix.
3. A is positive definite $\Rightarrow$ A is non-singular.

For a uniform mesh with size h:

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & \cdots \\ -1 & 2 & -1 & \cdots \\ 0 & -1 & 2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

1.5.2 Comparison with Finite Difference Method

The standard central difference approximation for $-u'' = f$ is:

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j).$$

In the FEM formulation, the j-th row of the system reads:

$$\frac{1}{h}(-u_{j-1} + 2u_j - u_{j+1}) = (f, \phi_j).$$

If we approximate the load integral using the trapezoidal rule, $(f, \phi_j) \approx h f(x_j)$, the FEM system becomes identical to the Finite Difference system.

Chapter 2 Sobolev Spaces

2.1 Sobolev Spaces

Let $\Omega \subset \mathbb{R}^n$ $(n \geq 1)$ be a bounded open domain.

2.1.1 $L^p(\Omega)$ spaces

Definition 2.1

A (real-valued) measurable function $f : \Omega \to \mathbb{R}$ belongs to $L^p(\Omega)$ if

$$\int_\Omega |f(x)|^p \, dx < \infty \quad (1 \leq p < \infty).$$

For $p = \infty$ we define

$$L^\infty(\Omega) = \Big\{ f : \Omega \to \mathbb{R} \text{ measurable} : \operatorname*{ess\,sup}_{x \in \Omega} |f(x)| < \infty \Big\}.$$

♣

For $1 \leq p < \infty$ we set

$$\|f\|_{L^p(\Omega)} = \Big(\int_\Omega |f(x)|^p \, dx \Big)^{1/p},$$

and for $p = \infty$ we set

$$\|f\|_{L^\infty(\Omega)} = \operatorname*{ess\,sup}_{x \in \Omega} |f(x)|.$$

If $\overline{\Omega}$ is compact and $f \in C(\overline{\Omega})$, then

$$\|f\|_{L^\infty(\Omega)} = \max_{x \in \overline{\Omega}} |f(x)|.$$

Theorem 2.1 (Minkowski inequality)

Let $1 \leq p \leq \infty$ and $f, g \in L^p(\Omega)$. Then

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

More generally, for $f_k \in L^p(\Omega)$, $k = 1, \ldots, N$,

$$\Big\| \sum_{k=1}^N f_k \Big\|_{L^p(\Omega)} \leq \sum_{k=1}^N \|f_k\|_{L^p(\Omega)}.$$

♡

Remark If $\int_\Omega |f|^p \, dx = 0$, then $f = 0$ almost everywhere in Ω. Hence $\| \cdot \|_{L^p(\Omega)}$ is indeed a norm on $L^p(\Omega)$.

Theorem 2.2 (Hölder inequality)

Let $1 < p, q < \infty$ with $1/p + 1/q = 1$ and $f \in L^p(\Omega)$, $g \in L^q(\Omega)$. Then

$$\int_\Omega |f(x) g(x)| \, dx \leq \|f\|_{L^p(\Omega)} \, \|g\|_{L^q(\Omega)}.$$

♡

For $p = q = 2$, Hölder inequality becomes the Cauchy–Schwarz inequality

$$\Big| \int_\Omega f(x) g(x) \, dx \Big| \leq \|f\|_{L^2(\Omega)} \, \|g\|_{L^2(\Omega)}.$$

Example 2.1

$$\left(\int_\Omega f(x)\cdot dx\right)^2 \leqslant \left(\int_\Omega |1\cdot f(x)|dx\right)^2$$
$$\leqslant \int_\Omega 1dx \int_\Omega |f(x)|^2 dx \leqslant \mu(\Omega)\int_\Omega |f|^2 dx$$

Definition 2.2

For $f, g \in L^2(\Omega)$ we define

$$(f,g)_{L^2(\Omega)} := \int_\Omega f(x)g(x)\,dx.$$

Then $(\cdot,\cdot)_{L^2(\Omega)}$ is an inner product on $L^2(\Omega)$.

Theorem 2.3

For $1 \leq p \leq \infty$, $L^p(\Omega)$ is complete with respect to the norm $\|\cdot\|_{L^p(\Omega)}$, i.e. $L^p(\Omega)$ is a Banach space. namely every Cauchy sequence (w.r.t. $\|\cdot\|_{L^p}$) in $L^p(\Omega)$, say $\{f_n\} \subset L^p(\Omega)$, converges to a function (unique) $f \in L^p(\Omega)$, i.e.

$$\lim_{n\to\infty} \|f - f_n\|_{L^p} = 0$$

For $p = 2$, $L^2(\Omega)$ equipped with the inner product $(\cdot,\cdot)_{L^2(\Omega)}$ is a Hilbert space.

We denote the Lebesgue measure of a measurable set $E \subset \Omega$ by $\mu(E)$.

2.1.2 Generalized or weak derivatives

Let $\Omega \subset \mathbb{R}^d$ and let $\alpha = (\alpha_1, \ldots, \alpha_d)$ be a multi-index with $|\alpha| = \alpha_1 + \cdots + \alpha_d$.

Definition 2.3

For a sufficiently smooth function u we define

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}},$$

and write $D^0 u = u$.

Definition 2.4

We denote by $C_c^\infty(\Omega)$ the space of infinitely differentiable functions $\varphi : \mathbb{R}^d \to \mathbb{R}$ whose support is contained in Ω.

Definition 2.5

For $\varphi \in C_c^\infty(\Omega)$ the support of φ is defined by

$$\operatorname{supp}(\varphi) := \overline{\{x \in \Omega : \varphi(x) \neq 0\}} \subset \overline{\Omega}.$$

Definition 2.6 (L^1_{loc})

A measurable function f is said to belong to $L^1_{\text{loc}}(\Omega)$ if

$$\int_K |f(x)|\,dx < \infty \quad \text{for every compact set } K \subset \Omega.$$

Example 2.2 The function $f(x) = \dfrac{1}{x}$ on $\Omega = (0,1)$ satisfies $f \in L^1_{\text{loc}}(0,1)$ but $f \notin L^1(0,1)$ since

$$\int_a^1 \frac{1}{x}\,dx < \infty \quad \text{for all } a > 0, \qquad \int_0^1 \frac{1}{x}\,dx = \infty.$$

✍ **Exercise 2.1** For what value of α Satisfying $x^{-\alpha} \in L^p(0,1)$?

Definition 2.7 (Weak derivative)

Let $f \in L^1_{\text{loc}}(\Omega)$ and let α be a multi-index. A function $g \in L^1_{\text{loc}}(\Omega)$ is called the weak derivative *of f of order α, written $g = D^\alpha f$, if*

$$\int_\Omega f(x) D^\alpha \varphi(x)\,dx = (-1)^{|\alpha|} \int_\Omega g(x)\varphi(x)\,dx \quad \textit{for all } \varphi \in C_c^\infty(\Omega).$$

♣

Examples in one dimension

Let $\Omega = \mathbb{R}$ for simplicity.

Example 2.3Smooth case If $f \in C^1(\mathbb{R})$ with compact support, then integration by parts shows

$$\int_{\mathbb{R}} f(x)\varphi'(x)\,dx = -\int_{\mathbb{R}} f'(x)\varphi(x)\,dx \quad \forall\, \varphi \in C_c^\infty(\mathbb{R}),$$

so the usual derivative f' is also the weak derivative.

Example 2.4A continuous, piecewise smooth function Define

$$g(x) = \begin{cases} 0, & x \le 0, \\ x, & 0 < x < 1, \\ 1, & x \ge 1. \end{cases}$$

Then g is continuous and piecewise C^1. A direct computation using integration by parts on $(0,1)$ shows that

$$\int_{\mathbb{R}} g(x)\varphi'(x)\,dx = -\int_0^1 \varphi(x)\,dx \quad \forall\, \varphi \in C_c^\infty(\mathbb{R}).$$

Hence the weak derivative of g is

$$g'(x) = \chi_{(0,1)}(x),$$

the characteristic function of the open interval $(0,1)$.

Example 2.5Absolute value Let $f(x) = |x|$ on $\mathbb{R}$. For $\varphi \in C_c^\infty(\mathbb{R})$ we have

$$\int_{\mathbb{R}} |x|\varphi'(x)\,dx = \int_0^\infty x\varphi'(x)\,dx - \int_{-\infty}^0 x\varphi'(x)\,dx = -\int_0^\infty \varphi(x)\,dx + \int_{-\infty}^0 \varphi(x)\,dx = -\int_{\mathbb{R}} \operatorname{sgn}(x)\varphi(x)\,dx.$$

Thus the weak derivative of $|x|$ is

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

which belongs to $L^1_{\text{loc}}(\mathbb{R})$ although the classical derivative does not exist at $x = 0$.

Example 2.6A step function without weak derivative Let

$$h(x) = \begin{cases} 0, & x < 0, \\ 1, & x \ge 0. \end{cases}$$

Then for $\varphi \in C_c^\infty(\mathbb{R})$,

$$\int_{\mathbb{R}} h(x)\varphi'(x)\,dx = -\varphi(0).$$

If a weak derivative $g \in L^1_{\text{loc}}(\mathbb{R})$ of h existed, we would have

$$-\varphi(0) = \int_{\mathbb{R}} h(x)\varphi'(x)\,dx = -\int_{\mathbb{R}} g(x)\varphi(x)\,dx \quad \forall\, \varphi \in C_c^\infty(\mathbb{R}).$$

Using a sequence of mollifiers $\{\varphi_\varepsilon\}$ that approximate the Dirac mass at 0 and the dominated convergence theorem leads to a contradiction. Hence h has no weak derivative in L^1_{loc}.

2.1.3 Sobolev spaces $W^{k,p}(\Omega)$

Definition 2.8

Let $k \in \mathbb{N}$ and $1 \le p \le \infty$. The Sobolev space $W^{k,p}(\Omega)$ is defined by

$$W^{k,p}(\Omega) = \Big\{ u \in L^p(\Omega) : D^\alpha u \in L^p(\Omega) \text{ for all multi-indices } \alpha \text{ with } |\alpha| \le k \Big\}.$$

♣

If $d = 1$, and $k = 1, \Omega = (a, b)$, $W^{1,p}(\Omega) = \{f \in L^p(\Omega) \mid f' \text{ exists and } f' \in L^p(\Omega)\}$.

Remark If $u \in C^k(\overline{\Omega})$, then the weak derivatives $D^\alpha u$ coincide with the classical derivatives; hence $C^k(\overline{\Omega}) \subset W^{k,p}(\Omega)$.

Definition 2.9 (Norm and seminorm)

For $1 \le p < \infty$ we define

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

For $p = \infty$ we use

$$\|u\|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| \le k} \|D^\alpha u\|_{L^\infty(\Omega)}, \qquad |u|_{W^{k,\infty}(\Omega)} := \max_{|\alpha| = k} \|D^\alpha u\|_{L^\infty(\Omega)}.$$

♣

Theorem 2.4

For each $k \in \mathbb{N}$ and $1 \le p \le \infty$, the space $W^{k,p}(\Omega)$ equipped with $\|\cdot\|_{W^{k,p}(\Omega)}$ is a Banach space.

♡

For $p = 2$ we write

$$H^k(\Omega) := W^{k,2}(\Omega).$$

Definition 2.10

For $u, v \in H^k(\Omega)$ we define the inner product

$$(u, v)_{H^k(\Omega)} := \sum_{|\alpha| \le k} \big(D^\alpha u, D^\alpha v\big)_{L^2(\Omega)}.$$

Then $H^k(\Omega)$ is a Hilbert space with this inner product.

♣

Example 2.7 $f \in H^1(\Omega) \Rightarrow f \in L^2(\Omega)$ and $\frac{\partial}{\partial x_i} f \in L^2(\Omega), i = 1, 2, \cdots, d.$

$$\|f\|_{H^1(\Omega)} = \big(\|f\|^2_{L^2} + \||\nabla f|\|^2_{L^2}\big)^{1/2}$$

$$\|\nabla f\|^2_{L^2} = \int_\Omega |\nabla f|^2 dx$$

2.1.4 A one-dimensional example and hat functions

Let $\Omega = (a, b) \subset \mathbb{R}$ and choose a partition

$$a = x_0 < x_1 < \cdots < x_N = b.$$

Define

$$\mathcal{M} = \left\{ v \in C([a,b]) : v\big|_{(x_{i-1},x_i)} \text{ is linear for } i = 1, \ldots, N \right\}.$$

For each node x_j $(j = 0, \ldots, N)$ we define the *hat function* $\varphi_j \in \mathcal{M}$ by

$$\varphi_j(x_i) = \delta_{ij}, \qquad i = 0, \ldots, N,$$

and on each interval (x_{i-1}, x_i) we let φ_j be the unique linear function interpolating these node values. Explicitly, for $j = 1, \ldots, N-1$,

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}}, & x \in [x_{j-1}, x_j], \\ \dfrac{x_{j+1} - x}{x_{j+1} - x_j}, & x \in [x_j, x_{j+1}], \\ 0, & \text{otherwise}, \end{cases}$$

and similar formulas hold for φ_0 and φ_N.

The derivatives φ_j' exist almost everywhere and are piecewise constant, hence $\varphi_j \in H^1(a,b)$ for all j. One can show that $\{\varphi_j\}_{j=0}^N$ is a basis of $\mathcal{M}$.

✍ **Exercise 2.2** The function ϕ_j has a weak derivative which equals $\phi_j'(x)$ for $x \in (x_l, x_{l+1}), l = 0, 1, \cdots, n-1. \phi_j \in H^1(a,b)$.

Remark In fact $\phi_j \in W^{1,p}(a,b), p \in [1, +\infty]. \Rightarrow \mathcal{M} \subset W^{1,p}(\Omega), p \in [1, +\infty]$.

2.1.5 Domains with continuous / Lipschitz boundary

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain and $x^0 \in \partial\Omega$.

Definition 2.11 (Continuous boundary)

We say that $\partial\Omega$ is continuous *(of class C^0) if for each $x^0 \in \partial\Omega$ there exist a rectangle $V \subset \mathbb{R}^2$ containing x^0, a new Cartesian coordinate system (x_1, x_2) with origin at x^0, and a continuous function $\varphi : \mathbb{R} \to \mathbb{R}$ such that*

$$\Omega \cap V = \{y \in V : y_n > \psi(y')\}, \qquad \partial\Omega \cap V = \{y \in V : y_n = \psi(y')\}.$$

♣

Definition 2.12 (Lipschitz boundary)

We say that $\partial\Omega$ is Lipschitz *if in the above definition the function ψ can be chosen to be Lipschitz continuous: there exists $L > 0$ such that*

$$|\psi(y_1') - \psi(y_2')| \le L\,|y_1' - y_2'| \quad \text{for all } y_1', y_2'.$$

♣

A typical example of a domain without continuous boundary is a "slit domain", e.g. a square with a line segment removed from its interior.

Theorem 2.5 (Density of smooth functions)

If Ω has a Lipschitz boundary, then $C^\infty(\overline{\Omega})$ is dense in $W^{1,p}(\Omega)$ for $1 \le p < \infty$, i.e. for every

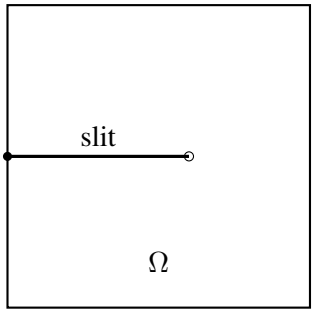


Figure 2.1: A slit domain: a square with a line segment removed from its interior.

$u \in W^{1,p}(\Omega)$ *there exists a sequence* $\{u_k\} \subset C^\infty(\overline{\Omega})$ *such that*

$$\|u_k - u\|_{W^{1,p}(\Omega)} \to 0 \quad \text{as } k \to \infty.$$

♡

2.2 Integration by parts

2.2.1 One-dimensional case

By the fundamental theorem of calculus, if $f \in C^1([a,b])$, then

$$f(b) - f(a) = \int_a^b f'(s)\,ds = f\big|_a^b.$$

Apply this to $f = uv$, where $u, v \in C^1([a,b])$. Then

$$uv\big|_a^b = \int_a^b (uv)'(s)\,ds = \int_a^b u'(s)v(s)\,ds + \int_a^b u(s)v'(s)\,ds.$$

Hence the one-dimensional integration-by-parts formula

$$\int_a^b u'(s)v(s)\,ds = uv\big|_a^b - \int_a^b v'(s)u(s)\,ds.$$

2.2.2 Gauss theorem in higher dimensions

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and $f : \Omega \to \mathbb{R}^d$ a vector field. Gauss (divergence) theorem states that

$$\int_{\partial\Omega} f \cdot n\,dA = \int_\Omega \nabla\cdot f\,dx,$$

where $f \in C^1(\overline{\Omega})^d$ and

$$\nabla\cdot f = \sum_{i=1}^d \frac{\partial f_i}{\partial x_i}.$$

The quantity

$$\int_{\partial\Omega} 1\,dA$$

is the $(d-1)$–dimensional measure (surface area) of $\partial\Omega$.

Assume that $\partial\Omega$ is of class C^1, namely that $\partial\Omega$ is the union of a finite number of patches U_i, $i = 1, \dots, k$, each of which can be mapped via an invertible C^1 function to a subset of $\mathbb{R}^{d-1}$.

2.2.3 Integration by parts in higher dimensions

Let $u : \Omega \to \mathbb{R}^d$ be a vector field with $u \in C^1(\overline{\Omega})^d$, and let $v : \Omega \to \mathbb{R}$ be a scalar function with $v \in C^1(\overline{\Omega})$. Apply Gauss theorem to the vector field

$$uv := (u_1 v, \dots, u_d v) : \Omega \to \mathbb{R}^d.$$

We have

$$\int_{\partial\Omega} v\, u \cdot n \, dA = \int_{\Omega} \nabla\cdot (uv)\, dx.$$

Using the product rule,

$$\nabla\cdot (uv) = (\nabla\cdot u)\, v + u \cdot \nabla v,$$

we obtain the integration–by–parts formula

$$\int_{\Omega} u \cdot \nabla v \, dx = \int_{\partial\Omega} v(u \cdot n)\, dA - \int_{\Omega} (\nabla\cdot u)\, v \, dx. \tag{2.1}$$

Now choose a special vector field whose only nonzero component is the jth one:

$$F(x) = (0, \dots, 0, u(x), 0, \dots, 0),$$

so that $(\nabla\cdot F)(x) = \dfrac{\partial u}{\partial x_j}(x)$ and $F \cdot \nabla v = u\, \dfrac{\partial v}{\partial x_j}$. Applying (2.1) with F in place of u gives

$$\int_{\Omega} u\, \frac{\partial v}{\partial x_j}\, dx = \int_{\partial\Omega} v\, u\, n_j \, dA - \int_{\Omega} \frac{\partial u}{\partial x_j}\, v \, dx. \tag{2.2}$$

2.2.4 Smooth functions and weak derivatives

Let α be a multi-index with $|\alpha| \le k$.

Theorem 2.6

If $u \in C^k(\overline{\Omega})$, then the weak derivative $D^\alpha u$ exists in $L^1_{loc}(\Omega)$ and coincides with the classical (strong) derivative. ♡

Proof Let $\varphi \in C_c^\infty(\Omega)$ be a test function and assume $|\alpha| \ge 1$ (the case $|\alpha| = 0$ is trivial). Write $\alpha = \beta + e_j$ with $|\beta| = |\alpha| - 1$ and e_j the jth coordinate unit vector. Then

$$(-1)^{|\alpha|} \int_{\Omega} u\, D^\alpha \varphi \, dx = (-1)^{|\alpha|} \int_{\Omega} u\, \frac{\partial}{\partial x_j} D^\beta \varphi \, dx.$$

Apply the one-coordinate integration-by-parts formula (2.2) with $v = D^\beta \varphi$. Since φ has compact support in Ω, the boundary term vanishes and we get

$$(-1)^{|\alpha|} \int_{\Omega} u\, D^\alpha \varphi \, dx = (-1)^{|\alpha|-1} \int_{\Omega} \frac{\partial u}{\partial x_j} D^\beta \varphi \, dx.$$

Repeating this procedure $|\alpha|$ times (moving one derivative at a time from φ to u), we obtain

$$(-1)^{|\alpha|} \int_{\Omega} u\, D^\alpha \varphi \, dx = \int_{\Omega} D^\alpha u\, \varphi \, dx.$$

Thus the classical derivative $D^\alpha u$ is also the weak derivative. Since $u \in C^k(\overline{\Omega})$, we have $D^\alpha u \in C(\overline{\Omega})$ and in particular $D^\alpha u \in L^\infty(\Omega)$.

2.3 Introduction: Trace on a Square

Assume Ω has a continuous boundary so that $C^\infty(\overline{\Omega})$ is dense in $H^1(\Omega)$.

2.3.1 Special Case: $\Omega = (0,1)^2$

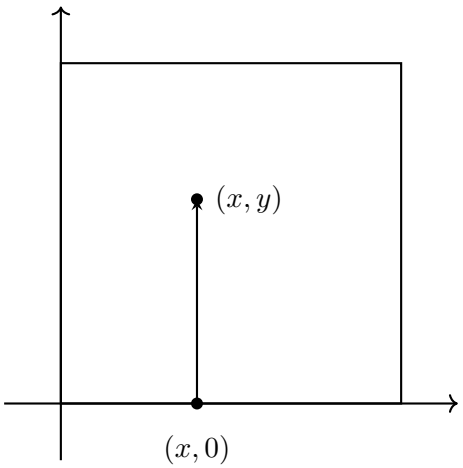


Figure 2.2: Special case of the trace theorem

We start with 1D integration by parts (Fundamental Theorem of Calculus). Let $u \in C^1(\overline{\Omega})$. Consider the boundary at $y = 0$. By the Fundamental Theorem of Calculus:

$$u(x,0) = -\int_0^y u_s(x,s)\,ds + u(x,y).$$

Recall the inequality $(a+b)^2 \le 2(a^2+b^2)$. Applying this:

$$|u(x,0)|^2 \le 2\left(\left|\int_0^y u_s(x,s)\,ds\right|^2 + |u(x,y)|^2\right).$$

Using Cauchy-Schwarz inequality on the integral term:

$$\begin{aligned}\left|\int_0^y u_s(x,s)\,ds\right|^2 &\le \left(\int_0^y 1^2\,ds\right)\left(\int_0^y |u_s(x,s)|^2\,ds\right)\\ &= y\int_0^y |u_s(x,s)|^2\,ds\\ &\le 1\cdot\int_0^1 |u_s(x,s)|^2\,ds \quad (\text{since } y \le 1).\end{aligned}$$

Substituting this back, we get:

$$|u(x,0)|^2 \le 2\left(\int_0^1 |u_s(x,s)|^2\,ds + |u(x,y)|^2\right).$$

Now, integrate with respect to y over $[0,1]$:

$$\begin{aligned}\int_0^1 |u(x,0)|^2\,dy &\le \int_0^1 2\left(\|u_s(x,\cdot)\|^2_{L^2(0,1)} + |u(x,y)|^2\right)dy\\ |u(x,0)|^2\cdot 1 &\le 2\|u_y(x,\cdot)\|^2_{L^2(0,1)} + 2\int_0^1 |u(x,y)|^2\,dy.\end{aligned}$$

Next, integrate with respect to x over $[0,1]$:

$$\begin{aligned}\int_0^1 |u(x,0)|^2\,dx &\le 2\int_0^1\int_0^1 |u_y(x,y)|^2\,dy\,dx + 2\int_0^1\int_0^1 |u(x,y)|^2\,dy\,dx\\ \|u(\cdot,0)\|^2_{L^2(0,1)} &\le 2\left(\|u_y\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)}\right)\\ &\le 2\|u\|^2_{H^1(\Omega)}.\end{aligned}$$

Repeating this argument for each boundary segment (4 sides), we get:

$$\|u\|^2_{L^2(\partial\Omega)} \leq 8\|u\|^2_{H^1(\Omega)}.$$

2.3.2 Density Argument

Since $C^\infty(\overline{\Omega})$ is dense in $H^1(\Omega)$, given $w \in H^1(\Omega)$, there exists a sequence $\{u_n\} \subset C^\infty(\overline{\Omega})$ such that $u_n \to w$ in $H^1(\Omega)$. Then $\{u_n|_{\partial\Omega}\}$ is Cauchy in $L^2(\partial\Omega)$:

$$\|u_n - u_m\|_{L^2(\partial\Omega)} \leq \sqrt{8}\|u_n - u_m\|_{H^1(\Omega)} \to 0. \tag{2.3}$$

Thus, $u_n|_{\partial\Omega}$ has a limit in $L^2(\partial\Omega)$, which we call the **trace** of w on $\partial\Omega$, denoted by $\gamma w = w|_{\partial\Omega}$.

2.4 General Case: Continuous Boundary

Theorem 2.7

Assume that Ω has a continuous boundary. The map $u \mapsto \gamma(u) = u|_{\partial\Omega}$ extends continuously from $W^{1,p}(\Omega)$ to $L^p(\partial\Omega)$ for $p \in [1, \infty)$. ♡

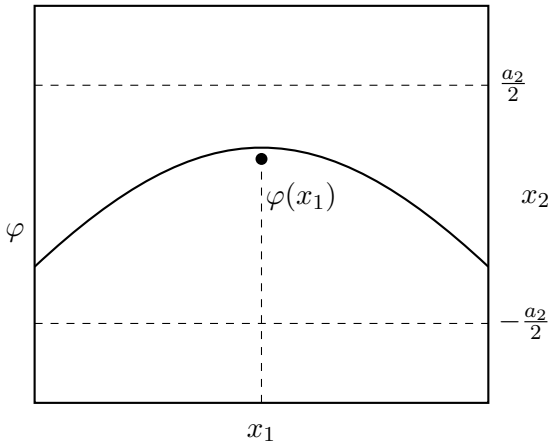


Figure 2.3: Generalized Boundary Patch

Proof **Setup:** Since $\partial\Omega$ is compact, we may reduce to a finite set of patches $V_1, \ldots, V_N$ covering $\partial\Omega$. We use a local coordinate system. Let the boundary be locally described by a graph $x_2 = \varphi(x_1)$. Let the domain within the patch be defined by $-a_1 \leq x_1 \leq a_1$ and $-a_2 \leq x_2 \leq \varphi(x_1)$.

For $u \in C^1(\overline{\Omega})$, using the Fundamental Theorem of Calculus along the x_2 direction:

$$u(x_1, \varphi(x_1)) = \int_{x_2}^{\varphi(x_1)} u_s(x_1, s)\, ds + u(x_1, x_2). \tag{2.4}$$

Taking the absolute value:

$$|u(x_1, \varphi(x_1))| \leq \int_{x_2}^{\varphi(x_1)} |u_s(x_1, s)|\, ds + |u(x_1, x_2)|.$$

Using the inequality $(a+b)^p \leq 2^p(a^p + b^p)$ (for $p \geq 1$):

$$|u(x_1, \varphi(x_1))|^p \leq 2^p \left(\left| \int_{x_2}^{\varphi(x_1)} u_s(x_1, s)\, ds \right|^p + |u(x_1, x_2)|^p \right).$$

Apply Hölder's inequality to the integral term. Let q be the conjugate exponent of p ($1/p + 1/q = 1$).

$$\left| \int_{x_2}^{\varphi(x_1)} 1 \cdot u_s \, ds \right| \leq \|1\|_{L^q(x_2,\varphi_1(x_1))} \|u_s(x_1, s)\|_{L^p(x_2,\varphi_1(x_1))}$$

And

$$\|1\|_{L^q(x_2,\varphi(x_1))} = \begin{cases} 1 & \text{if } q = \infty \\ (\varphi(x_1) - x_2)^{\frac{1}{q}} \leqslant \left(\frac{3}{2}a_2\right)^{\frac{1}{q}} \leqslant C. \end{cases}$$

Note that the length is bounded by the size of the patch, say C. Thus,

$$|u(x_1, \varphi(x_1))|^p \leq C \left(\int_{-a_2}^{\varphi(x_1)} |u_s(x_1, s)|^p \, ds + |u(x_1, x_2)|^p \right).$$

Integration: Now integrate with respect to x_2 over the vertical segment $[-a_2, \varphi(x_1)]$. Note that the Left Hand Side (LHS), $|u(x_1, \varphi(x_1))|^p$, does not depend on x_2.

$$\begin{aligned} \text{Integrated LHS} &= \int_{-a_2}^{\varphi(x_1)} |u(x_1, \varphi(x_1))|^p \, dx_2 \\ &= (\varphi(x_1) - (-a_2))|u(x_1, \varphi(x_1))|^p \\ &\geq \frac{a_2}{2}|u(x_1, \varphi(x_1))|^p \quad \text{(assuming patch geometry is bounded away from 0).} \end{aligned}$$

Now integrate with respect to x_1 over $[-a_1, a_1]$:

$$\int_{-a_1}^{a_1} \frac{a_2}{2}|u(x_1, \varphi(x_1))|^p \, dx_1 \leq C \int_{-a_1}^{a_1} \int_{-a_2}^{\varphi(x_1)} (|u_s|^p + |u|^p) \, dx_2 \, dx_1.$$

This simplifies to:

$$C_0 \|u\|^p_{L^p(\partial\Omega \cap V)} \leq C \left(\|u_{x_2}\|^p_{L^p(\Omega \cap V)} + \|u\|^p_{L^p(\Omega \cap V)} \right).$$

In standard coordinates, this implies:

$$\|u\|^p_{L^p(\partial\Omega \cap V)} \leq C \|u\|^p_{W^{1,p}(\Omega \cap V)}.$$

Conclusion: Summing up for all finite number of patches covering the boundary, there holds:

$$\|u\|_{L^p(\partial\Omega)} \leq C \|u\|_{W^{1,p}(\Omega)}.$$

Chapter 3 Finite Element Space

3.1 The Finite Element

3.1.1 Admissible Triangulation $(\mathbb{R}^2)$

Assume that Ω is polygonal in 2D or polyhedral in 3D, i.e. $\Omega = \left(\bigcup_{\tau_i \in \mathcal{J}} \tau_i\right)^\circ$

1. τ_i is a (closed) rectangle or triangle. $\left(\Omega \subset \mathbb{R}^2\right)$
2. τ_i is a (closed) brick shaped element on. a tetrahedra ($\Omega \subseteq \mathbb{R}^3$)
3. If τ_i and τ_j are distinct elements in $\mathcal{J}$, then.
   (a). $\tau_i \cap \tau_j = \varnothing$;
   (b). or $\tau_i \cap \tau_j$ is a vertex or an edge of both τ_i and τ_j. $\left(\Omega \subseteq \mathbb{R}^2\right)$
   (c). $\tau_i \cap \tau_j$ is a vertex, edge, face of both τ_i and τ_j. $\left(\Omega \subseteq \mathbb{R}^3\right)$.

Remark Hanging node is not allowed.

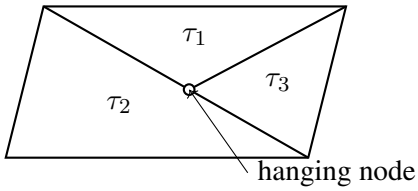


Figure 3.1: Hanging node

Theorem 3.1

Suppose that $\mathcal{J}$ is a triangulation and Ω satisfying $\mu\left(\tau_i \cap \tau_j\right) = 0$ if τ_i and τ_j are distinct. Suppose that $f|_{\tau_i}$ is smooth for each $\tau_i \in \mathcal{J}$. Then $f \in H^k(\Omega)$ iff $f^{(i)}$ is continuous for $j = 0, 1, \cdots, k-1$. ♡

Remark DoF: Degrees of freedom allow for local basis, global basis, interpolation and continuity across element interface.

3.1.2 General finite element

General finite element (K, P, Σ).

- K - region
- P - shape function $\subseteq C^\ell(K)$ with finite dimension m
- $\Sigma = \{\ell_1, \cdots, \ell_m\}$ linear functionals which are unisolvent on P

Definition 3.1 (Unisolvent)

$\{\ell_1, \cdots, \ell_m\}$ linear functionals on P(dimension m) is unisolvent iff the only solution q to $\ell_i(q) = 0 (i = 1, \cdots, m)$ is $q = 0$. ♣

Example 3.1 $K =$ triangle in $\mathbb{R}^2$

$$P = \mathbb{P}^k = \left\{ \sum_{i=0}^{k} \sum_{j=0}^{k-i} c_{ij} x_1^i x_2^j, c_{ij} \in \mathbb{R} \right\}, \quad \dim P = 1 + 2 + \ldots k + 1 = \frac{(k+1)(k+2)}{2}$$

$$\Sigma = \{\ell_{ij}(q) = q(x_{ij}), i = 0, \cdots, k, j = 0, \cdots, k-i, x_{ij} \in K\}.$$

We need to define $\{x_{ij}\}$.

- Case 1: $k = 0, \mathbb{P}^0 =$ constant functions, $\dim\left(\mathbb{P}^0\right) = 1$.

  $$x_{0,0} = \frac{V_1 + V_2 + V_3}{3} (\text{ barycenter of } \tau), \quad \ell_{00}(q) = q(x_{00}).$$

  We cannot use this element to get a nontrivial subspace of $H^1(\Omega)$. But it works for approximation in $L^2(\Omega)$.
- Case 2: $k = 1, \mathbb{P}^1 = \{\alpha + \beta x_1 + \gamma x_2, \alpha, \beta, \gamma \in \mathbb{R}\}$, $\dim\left(\mathbb{P}^1\right) = 3$.

  $$\Sigma = \{\ell_i(q) = q(v_i), i = 1, 2, 3\}.$$

  Question: Is this a unisolvent set of functionals for $\mathbb{P}^1$?

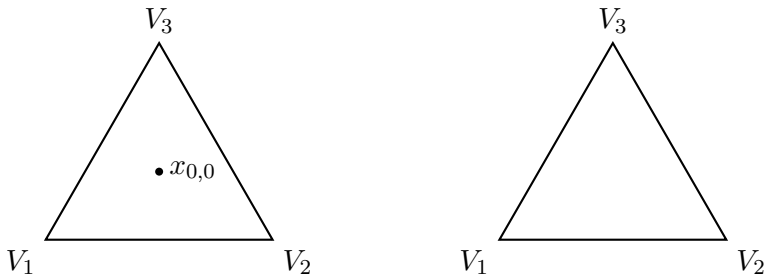


Figure 3.2: finite element for example(3.1)

3.1.3 Reference Element

Lemma 3.1

If $f \in \mathbb{P}^k$, and $\ell(S) = v_1 + (v_2 - v_1)\, s$, then $\ell(0) = v_1, \ell(1) = v_2$. Define $q_k(s) = f(\ell(s))$. Then q_k is a polynomial of degree k. ♡

Proof $\ell(s) = \begin{pmatrix} v_{11} + (v_{21} - v_{11})\, s \\ v_{12} + (v_{22} - v_{12})\, s \end{pmatrix}$

$$f(\ell(s)) = \sum_{0 \le i+j \le k} c_{ij} \left(v_{1,1} + (v_{2,1} - v_{1,1})\, s\right)^i \left(v_{1,2} + (v_{2,2} - v_{1,2})\, s\right)^j$$

This is the polynomial in s of degree k.

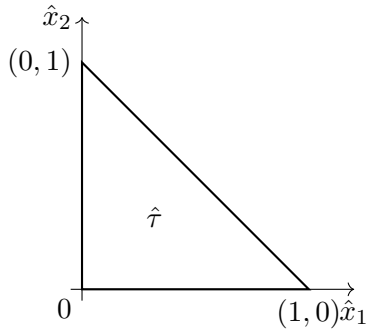


Figure 3.3: Reference element

Now if $p \in \mathbb{P}^1$ satisfies $p(v_i) = 0$ where $i = 0, 1, 2, v_0 = (0,0), v_1 = (1,0), v_2 = (0,1)$.

Restricted to the edge $\hat{x}_2 = 0$ is linear in x_1, which vanishes at 0 and 1 . Therefore, it vanishes on the edge $x_2 = 0$. So $p = \alpha + \beta x_1 + \gamma x_2 = 0$ when $x_2 = 0$. Thus $\alpha = \beta = 0$. Finally, $p(v_2) = 0 = \gamma x_2$, we get $\gamma = 0$ i.e. $p = 0$. we get the unisolvence.

Definition 3.2

An affine mapping of $\mathbb{R}^d$ is one of the form

$$Bx = C_0 + Mx, \quad x \in \mathbb{R}^d.$$

Here $C_0 \in \mathbb{R}^d, M \in \mathbb{R}^{d\times d}$.

♣

Remark The transform is invertible if M is nonsingular.

$$\begin{aligned} B^{-1}y &= M^{-1}(y - C_0) \\ &= -M^{-1}C_0 + M^{-1}y \end{aligned}$$

and B^{-1} is also affine.

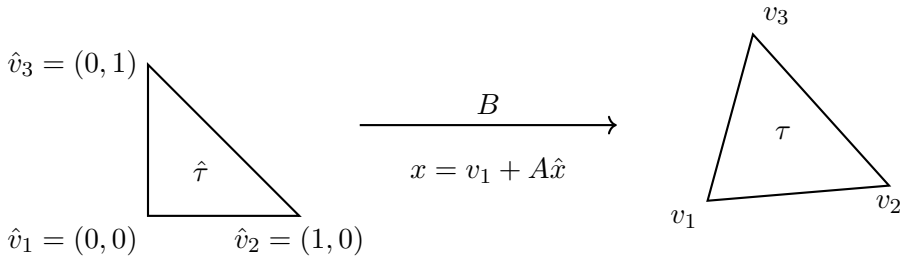


Figure 3.4: Affine mapping

In our case, $x = B\hat{x} = v_1 + \left(v_2 - v_1 \vdots v_3 - v_1\right)\hat{x}$ and there holds $B(\hat{v}_i) = v_i$.

✍ **Exercise 3.1** B is invertible if v_1, v_2, v_3 are not on a line. (or τ has positive area).

Proposition 3.1

B maps $\hat{\tau}$ l:1 onto τ.

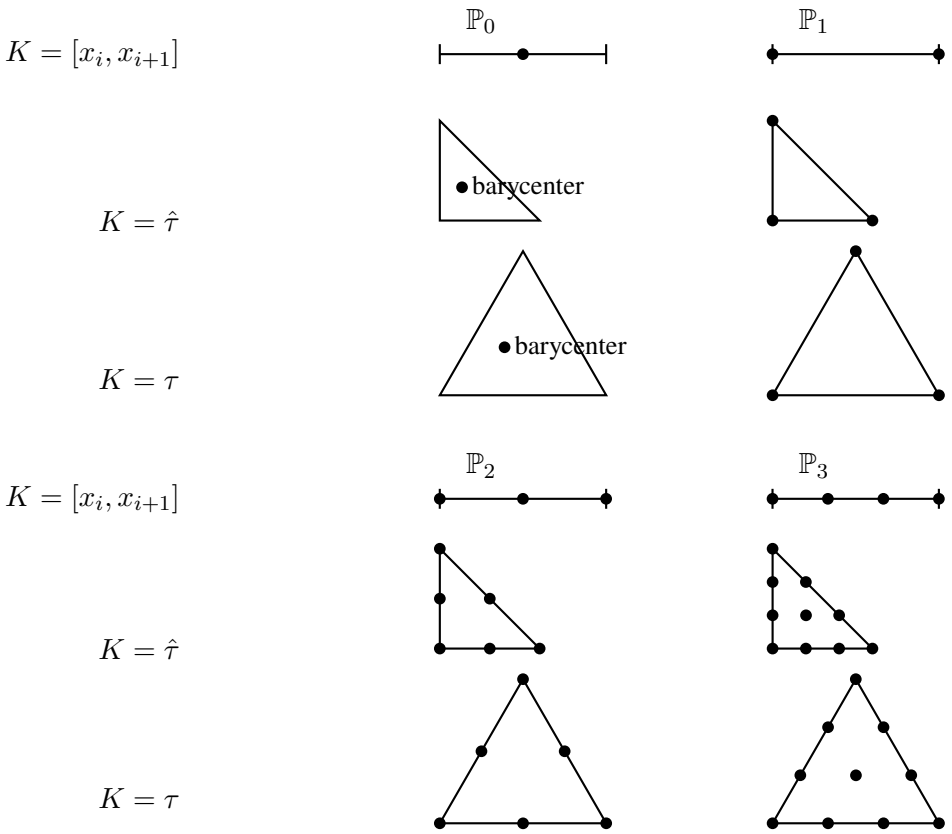
♠

Lemma 3.2

B $1:1$ map from $\mathbb{P}^k(\hat{\tau})$ to $\mathbb{P}^k(\tau)$. (map 0 to 0 one-to-one)

♡

3.2 Lagrange Element



$$\mathbb{P}_4, \mathbb{P}_5 \cdots, \mathbb{P}^k = \left\{ \sum_{0 \le i+j \le k} C_{ij} x_1^i x_2^j \right\},$$

$$\dim \mathbb{P}^k = (k+1) + k + \cdots + 1 = \frac{1}{2}(k+1)(k+2).$$

Proposition 3.2

$\mathbb{P}^k$ Lagrange elements are unisolvent (2D Triangle case).

Proof $B : \hat{\tau} \mapsto \tau$, is an affine and invertible when τ is nondegenerate. B maps the $\left(\mathbb{P}^k\right)$ nodes $\hat{N}_k$ of $\hat{\tau}$ onto N_k of τ. If the nodes of $\hat{N}_k$ are unisolvent, then the nodes of N_k are unisolvent, namely if $p \in \mathbb{P}^k$ and satisfies $P(x_j) = 0, x_j \in N_k$, then $p = 0$. Consider $q(\hat{x}) = p(B(\hat{x}))$, sine $p \mapsto p \circ B$ is a bijection of $\mathbb{P}^k$ onto $\mathbb{P}^k$ and since B maps nodes to nodes and p vanishes on the nodes of N_k, q vanishes on the nodes $\hat{N}_k$. The unisolvence of $\hat{N}_k$ implies that $q = 0$, hence $p = 0$. Therefore. it suffices to verify the result only on the reference triangle. Nest we show by induction on k.

$k = 0, \mathbb{P}^k =$ constants DONE!

Now we assume unisolvence holds for $\mathbb{P}^{k-1}$ in $\hat{\tau}$, we need only show the case on $\mathbb{P}^k$ for $\hat{N}_k$.

Assume $p(x_i) = 0, x_i \in \hat{N}_k$, consider $q = p|_{\hat{x}_2=0} = q(\hat{x}_1)$. So $q(\hat{x}_1) \in \mathbb{P}^k(0,1)$, and q vanishes at $k+1$ nodes. Thus $q = p|_{\hat{x}_2=0} = 0$ (1D case). Let $p = \sum_{0 \le i+j \le k} c_{ij} \hat{x}_1^i \hat{x}_2^j$. Thus

$$p|_{\hat{x}_2=0} = \sum_{i=0}^{k} c_{i0} \hat{x}_1^i = 0 \Rightarrow c_{i0} = 0, \quad i = 0 \ldots, k.$$

$$
\begin{aligned}
p &= \sum_{i=0}^{k}\sum_{j=1}^{k-i} c_{ij}\hat{x}_1^i\hat{x}_2^j = \hat{x}_2\sum_{i=0}^{k}\sum_{j=1}^{k-i} c_{ij}\hat{x}_1^i\hat{x}_2^{j-1} \\
&= \hat{x}_2\sum_{i=0}^{k}\sum_{j=0}^{k-i-1} c_{i,j+1}\hat{x}_1^i\hat{x}_2^j = \hat{x}_2\sum_{0\le i+j\le k-1} c_{i,j+1}\hat{x}_1^i\hat{x}_2^j \\
&= \hat{x}_2\tilde{q}(\hat{x}_1,\hat{x}_2).
\end{aligned}
$$

and $\tilde{q}(\hat{x}_1,\hat{x}_2)\in\mathbb{P}^{k-1}$. As $\hat{x}_2\neq 0$ on $N_{k-1}(\hat{\tau}),\tilde{q}\left(x_j\right)=0$, $x_j\in N_{k-1}(\hat{\tau})$. By the induction assumption, $\tilde{q}(\hat{x})=0$, we have $p(\hat{x})=0$, DONE!

Remark The proof easily extends to the case of $\mathbb{R}^d$. (an additional induction overd)

Let $S_h=\left\{\varphi\in C(\bar{\Omega}),\varphi|_\tau\in\mathbb{P}^k,\tau\in\mathcal{J}\right\}\subset H^1(\Omega)$, where $\mathcal{J}$ is an admissible triangulation of Ω.

Other characterization

$\mathcal{N}_k(\mathcal{J})=\bigcup_{\tau\in\mathcal{J}}\mathcal{N}(\tau)$, where $\mathcal{N}(\tau)$ denotes the set of nodes(DoFs) in τ. We set $\widetilde{S_h}$ to be the collection of all piecewise $\mathbb{P}^k$ functions determined by these nodes. So given $v\left(x_j\right),x_j\in\mathcal{N}$. There exists a unique $u\in\widetilde{S_h}$ satisfying $u\left(x_j\right)=v\left(x_j\right),x_j\in\mathcal{N}$.

Proposition 3.3

$\widetilde{S_h}=S_h$.

Proof

- $\widetilde{S_h}\subset S_h$, we need to show that $\varphi\in\widetilde{S_h}$ is continuous. Let $\varphi_1=\varphi|_{\tau_1}$ and $\varphi_2=\varphi|_{\tau_2}$. So $\varphi_2|_e$ and $\varphi_1|_e$ are 1D polynomials of degree k, which share the same values at the $k+1$ distinct nodes on e. Both must equal the unique polynomial interpolating the $k+1$ values on the edge and hence $\varphi_2|_e=\varphi_1|_e$, namely φ is continuous on e.
- $S_h\subset\widetilde{S_h}$ is clear since given $\varphi\in S_h$, we set $\widetilde{\varphi}$ by $\widetilde{\varphi}\left(x_i\right)=\varphi\left(x_i\right)\left(x_i\in\mathcal{N}\right)$ as both φ and $\widetilde{\varphi}$ are in $\mathbb{P}^k$ on each triangle and agree on the nodes of the triangle. Thus $\widetilde{\varphi}=\varphi(\forall\tau\in\mathcal{J})$, we get $\varphi\in\widetilde{S_h}$.

We then have the finite element basis $\varphi_i\in S_h$ such that.

$$\varphi_i\left(x_j\right)=\delta_{ij},x_j\in\mathcal{N}$$

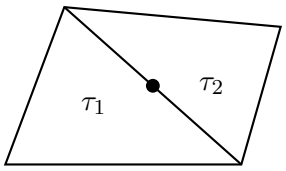
and define the interpolation operator

$$\mathcal{I}_h(u)=\sum_{x_i\in\mathcal{N}}u\left(x_i\right)\varphi_i(x).$$

Example 3.2 $K=\tau$ (triangle),$P=\mathbb{P}^1$(linear).$\Sigma=\{\ell_i(p)=p\left(m_i\right),i=1,2,3\}$ Is Σ unisolvent?

Figure 3.5: $\mathbb{P}^1$ element on a triangle

As in the earlier proof, this will be unisolvent if is unisolvent on $\hat{\tau}$.



✍ **Exercise 3.2** check that when plug in the points in $p = \alpha + \beta\hat{x}_1 + \gamma\hat{x}_2$, the corresponding 3×3 matrix is nonsingular.

$N =$ midpoints of edges of $\tau, \tau \in \mathcal{J}$. $S_h =$ Set of piecewise linear functions determined by the nodes.

$\varphi \in S_h, \varphi_1 = \varphi|_{\tau_1}$, $\varphi_2 = \varphi|_{\tau_2}$. This is not enough to impose continuity. The resulting space is not a subset of $H^1(\Omega)$. But it is useful to solve other problems. We call such finite element as "Crouvix - Raviart elements".

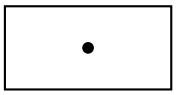
Example 3.3

$K =$ rectangle.

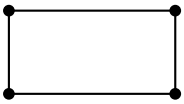
$$P = \mathbb{Q}^k = \left\{ \sum_{i,j=0}^{k} c_{ij} x_1^i x_2^j \right\} \subset \mathbb{P}^{2k} \quad \text{dim } = (k+1)^2.$$

$\mathbb{P}^k \subseteq \mathbb{Q}^k$

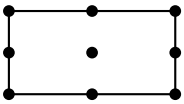
$k = 0$



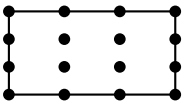
$k = 1$: bilinears



$k = 2$: bi-quadratics



$k = 3$



etc.

Figure 3.6: Tensor product element

Consider B an affine mapping, $\hat{\tau} = (0,1)^2$. $B : \hat{\tau} \to \tau$.

$$p \in \mathbb{Q}^k \to q = p \circ B \in \mathbb{Q}^{2k}$$

$$p = \sum_{i,j=0}^{k} c_{ij} x_1^i x_2^j \Rightarrow p \circ B = \sum_{i,j=0}^{k} c_{ij} \left(b_{11}\hat{x}_1 + b_{12}\hat{x}_2\right)^i \left(b_{21}\hat{x}_1 + b_{22}\hat{x}_2\right)^j$$

So B does not necessarily map $\mathbb{Q}^k \to \mathbb{Q}^k$.

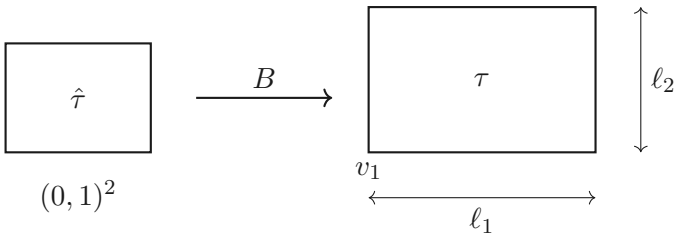
Proposition 3.4

The nodes are unisolvent. ♠

Notation: τ is a rectangle and $\mathcal{N}_k(\tau)$ is the set of nodes for $\mathbb{Q}^k$ on τ.

Proof Given τ, we can map $\hat{\tau}$ onto τ by translating and dialation in each direction.

$$B\hat{x} = v_0 + \begin{pmatrix} \ell_1 & \\ & \ell_2 \end{pmatrix} \begin{pmatrix} \hat{x_1} \\ \hat{x_2} \end{pmatrix}$$



This transformation $P \to P \circ B$ maps $\mathbb{Q}^k$ onto $\mathbb{Q}^k$. This means that it suffices to show the unisolvence for the reference $\hat{\tau} = (0,1)^2$ and implies unisolvence for all τ.

Do induction on k. For $k = 0$, DONE!

Assume it holds for $k-1$. (all τ). We need only show that it works for the reference $\hat{\tau}$. If $p \in \mathbb{Q}^k$ then $p|_{\text{edge}}$ of τ is a polynomial $q \in \mathbb{P}^k(1\text{D})$.

Assume $p \in \mathbb{Q}^k$ and $p(x_j) = 0$ for $x_j \in \mathcal{N}_k(\hat{\tau}), p|_{x_2=0} \in \mathbb{P}^k(1\text{D})$ and vanishes at $k+1$ distinct nodes $\Rightarrow p = 0$ on $x_2 = 0$. Also $p = 0$ on $x_1 = 0$. Then, if follows that

$$p = x_1 x_2 \sum_{i,j=1}^{k} C_{ij} x_1^{i-1} x_j^{j-1} \in \mathbb{Q}^{k-1}.$$

i.e $p = x_1 x_2 q$, with $q \in \mathbb{Q}^{k-1}$. And q vanishes for $x_j \in \mathcal{N}_{k-1}(\hat{\tau})$, By induction $q = 0$, so $p = 0$.

3.3 $H^2(\Omega)$ conforming elements

In this section, we discuss about finite element spaces that are subspace of $H^2(\Omega)$. As the previous lemma. for H^2-conforming elements, we require that the finite element functions are not only continuous, but its gradient are also continuous in the entire region.

$$K = (x_0, x_1), \quad P = \mathbb{P}^3, \quad \Sigma = \{\ell_0(p) = p(x_0), \ell_1(p) = p(x_1), \ell_2(p) = p'(x_0), \ell_3(p) = p'(x_1)\}$$

Proposition 3.5 (unisolvence)

There is a unique cubic satisfying the interpolation problem: $p \in \mathbb{P}^3$ *so that*

$$p(x_i) = f_i, \quad i = 0, 1 \quad \textit{and} \quad p'(x_i) = g_i, \quad i = 0, 1.$$

Note that x_0, *and* x_1 *are distinct.*

Proof We need only to show that the only solution to $p(x_i) = p'(x_i) = 0$ in $\mathbb{P}^3$ is the zero polynomial. if $p(x_i) = p'(x_i) = 0$, then $(x - x_i)^2$ devides p. So $(x - x_0)^2 (x - x_1)^2$ devides p, but p is cubic. So p must be 0.

Now we consider the two-dimensional case. The corresponding finite element space is called Argyris

elements:

$$
\begin{aligned}
K &= \text{triangle.} \\
P = \mathbb{P}^5 &= \left\{ \sum_{0 \le i+j \le 5} C_{ij} x_1^i x_2^j \right\} \\
\Sigma &= \left\{ D^\alpha P(v_i), i = 1,2,3, |\alpha| \le 2, \frac{\partial P}{\partial \mathbf{n}}(m_i), i = 1,2,3 \right\} \\
\dim(\Sigma) &= 3 \times 6 + 3 = 21 = \dim\left(\mathbb{P}^5\right).
\end{aligned}
$$

Proposition 3.6

The nodes are unsolvent.

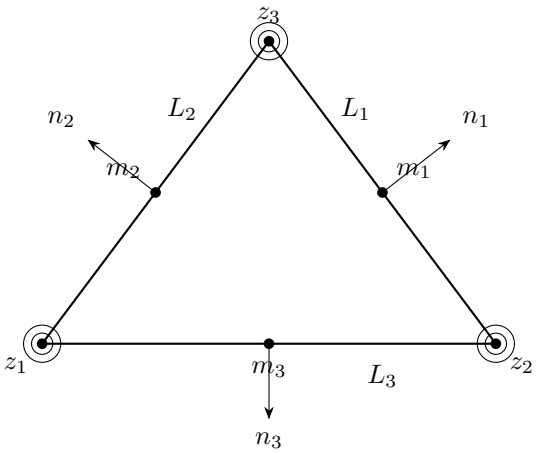
Remark The normal derivative does not map to the that in the reference triangle by affine mapping.

Lemma 3.3

1. *In the one dimensional case, consider* $K = [x_0, x_1], p = \mathbb{P}^4$ *and* $\Sigma = \{P(x_i), P'(x_i), i = 0,1$ *and* $P\left(\frac{x_0+x_1}{2}\right)\}$*, we can show that* Σ *is unisolvent for* $\mathbb{P}^4$*. (Exercise).*
2. *If* $\Sigma = \{P(x_i), P'(x_i), P''(x_i)$ *with* $i = 0,1\}$*. Then* $\sum$ *is unisolvent for* $\mathbb{P}^5$*. (Exercise).*

Lemma 3.4

Let $B : \hat{\tau} \to \tau$ *be an affine map with* $B(\hat{e}_i) = e_i$*. If* $D^\alpha f|_{e_i} = 0, |\alpha| \leqslant 1$*, then* $D^\alpha (f \circ B)|_{\hat{e}_i} = 0, |\alpha| \leqslant 1$*. (Exercise; chain rule).*



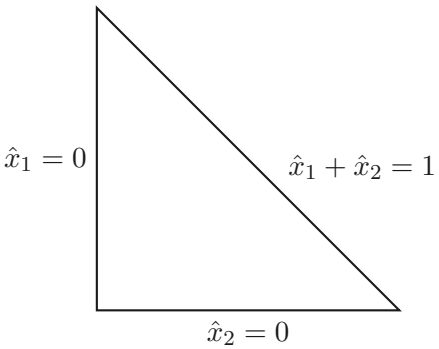
Proof [Proof of the proposition]

Assume $P \in \mathbb{P}^5$ and vanishes at the nodes of τ. Then: $P(v_i) = 0$, $\frac{\partial P}{\partial \tau}(v_i) = 0$, and $\frac{\partial^2 P}{\partial \tau^2}(v_i) = 0$ for $i = 1,2$. As $P|_e$ is a polynomial of degree 5 . Using Lemma 3.3 implies that $P = 0$ on e. and $P_\tau = \nabla P \cdot \tau$ vanishes on e.

On the other hand, $\frac{\partial P}{\partial \mathbf{n}}$ is in $\mathbb{P}^{k-1}$ (Exercise). $\frac{\partial P}{\partial n}\Big|_e$ is in ord $\mathbb{P}^{k-1}$. and $\frac{\partial P}{\partial \mathbf{n}}(m) = 0$, $\frac{\partial P}{\partial \mathbf{n}}(v_i) = 0, i = 1,2.$ and $\frac{\partial}{\partial \tau}\left(\frac{\partial p}{\partial \mathbf{n}}\right)(v_i) = 0, i = 1.2$ (combination of free 2^{nd} derivative). By Lemme 3.3, $\frac{\partial P}{\partial \mathbf{n}} = 0$ on e. So p and its first derivative vanish on each edge. Lemma 3.4 implies that $q = P \circ B$ and its derivatives vanishes on each

edge of $\hat{\tau}$. $q = \sum_{i+j=k} a_j \widehat{x}_1^i x_2^j$. So q vanishes on $\widehat{x}_2 = 0$.

$$\Rightarrow \quad C_{i,0} = 0, \quad i = 0, \cdots, k.$$
$$\frac{\partial P}{\partial \widehat{x}_2} = 0 \quad \Rightarrow \quad C_{i1} = 0, \quad i = 0, \cdots, k-1.$$



Thus $\widehat{x}_2^2$ devides q. Similarly $\widehat{x}_1^2$ devides q. Therefore $q = \widehat{x}_1^2 \widehat{x}_2 (\alpha + \beta \widehat{x}_1 + \gamma \widehat{x}_2)$. Apply to $\widehat{x}_1 + \widehat{x}_2 = 1 \Rightarrow (1 - \widehat{x}_1 - \widehat{x}_2)^2$ devides $q \Rightarrow \alpha = \beta = \gamma = 0$.

Theorem 3.2

Let $\mathcal{J}$ be a triangulation of $\Omega \subseteq \mathbb{R}^2$. The Argyris finite element space is in $C^1(\bar{\Omega})$. i.e. it is a subspace of $H^2(\Omega)$. ($H^2(\Omega)$-conforming)

♡

Proof If triangles meet at a vertex, then the function and its 4^{st} derivative values are the same at the vertex.

Suppose we have a piecewise quintic on τ_1 and τ_2 which share the DoFs on V_1, V_2, and m.

Let $p|_{\tau_1} = p_1$, $p|_{\tau_2} = p_2$. Since any first derivative is a linear combination of p_{x_1} and p_{x_2}.

So $\frac{\partial p_1}{\partial \tau}(vi) = \frac{\partial p_2}{\partial \tau}(v_i), i = 1, 2$. Similarly. $\frac{\partial^2 p}{\partial x_1^2}$, $\frac{\partial^2 p}{\partial x_1 \partial x_2}$ and $\frac{\partial^2 p}{\partial x_2^2}$ imply that

$$\frac{\partial^2 p_1}{\partial \tau^2}(v_i) = \frac{\partial^2 p_2}{\partial \tau^2}(v_i) \quad i = 1, 2$$
$$\frac{\partial^2 p_1}{\partial \tau \partial \mathbf{n}}(v_i) = \frac{\partial^2 p_2}{\partial \tau \partial \mathbf{n}}(v_i) \quad i = 1, 2.$$

Now let $q_j = p_j|_e \in \mathbb{P}^5$ (1D). $j = 1, 2$. We have

$$q_1(v_i) = q_2(v_i) \quad i = 1, 2$$
$$q_1'(v_i) = q_2'(v_i) \quad i = 1, 2.$$
$$q_1''(v_i) = q_2''(v_i) \quad i = 1, 2.$$

These DoFs are unisolvent on $\mathbb{P}^5(1D) \Rightarrow q_1 = q_2$.

So the function is continuous.

Next we have to check its derivatives are continuous. It suffices to check normal derivatives on e are continuous. Let $q_i = \left.\frac{\partial p_i}{\partial \mathbf{n}}\right|_e \in \mathbb{P}^4(1D), i = 1, 2$. So

$$q_1(v_i) = q_2(v_i) \quad i = 1, 2$$
$$\frac{\partial^2 p_1}{\partial \tau \partial \mathbf{n}}(v_i) = \frac{\partial q_1}{\partial \tau}(v_i) = \frac{\partial q_2}{\partial \tau}(v_i) = \frac{\partial^2 p_2}{\partial \tau \partial \mathbf{n}}(v_i).$$
$$q_1(m) = q_2(m).$$

By Lemma 3.3, we have $q_1 = q_2$. So the 1^{st} derivative of f is continuous.

Chapter 4 Finite Element Method

4.1 Finite element approximation properties

Consider the reference triangulation element $\hat{\tau}$ in $\mathbb{R}^2$.

We shall show that

$$\inf_{p\in\mathbb{P}^0} \|u-p\|_{H^1(\hat{\tau})} \leqslant C|u|_{H^1(\hat{\tau})} = C\left(\int_{\hat{\tau}} |\nabla u|^2 \,\mathrm{d}x\right)^{1/2}.$$

C doesn't depend on u.

$$\begin{aligned}\|u-p\|^2_{H^1(\hat{\tau})} &= \|u-p\|^2_{L^2(\hat{\tau})} + \|\nabla(u-p)\|^2_{L^2(\hat{\tau})} \\ &= \|u-p\|^2_{L^2(\hat{\tau})} + |u|^2_{H^1(\hat{\tau})}\end{aligned}$$

We need only to bound

$$\inf_{p\in\mathbb{P}^0} \|u-p\|_{L^2(\hat{t})} \leqslant c|u|_{H^1(\hat{\tau})}.$$

Next we shall show that for $u \in C^\infty(\hat{\tau})$. there holds

$$\|u-\bar{u}\|_{L^2(\hat{\tau})} \leqslant c|u|_{H^1(\hat{\tau})}.$$

with

$$\bar{u} = \frac{1}{|\hat{t}|}\int_{\hat{\tau}} u \,\mathrm{d}x = 2\int_{\hat{\tau}} u \,\mathrm{d}x = \fint_{\hat{\tau}} u \,\mathrm{d}x \quad \text{(average)}.$$

$$\begin{aligned}u(x) &= u(x) - u(x_1,y_2) + u(x_1,y_2) - u(y) + u(y) \\ &= \underbrace{\int_{y_2}^{x_2} u_s(x_1,s)\,\mathrm{d}s + \int_{y_1}^{x_1} u_t(t,y_2)\,\mathrm{d}t}_{J(x,y)} + u(y).\end{aligned}$$

Now we integrate y over $\hat{\tau}$:

$$\frac{1}{2}u(x) = \int_{\hat{\tau}} J(x,y)\,\mathrm{d}y + \int_{\hat{\tau}} u(y)\,\mathrm{d}y.$$

Thus

$$u(x) = 2\int_{\hat{\tau}} J(x,y) + \bar{u}.$$

$$\Rightarrow \quad u(x) - \bar{u} = 2\int_{\hat{\tau}} J(x,y)\,\mathrm{d}y.$$

Square and integrate over x in $\hat{\tau}$ to get

$$\|u-\bar{u}\|^2_{L^2(\hat{\tau})} = \int_{\hat{\tau}} |u(x)-\bar{u}|^2 \,\mathrm{d}x = 4\int_{\hat{\tau}}\left(\int_{\hat{\tau}} J(x,y)\,\mathrm{d}y\right)^2 \mathrm{d}x.$$

Bound the right-hand side using $(a+b)^2 \leqslant 2\left(a^2+b^2\right)$ and Schwarz inequality. (Exercise).

$$\|u-\bar{u}\|^2_{L^2(\hat{\tau})} \leqslant c|u|^2_{H^1(\hat{\tau})}$$

Lemma 4.1 (Deny-Lions)

Let B be connected bounded domain with a Lipschitz continuous boundary, then there texists a constant

C, such that.

$$\inf_{p\in\mathbb{P}^k} \|u-p\|_{H^{k+1}(B)} \leqslant C|u|_{H^{k+1}(B)} \equiv C\left(\sum_{|\alpha|=k+1} \|D^\alpha u\|^2_{L^2(B)}\right)^{1/2}.$$

♡

Remark This lemma sometimes is called the Bramble-Hilbert lemma.

Sketch of proof:

Consider $p\in\mathbb{P}^k$, $\ell_{ij}(p)=c_{ij}$ where $p=\sum_{0\leqslant i+j\leqslant k} C_{ij}x_1^i x_2^j$. So ℓ_{ij} is a bonded linear functional on $\mathbb{P}^k$ (a finite dimensional space). By the Hahn-Banach thm, ℓ_{ij} can be extended to the linear functional on $H^{k+1}(B)$. We will show that

$$\|v\|_{H^{k+1}(B)} \leq C(B)\left\{|v|_{H^{k+1}(B)} + \sum_{0\leqslant i+j<k} |\ell_{ij}(v)|\right\} \tag{4.1}$$

If we verify equation 4.1, then we choose

$$q=\sum_{0\leq i+j\leq k} \ell_{ij}(v)\cdot x_1^i x_2^j \in \mathbb{P}^k$$

So that

$$\inf_{p\in\mathbb{P}^k} \|v-p\|_{H^{k+1}(B)} \leqslant \|v-q\|_{H^{k+1}(B)}$$

Note that $\ell_{ij}(v-q)=\ell_{ij}(v)-\ell_{ij}(q)=0$. By equation (4.1), we have

$$\|v-q\|_{H^{k+1}(B)} \leqslant C(B)|v-q|_{H^{k+1}(B)} = C(B)|v|_{H^{k+1}(B)}$$

(Since $D^\alpha q=0$, for $|\alpha|=k+1$.)

We already had shown that (4.1) implied the Deny-Lion Lemma. Suppose that (4.1) does not hold. This imply that there exists v_n with $\|v_n\|_{H^{k+1}(B)}=1$ and

$$\left\{|v_n|_{H^{k+1}(B)} + \sum_{0\leq i+j\leq k} |\ell_{ij}(v_n)|\right\} < \frac{1}{n} \tag{4.2}$$

The compactness of $H^1(B)$ in $L^2(B)$ implies the compactness of $H^{k+1}(B)$ in $H^k(B)$. This means that there is a subsequence of $\{v_n\}$ which converges in $H^k(B)$. We still denote the subsequence by $\{v_n\}$ which still satisfies (4.2).

(4.2) implies the sequence is Cauchy in $H^{k+1}(B)$ and v_n convegent in $H^k(B)$ implies Cauchy in the $H^k(B)$-norm $\Rightarrow$ convegent in $H^{k+1}(B)$, i.e, there exists $v\in H^{k+1}(B)$ with $v_n\to v$ in $H^{k+1}(B)$. Taking the limit in (4.1) implies $|v|_{H^{k+1}(B)}=0 \Rightarrow D^\alpha v=0, |\alpha|=k+1$. So for functions in Sobolev spaces $D^\alpha v=0$ for $|\alpha|=k+1$ on a connected domain $B\Rightarrow v\in\mathbb{P}^k$. $v=\sum l_{ij}(v)x_1^i x_2^j$, $0\leq i+j\leq k$. and taking the limit in (4.1) implies $\sum_{0\leq i+j\leq k}|l_{ij}(v)|=0. \Rightarrow v=0$. This is a contradiction since $\|v_n\|_{H^{k+1}(B)}=1$ cannot converge in H^{k+1} to a function with norm 0 .

Remark This is also called the Bramble-Hilbert Lemma.

Definition 4.1

$\ell: V\to\mathbb{R}$ is sublinear and bounded if

1. $|\ell(v+w)|\leq|\ell(v)|+|\ell(w)|, \quad \forall v,w\in V$
2. $|\ell(\alpha v)|=|\alpha||\ell(v)|$.
3. $|\ell(v)|\leqslant c\|v\|_V$. *(Bounded)*

♣

Lemma 4.2 (Bramble - Hilbert Lemma)

Assume B is bounded connected domain with Lipschitz domain. Let q be a bounded sublinear functional satisfying

$$q(p) = 0, \quad \forall p \in \mathbb{P}^k \text{ on } H^{k+1}(B).$$

Then there exists $C(B)$ such that

$$|q(v)| \leqslant C(B)\|q\||v|_{H^{k+1}(B)}.$$

Here $\|q\| = \sup_{v \in H^{k+1}(B)} \dfrac{|q(v)|}{\|v\|_{H^{k+1}(B)}}$.

♡

Example 4.1 $q(v) = \|v - \hat{\mathcal{I}}(v)\|_{H^{k+1}(B)}$ where $\hat{\mathcal{I}} : H^{k+1}(B) \to \mathbb{P}^k(B)$ and satisfies $\hat{\mathcal{I}}(p) = p$ for all $p \in \mathbb{P}^k$. $\hat{\mathcal{I}}$ is linear and bounded. And $q(p) = \left\|p - \hat{\mathcal{I}}p\right\|_{H^{k+1}} = 0 \quad \forall p \in \mathbb{P}^k$.

$$\begin{aligned} |q(v+w)| =& \|(v+w) - \hat{\mathcal{I}}(v+w)\|_{H^{k+1}(B)} \\ \leqslant& \|v - \hat{\mathcal{I}}(v)\|_{H^{k+1}(B)} + \|w - \hat{\mathcal{I}}(w)\|_{H^{k+1}(B)} \\ =& |q(v)| + |q(w)|. \end{aligned}$$

Proof [Bramble - Hilbert Lemma] $v \in H^{k+1}(B)$ and $p \in \mathbb{P}^k$

$$\begin{aligned} |q(v)| &= |q(v - p + p)| \\ &\leq |q(v-p)| + |q(p)| \\ &= |q(v-p)| \\ &\leq \|q\| \cdot \|v - p\|_{H^{k+1}(B)} \\ &\leq \|q\| \cdot c(B) \cdot |v|_{H^{k+1}(B)} \quad \text{(Deny-Lions)} \end{aligned}$$

4.2 Interpolation error on triangular

Let h_τ = diameter of τ = length of the largest edge , $\underline{h}_\tau$ = diameter of the largest ball which can be included in τ.

Let $h = \max_{\tau \in \mathcal{J}} h_\tau$ and $\rho^{-1} = C_0(\mathcal{J}) = \min_{\tau \in \mathcal{J}} \dfrac{\underline{h}_\tau}{h_\tau}$. If $\rho(\mathcal{J}) = \max_{\tau \in \mathcal{J}} \dfrac{h_\tau}{\underline{h}_\tau} \leq c$, we say $\mathcal{J}$ is shape regular.

Proposition 4.1

Assume triangle $\tau \subset \mathcal{J}$, if its smallest angle is bounded below by a constant. Then $\mathcal{J}$ is shape-regular.

♠

T is an affine mapping of $\hat{\tau}$ onto τ just like Figure (4.1)

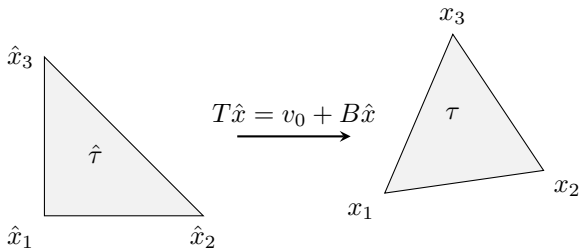


Figure 4.1: Affine mapping from the reference triangle $\hat{\tau}$ to a physical triangle τ

Lemma 4.3

$$|v|_{H^m(\tau)} \leqslant C|\det B|^{\frac{1}{2}} \cdot \left\|B^{-1}\right\|_2^m \cdot |\hat{v}|_{H^m(\hat{\tau})}$$

Here v is defined on τ and $\hat{v} = v \circ T$,

$$\|B\| = \sup_{x \in \mathbb{R}^d} \frac{|Bx|}{|x|} \quad \left(d = 2 \text{ for } \mathbb{R}^2\right)$$

Also,

$$|\hat{v}|_{H^m(\hat{\tau})} \leqslant C|\det B|^{-\frac{1}{2}} \|B\|_2^m |v|_{H^m(\tau)}$$

♡

Proof We just proof the condition for H^1. Idea: change of variable end chain rule.

Note that $v = \hat{v} \circ T^{-1}$ $\left(v(x) = \hat{v}\left(T^{-1}x\right)\right)$

$T^{-1}x = \hat{v}_0 + B^{-1}x$

$\nabla_x v = \nabla_{\hat{x}} \hat{v} \cdot D\left(T_x^{-1}\right) = \nabla_{\hat{x}} \hat{v} \cdot B^{-1}$

$$\begin{aligned}
\int_\tau |\nabla v|^2 \, \mathrm{d}x &= \int_{\hat{\tau}} |\det B| \cdot \left|\nabla \hat{v} \cdot B^{-1}\right|^2 \, \mathrm{d}\hat{x} \\
&= \int_{\hat{\tau}} \left|B^{-1} \nabla \hat{v}\right|^2 \, \mathrm{d}\hat{x} \cdot |\det B| \\
&\leqslant |\det B| \cdot \left\|B^{-1}\right\|^2 \int_{\hat{\tau}} |\nabla \hat{v}|^2 \, \mathrm{d}\hat{x}
\end{aligned}$$

Thus

$$|v|^2_{H^1(\tau)} \leq |\det B| \left\|B^{-1}\right\|_2^2 |\hat{v}|^2_{H^1(\hat{\tau})}$$

Here we note. $\|B\|_2 = \left\|B^\top\right\|_2$, $\left\|B^{-1}\right\|_2 = \left\|B^{-\top}\right\|_2$. (Exercise). and

$$\left\|B^{-\top} w\right\| \leq \left\|B^{-\top}\right\| \|w\| = \left\|B^{-1}\right\| |\omega|.$$

So

$$|v|_{H^1(\tau)} \leqslant \left\|B^{-1}\right\| \cdot |\det B|^{\frac{1}{2}} |\hat{v}|_{H^1(\hat{\tau})}.$$

We can thus similarly show the inequality for m.

Lemma 4.4

$\|B\| \leqslant C h_\tau$ *and* $\left\|B^{-1}\right\| \leqslant C_0 \underline{h}_\tau^{-1}$ *(C_0 depends on $C_0(\tau)$).*

♡

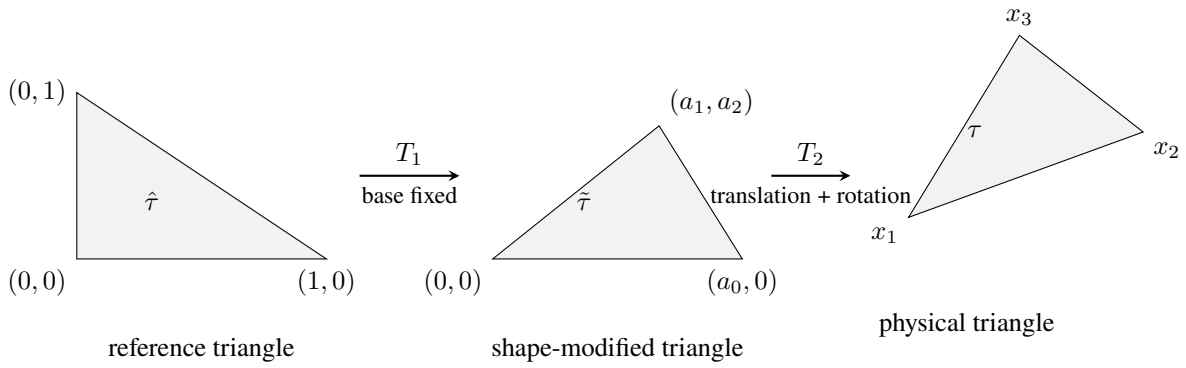


Figure 4.2: Two-step decomposition of the affine mapping

Proof $T_1\hat{x} = B_1\hat{x} = y$ where $B_1 = \begin{pmatrix} a_0 & a_1 \\ 0 & a_2 \end{pmatrix}$, $T_2 y = B_2 y + v_0$, thus $B = B_2 \circ B_1$.

Check B_2 is unitary i.e. $|B_2 x| = |x|, \|B_2\| = \left\|B_2^{-1}\right\| = 1$.

$$\|B\| = \|B_2 B_1\| \leq \|B_2\| \, \|B_1\| = \|B_1\|.$$

So $\|B_1\| \leqslant C \, \|B_1\|_\infty$ where $\|B_1\|_\infty = \max\limits_{i,j=1,2.} \left|(b_1)_{ij}\right|$. Since $a_0, a_1, a_2 \leqslant \operatorname{diam} \tau = h_\tau$. We get

$$\|B\| \leq \|B_1\| \leqslant c \cdot h_\tau$$

On the otherhand

$$B_1^{-1} = \begin{pmatrix} a_0^{-1} & -a_1/a_2 \\ 0 & a_2^{-1} \end{pmatrix} = (a_0 a_2)^{-1} \begin{pmatrix} a_2 & -a_1 \\ & a_0 \end{pmatrix} \Longrightarrow \left\|B_1^{-1}\right\| \leq C \cdot \left|(a_0 a_2)^{-1}\right| h_\tau$$

Note that Area $(\tau) = \frac{1}{2} a_0 a_2 \geqslant \pi \left(\frac{1}{2} \underline{h}_\tau\right)^2 \geqslant \frac{\pi}{4} c_0^2(\tau) \cdot h_\tau^2$. $\left(c_0(\tau) \leqslant \frac{h_\tau}{\underline{h}_\tau}\right)$.

So $(a_0 a_2)^{-1} \leqslant c h_\tau^2 \Rightarrow \left\|B_1^{-1}\right\| \leqslant c h_\tau^{-1}$

Here the constant $c.$ depends $C_0^{-1}(\tau)$.

$$C_0^{-1}(\tau) = \max_{\tau \in \mathcal{J}} \frac{h_\tau}{\underline{h}_\tau}$$

if $c_0^{-1}(\tau) \leqslant c$, we say $\mathcal{J}$ is shape-regular.

Theorem 4.1 (Example of Sobolev embedaling)

For $k > \frac{d}{2}, H^k(\Omega) \subseteq C(\bar{\Omega})$ and there exists a constant $C > 0$ such that

$$\|u\|_{L^\infty(\Omega)} \leqslant C \|u\|_{H^k(\Omega)} \quad \textit{for all} \quad u \in H^k(\Omega).$$

♡

Remark For Lagrange interpolant, we need function evaluation,

$$\{x_i\} \subset \mathcal{N}_k(\tau), \tau \in \mathcal{J}. \quad f(x_i) < +\infty \Rightarrow H^k(\Omega) \subseteq C(\bar{\Omega})$$

Theorem 4.2 (Interpolation error)

Let Ω be a polygonal domain. $\mathcal{J}$ is a shape regular triangulation of Ω (admissible).

$$S_h = \left\{\phi \in C(\Omega) \textit{ satisfying } \phi|_\tau \in \mathbb{P}^k \textit{ for } \tau \in \mathcal{J}\right\}.$$

For $u \in H^m(\Omega), k+1 \geqslant m > \frac{d}{2}$ and $l = 0, 1.$

$$\|u - \mathcal{I}_h u\|_{H^l(\Omega)} \leqslant c h^{m-l} \|u\|_{H^m(\Omega)}.$$

where $\mathcal{I}_h$ denotes the Lagrange finite element interpolation operator into S_h.

♡

Proof

$$\int_\Omega \left|D^\alpha (u - \mathcal{I}_h u)\right|^2 \, \mathrm{d}x = \sum_{\tau \in \mathcal{J}} \int_\tau \left|D^\alpha (u - \mathcal{I}_h u)\right|^2 \, \mathrm{d}x$$

Consider

$$\begin{aligned} I_1 &= \int_\Omega |u - \mathcal{I}_h u|^2 \, \mathrm{d}x = \sum_{\tau \in \mathcal{J}} \int_\tau |u - \mathcal{I}_h u|^2 \, \mathrm{d}x \\ &= \sum_{\tau \in \mathcal{J}} \int_{\hat{\tau}} |\det B| \left|\widehat{u - \mathcal{I}_h u}\right|^2 \, \mathrm{d}\hat{x} \quad (\hat{f} = f \circ T) \\ I_2 &= \int_\Omega |\nabla (u - \mathcal{I}_h u)|^2 \, \mathrm{d}x \leq \sum_{\tau \in \mathcal{J}} |\det B| \left\|B^{-1}\right\|_2^2 \int_{\hat{\tau}} \left|\nabla \left(\widehat{u - \mathcal{I}_h u}\right)\right|^2 \, \mathrm{d}\hat{x}. \end{aligned}$$

As T is linear, $\widehat{u - \mathcal{I}_h u} = \hat{u} - \widehat{\mathcal{I}_h u}$. Let $\mathcal{I}$ be the interpolation with respect to the nodes on $\hat{\tau}$, then

$$\widehat{\mathcal{I}_h u} = \mathcal{I} \hat{u} \quad (\text{ Exercise })$$

So

$$I_2 \leqslant \sum_{\tau\in\mathcal{J}} |\det B| \left\|B^{-1}\right\|_2^2 \int_{\hat{\tau}} |\nabla(\hat{u}-\mathcal{I}\hat{u})|^2 \,\mathrm{d}\hat{x}$$

$$I_1 = \sum_{\tau\in\mathcal{J}} |\det B| \int_{\hat{\tau}} |\hat{u}-\mathcal{I}\hat{u}|^2 \,\mathrm{d}\hat{x}.$$

In fact $\mathcal{I}$ is stable in $H^m(\Omega)$.

$$\|\mathcal{I}\hat{w}\|_{H^m(\hat{\tau})} \leq C\|\hat{w}\|_{H^m(\hat{\tau})}.$$

Now we verify the stability of $\mathcal{I}$. Let $\hat{\varphi}_i$ be the finite element basis function for $\mathbb{P}^k$ on $\hat{\tau}$.

$$\mathcal{I}\hat{w} = \sum_{i=1}^{\frac{(k+1)(k+2)}{2}} w\left(\hat{x}_i\right)\hat{\varphi}_i(\hat{x}) \quad \hat{x}_i \in \mathcal{N}_{m-1}(\hat{\tau}) \quad \hat{\varphi}_i\left(\hat{x}_j\right) = \delta_{ij}.$$

$$\begin{aligned}\|\mathcal{I}\hat{w}\|_{H^m(\hat{\tau})} &\leqslant \sum_{i=1}^{\dim\left(\mathbb{P}^k\right)} \left|\hat{w}\left(\hat{x}_i\right)\right| \|\hat{\varphi}_i\|_{H^m(\hat{\tau})} \\ &\leqslant \underbrace{\max\|\hat{\varphi}_i\|_{H^m(\hat{\tau})} \cdot \dim\left(\mathbb{P}^k\right)}_{\text{constant}} \cdot c\|\hat{w}\|_{H^m(\hat{\tau})}\end{aligned}$$

Note that $\mathcal{I}p = p$ for any $p \in \mathbb{P}^{m-1} \quad (m-1 \leqslant k)$. Thus

$$\begin{aligned}|\hat{u}-\mathcal{I}\hat{u}| &= |\hat{u}-p+\mathcal{I}p-\mathcal{I}\hat{u}| \\ &\leqslant |\hat{u}-p| + |\mathcal{I}(\hat{u}-p)|.\end{aligned}$$

$$\begin{aligned}\int_{\hat{\tau}} |\hat{u}-\mathcal{I}\hat{u}|^2 \,\mathrm{d}\hat{x} &\leqslant 2\int_{\hat{\tau}} |\hat{u}-p|^2 \,\mathrm{d}\hat{x} + 2\int_{\hat{\tau}} |\mathcal{I}(\hat{u}-p)|^2 \,\mathrm{d}\hat{x} \\ &\leqslant C\int_{\hat{\tau}} |\hat{u}-p|^2 \,\mathrm{d}\hat{x} \leqslant c\|\hat{u}-p\|_{H^m(\hat{\tau})}^2.\end{aligned}$$

Similarly.

$$\int_{\hat{\tau}} |\nabla(\hat{u}-\mathcal{I}\hat{u})|^2 \,\mathrm{d}\hat{x} \leqslant C\cdot\|\hat{u}-p\|_{H^m(\hat{\tau})}$$

Take the infimum over p and apply the Deny-Lions Lemma.

$$\int_{\hat{\tau}} |D^\alpha(\hat{u}-\mathcal{I}\hat{u})|^2 \,\mathrm{d}\hat{x} \leqslant C \inf_{p\in\mathbb{P}^k} \|\hat{u}-p\|_{H^m(\hat{\tau})}^2 \leqslant C|\hat{u}|_{H^m(\hat{\tau})}^2 \quad (|\alpha| \leqslant 1).$$

Thus

$$I_1 \leqslant C\sum_{\tau\in\mathcal{J}} |\det B||\hat{u}|_{H^m(\hat{\tau})}^2 \leqslant C\sum_{\tau\in\mathcal{J}} \|B\|^{2m}\cdot|u|_{H^m(\tau)}^2 \leqslant Ch^{2m}|u|_{H^m(\Omega)}^2$$

So

$$\|u-\mathcal{I}_h\|_{L^2(\Omega)} \leq ch^m\|u\|_{H^m(\Omega)}$$

Then

$$\begin{aligned}I_2 &\leqslant \sum_{\tau\in J} C|\det B| \left\|B^{-1}\right\|^2 |\hat{u}|_{H^m(\hat{\tau})}^2 \leqslant \sum_{\tau\in\mathcal{J}} \left\|B^{-1}\right\|_2^2 \|B\|_2^{2m}|u|_{H^m(\tau)}^2 \\ &\leqslant C\sum_{\tau\in\mathcal{J}} \underline{h}_\tau^{-2}h_\tau^{2m}|u|_{H^m(\tau)}^2 \leqslant C\sum_{\tau\in\mathcal{J}} h_\tau^{2m-2}|u|_{H^m(\tau)}^2 \\ &\leqslant Ch^{2m-2}|u|_{H^m(\Omega)}^2\end{aligned}$$

We get

$$|\nabla(u-\mathcal{I}_h u)|_{H^1(\Omega)} \leq ch^{m-1}\|u\|_{H^m(\Omega)}.$$

Summary.

$$\|u - \mathcal{I}_h u\|_{L^2(\Omega)} + h\,\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \leqslant Ch^l \|u\|_{H^l(\Omega)} \quad 2 \leq l \leq k+1.$$

4.3 Inverse estimate

Lemma 4.5

Let τ be a triangle. p is a polynomial defined on τ. with $p \in \mathbb{P}^k$. Then

$$|p|_{H^1(\tau)} \leqslant Ch_\tau^{-1} \|p\|_{L^2(\tau)}.$$

♡

Proof Since $\mathbb{P}^k$ is a finite dimensional space in τ. We know that any norm for $\mathbb{P}^k$ are equivalent. So for $\hat{p} \in \mathbb{P}_k(\hat{\tau})$

$$C_1 \|\hat{p}\|_{L^2(\hat{\tau})} \leqslant \|\hat{p}\|_{H^1(\hat{\tau})} \leqslant C_2 \|\hat{p}\|_{L^2(\hat{\tau})}$$

where C_1, C_2, depends on k and $\hat{\tau}$. This implies that

$$|\widehat{p}|_{H^1(\hat{\tau})} \leqslant C \|\hat{p}\|_{L^2(\hat{\tau})}.$$

By the previous Lemma, let $\hat{p} = p \circ T$ and we get

$$\begin{aligned}
|p|^2_{H^1(\tau)} &\leq |\det B| \left\|B^{-1}\right\|_2^2 |\hat{p}|^2_{H^1(\hat{\tau})} \\
&\leq C_2 |\det B| \left\|B^{-1}\right\|_2^2 \|\hat{p}\|^2_{L^2(\hat{\tau})} \\
&\leq C_2 |\det B| \left\|B^{-1}\right\|_2^2 |\det B|^{-1} \|p\|^2_{L^2(\tau)} \\
&= C_2 \left\|B^{-1}\right\|_2^2 \|p\|^2_{L^2(\tau)}
\end{aligned}$$

Therefore, since $\mathcal{J}$ is shape regular, we have

$$|p|_{H^1(\tau)} \leqslant C \left\|B^{-1}\right\|_2 \|p\|_{L^2(\tau)} \leqslant ch_\tau^{-1} \|p\|_{L^2(\tau)}$$

Remark This also imples that

$$\|p\|_{H^1(\tau)} \leq ch_\tau^{-1} \|p\|_{L^2(\tau)}$$

It is trivial to check that

$$\|p\|^2_{H^1(\tau)} = \|p\|^2_{L^2(\tau)} + |p|^2_{H^1(\tau)} \leq \|p\|^2_{L^2(\tau)} + C^2 h_\tau^{-2} \|p\|^2_{L^2(\tau)} \leq Ch_\tau^{-2} \|p\|^2_{L^2(\tau)},$$

Define $h = \max\limits_{\tau \in \mathcal{J}} h_\tau$. If for all $\tau \in \mathcal{J}, \min\limits_{\tau} h_\tau \geqslant ch$, we say $\mathcal{J}$ is quasi-uniform. It is equivalent to $\exists c_1, c_2 > 0, \forall \tau \in \mathcal{J}, c_1 h \leq h_\tau \leq c_2 h$.

We can further show that for $p \in S_h$.

$$\begin{aligned}
|p|^2_{H^1(\Omega)} \leq \|p\|^2_{H^1(\Omega)} &= \sum_{\tau \in \mathcal{J}} \|p\|^2_{H^1(\tau)} \leqslant C \sum_{\tau \in \mathcal{J}} h_\tau^{-2} \|p\|^2_{L^2(\tau)} \\
&\leqslant c \sum_{\tau \in \mathcal{J}} h^{-2} \|p\|^2_{L^2(\tau)} = ch^{-2} \|p\|^2_{L^2(\Omega)}
\end{aligned}$$

i.e.

$$\|p\|_{H^1(\Omega)} \leq Ch^{-1} \|p\|_{L^2(\Omega)}, \quad \forall p \in S_h$$

This is the global inverse inequality.

Chapter 5 high dimensional problem

5.1 Two-Dimensional Problem

Consider the following boundary value problem:

$$\begin{cases} u - \Delta u = f & \text{in } \Omega \\ \dfrac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n} = g & \text{on } \partial\Omega \end{cases} \tag{5.1}$$

Multiply (5.1) by a test function $\varphi \in C^\infty(\bar{\Omega})$:

$$\int_\Omega u\varphi - \int_\Omega \Delta u\varphi = \int_\Omega f\varphi.$$

Integrate by parts (Green's formula):

$$-\int_\Omega \Delta u \cdot \varphi = -\int_{\partial\Omega} (\nabla u \cdot \mathbf{n})\varphi \,\mathrm{d}s + \int_\Omega \nabla u \cdot \nabla\varphi \,\mathrm{d}x.$$

Since $\nabla u \cdot \mathbf{n} = g$ on $\partial\Omega$, substituting this back gives:

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

Rearranging the terms:

$$\underbrace{\int_\Omega (u\varphi + \nabla u \cdot \nabla\varphi)\,\mathrm{d}x}_{A(u,\varphi)} = \underbrace{\int_\Omega f\varphi \,\mathrm{d}x + \int_{\partial\Omega} g\varphi \,\mathrm{d}s}_{\langle F,\varphi\rangle}.$$

Note that $A(u,\varphi) = (u,\varphi)_{H^1(\Omega)}$, which is the standard inner product in H^1.

So our weak formulation is: find $u \in H^1(\Omega)$ satisfying

$$A(u,\varphi) = \langle F,\varphi\rangle, \quad \forall \varphi \in H^1(\Omega).$$

Definition 5.1

A linear map $F : H \to \mathbb{R}$ on a Hilbert space H is called a linear functional. It is bounded (or continuous) if there exists a constant $C \geqslant 0$ satisfying

$$|\langle F,\varphi\rangle| \leqslant C\|\varphi\|_H, \quad \forall \varphi \in H.$$

♣

Theorem 5.1 (Riesz Representation Theorem)

Let H be a Hilbert space and F be a bounded linear functional on H. Then there exists a unique element $u \in H$ such that

$$(u,\varphi)_H = \langle F,\varphi\rangle, \quad \forall \varphi \in H.$$

Moreover, $\|u\|_H = \|F\|_{H'}$.

♡

Since $A(u,\varphi) = (u,\varphi)_{H^1}$ is exactly the inner product, we have shown that (5.1) admits a unique weak solution.

5.1.1 Generalization

Consider the general elliptic operator:

$$Lu = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial}{\partial x_j}u\right) + q(x)u. \tag{5.2}$$

Assumptions:

1. $a_{ij} \in L^\infty(\Omega)$ and satisfy the **Uniform Ellipticity Condition**: There exists a constant $\nu > 0$ such that
$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geqslant \nu|\xi|^2, \quad \forall \xi \in \mathbb{R}^d, \text{ a.e. } x \in \Omega.$$
(Note: Do not confuse the vector $\xi \in \mathbb{R}^d$ with the test function v).
2. $q \in L^\infty(\Omega)$ and there is a constant $q_0 > 0$ such that $q(x) \geqslant q_0$ a.e. in Ω.

Define the co-normal derivative on $\partial\Omega$:

$$\frac{\partial u}{\partial \nu_A} = \sum_{i,j=1}^{d} n_i a_{ij}\frac{\partial u}{\partial x_j} = (\mathbf{n}^T A \nabla u).$$

We want to solve the Neumann problem:

$$\begin{cases} Lu = f & \text{in} \quad \Omega, \\ \dfrac{\partial u}{\partial \nu_A} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{5.3}$$

(Given $f \in L^2(\Omega), g \in L^2(\partial\Omega)$).

Derivation: Take $\varphi \in C^\infty(\bar{\Omega})$.

$$\int_\Omega Lu\varphi \,\mathrm{d}x = \int_\Omega f\varphi \,\mathrm{d}x.$$

Analyze the Left Hand Side (LHS):

$$\begin{aligned} \text{LHS} &= -\sum_{i,j=1}^{d}\int_\Omega \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right)\varphi \,\mathrm{d}x + \int_\Omega qu\varphi \,\mathrm{d}x \\ &= \sum_{i,j=1}^{d}\int_\Omega a_{ij}(x)\frac{\partial u}{\partial x_j}\frac{\partial \varphi}{\partial x_i}\,\mathrm{d}x - \sum_{i,j=1}^{d}\int_{\partial\Omega} n_i a_{ij}(x)\frac{\partial u}{\partial x_j}\varphi \,\mathrm{d}s + \int_\Omega qu\varphi \,\mathrm{d}x \\ &= \sum_{i,j=1}^{d}\int_\Omega a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial \varphi}{\partial x_i}\,\mathrm{d}x - \int_{\partial\Omega} \underbrace{\left(\sum_{i,j=1}^{d} n_i a_{ij}\frac{\partial u}{\partial x_j}\right)}_{\frac{\partial u}{\partial \nu_A} = g} \varphi \,\mathrm{d}s + \int_\Omega qu\varphi \,\mathrm{d}x. \end{aligned}$$

Variational formulation: Find $u \in H^1(\Omega)$ such that:

$$A(u,\varphi) = \langle F, \varphi\rangle, \quad \forall \varphi \in H^1(\Omega).$$

Here:

$$\begin{aligned} A(u,\varphi) &= \int_\Omega qu\varphi \,\mathrm{d}x + \sum_{i,j=1}^{d}\int_\Omega a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial \varphi}{\partial x_i}\,\mathrm{d}x, \\ \langle F, \varphi\rangle &= \int_\Omega f\varphi \,\mathrm{d}x + \int_{\partial\Omega} g\varphi \,\mathrm{d}s. \end{aligned}$$

5.1.2 Properties of the Bilinear Form

1. **Boundedness (Continuity):** A is bounded on $H^1(\Omega) \times H^1(\Omega)$.

$$\left|\int_\Omega quv\right| \leqslant \|q\|_{L^\infty}\|u\|_{L^2}\|v\|_{L^2} \leqslant \|q\|_{L^\infty}\|u\|_{H^1}\|v\|_{H^1}.$$

$$\begin{aligned}\left|\sum_{i,j}\int_\Omega a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\right| &\leqslant \sum_{i,j}\|a_{ij}\|_{L^\infty}\left\|\frac{\partial u}{\partial x_j}\right\|_{L^2}\left\|\frac{\partial v}{\partial x_i}\right\|_{L^2} \\ &\leqslant C\sum_{i,j}\|\nabla u\|_{L^2}\|\nabla v\|_{L^2} \leqslant C'\|u\|_{H^1}\|v\|_{H^1}.\end{aligned}$$

   Thus, $|A(u,v)| \leq M\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}$.

2. **Boundedness of Functional F:** By Cauchy-Schwarz and Trace Theorem:

$$|\langle F, \varphi\rangle| \leq \|f\|_{L^2}\|\varphi\|_{L^2} + \|g\|_{L^2(\partial\Omega)}\|\varphi\|_{L^2(\partial\Omega)} \leq C\|\varphi\|_{H^1(\Omega)}.$$

3. **Coercivity:** We verify that $A(u,u) \geqslant \alpha\|u\|^2_{H^1(\Omega)}$.

$$\begin{aligned}A(u,u) &= \int_\Omega qu^2\,\mathrm{d}x + \int_\Omega \sum_{i,j=1}^d a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial u}{\partial x_j}\,\mathrm{d}x \\ &\geqslant q_0\|u\|^2_{L^2} + \nu\|\nabla u\|^2_{L^2} \quad \text{(using ellipticity assumption)} \\ &\geqslant \min(q_0,\nu)\left(\|u\|^2_{L^2} + \|\nabla u\|^2_{L^2}\right) \\ &= \alpha\|u\|^2_{H^1(\Omega)}.\end{aligned}$$

5.1.3 Symmetry and Existence

Assume that the coefficient matrix is symmetric, i.e., $a_{ij}(x) = a_{ji}(x)$ for $i,j = 1\cdots,d$. Check symmetry of A:

$$\begin{aligned}A(v,u) &= \int_\Omega qvu + \sum_{i,j=1}^d\int_\Omega a_{ij}\frac{\partial v}{\partial x_j}\frac{\partial u}{\partial x_i} \\ &\quad \text{Let } k = j, l = i \text{ (rename indices)} \\ &= \int_\Omega quv + \sum_{l,k=1}^d\int_\Omega a_{lk}\frac{\partial v}{\partial x_k}\frac{\partial u}{\partial x_l} \\ &\quad \text{Since } a_{lk} = a_{kl} \text{ (symmetry assumption)} \\ &= \int_\Omega quv + \sum_{k,l=1}^d\int_\Omega a_{kl}\frac{\partial u}{\partial x_l}\frac{\partial v}{\partial x_k} = A(u,v).\end{aligned}$$

Conclusion: Since $A(\cdot,\cdot)$ is a **symmetric**, **bounded**, and **coercive** bilinear form, it defines an inner product on $H^1(\Omega)$ equivalent to the standard one. By the Riesz Representation Theorem applied to the Hilbert space $(H^1(\Omega), A(\cdot,\cdot))$, there exists a unique solution $u \in H^1(\Omega)$ to the problem.

Remark If $a_{ij} \neq a_{ji}$, A is not symmetric. We cannot use Riesz directly. Instead, we must use the **Lax-Milgram Theorem**, which only requires boundedness and coercivity (not symmetry).

5.2 Lax-Milgram

Theorem 5.2 (Lax-Milgram)

Let V be a Hilbert space. Let $A : V \times V \to \mathbb{R}$ be a bounded, coercive bilinear form, and $F : V \to \mathbb{R}$ be a bounded linear functional. Then there exists a unique $u \in V$ satisfying

$$A(u, \varphi) = F(\varphi), \quad \forall \varphi \in V.$$

♡

Remark We do not need to assume the symmetry of the coefficients, i.e., $a_{ij}(x) = a_{ji}(x)$ is not required. This is the main difference from the Riesz Representation Theorem.

Example 5.1 Consider the following boundary value problem:

$$\begin{cases} u - \Delta u + \alpha u_x &= f \qquad \text{in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0 \qquad \text{on } \partial\Omega. \end{cases}$$

Multiplying by a test function $\varphi \in H^1(\Omega)$ and integrating by parts, we obtain the variational formulation. Note that the boundary term vanishes due to the Neumann boundary condition:

$$\begin{aligned} \int_\Omega (u - \Delta u + \alpha u_x)\varphi &= \int_\Omega u\varphi + \int_\Omega \nabla u \cdot \nabla \varphi - \int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{n}} \varphi + \alpha \int_\Omega u_x \varphi \\ &= \int_\Omega u\varphi + \int_\Omega \nabla u \cdot \nabla \varphi + \alpha \int_\Omega u_x \varphi. \end{aligned}$$

Define the bilinear form $A(\cdot, \cdot)$ and linear functional $F(\cdot)$ as:

$$A(u, \varphi) = \int_\Omega u\varphi + \int_\Omega \nabla u \cdot \nabla \varphi + \alpha \int_\Omega u_x \varphi, \quad F(\varphi) = \int_\Omega f\varphi.$$

1. Boundedness: Using the Cauchy-Schwarz inequality:

$$\left| \int_\Omega u_x \varphi \right| \leqslant \|u_x\|_{L^2} \|\varphi\|_{L^2} \leqslant \|u\|_{H^1} \|\varphi\|_{H^1}.$$

It is easy to check that other terms are also bounded, so A is bounded on $H^1(\Omega) \times H^1(\Omega)$.

2. Coercivity: We need to estimate the convection term $\int_\Omega u_x u$.

Method 1 (Young's Inequality):

$$\begin{aligned} \left| \int_\Omega u_x u \right| &\leqslant \|u_x\|_{L^2} \|u\|_{L^2} \leqslant \frac{1}{2}\|u_x\|_{L^2}^2 + \frac{1}{2}\|u\|_{L^2}^2 \\ &\leqslant \frac{1}{2}\left(\|\nabla u\|_{L^2}^2 + \|u\|_{L^2}^2\right) = \frac{1}{2}\|u\|_{H^1}^2. \end{aligned}$$

Thus,

$$\begin{aligned} A(u, u) &= \|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2 + \alpha \int_\Omega u_x u \\ &\geqslant \|u\|_{H^1}^2 - |\alpha| \left| \int_\Omega u_x u \right| \\ &\geqslant \|u\|_{H^1}^2 - \frac{|\alpha|}{2}\|u\|_{H^1}^2 = \left(1 - \frac{|\alpha|}{2}\right)\|u\|_{H^1}^2. \end{aligned}$$

Therefore, if $|\alpha| < 2$, the bilinear form is coercive.

Method 2 (Integration by parts / Trace Theorem): Alternatively, use the divergence theorem noting that $uu_x = \frac{1}{2}\partial_x(u^2)$:

$$\begin{aligned} \int_\Omega u_x u \, dx &= \frac{1}{2} \int_\Omega \frac{\partial}{\partial x}\left(u^2\right) \, dx \\ &= \frac{1}{2} \int_{\partial\Omega} u^2 n_x \, ds \quad \text{(where } n_x \text{ is the x-component of } \mathbf{n}\text{)}. \end{aligned}$$

Since $|n_x| \leq 1$, we have:

$$\left| \int_\Omega u_x u \right| \leqslant \frac{1}{2} \int_{\partial\Omega} u^2 \, ds \leqslant \frac{C_{tr}^2}{2} \|u\|_{H^1(\Omega)}^2,$$

where C_{tr} is the constant from the trace inequality $\|v\|_{L^2(\partial\Omega)} \leq C_{tr}\|v\|_{H^1(\Omega)}$.

This leads to:

$$A(u,u) \geqslant \left(1 - \frac{|\alpha| C_{tr}^2}{2}\right) \|u\|_{H^1}^2.$$

This provides coercivity if α is sufficiently small relative to the domain geometry (which determines C_{tr}).

5.3 inf-sup condition

Ladyzhenskaya-Babuška-Brezzi condition.

5.3.1 Matrix Case (On a finite dimensional case)

1. All nom on finite dimensional space are equivalent.
   Let space V (finite dimensional) has two norms $\|\cdot\|_i, i = 1, 2$. Then there are constants C_1, C_2, such that $(C_1 > 0)$
   $$C_1 \|u\|_2 \leqslant \|u\|_1 \leqslant C_2 \|u\|_2 .$$
2. All linear operators are bounded.
3. All bilinear forms are bounded.
   Consider $M : \mathbb{R}^m \to \mathbb{R}^n$. M is a $n \times m$ matrix.
   $$A(u,v) = (Mu) \cdot v, \quad u \in \mathbb{R}^m, v \in \mathbb{R}^n.$$
4. All linear functionals are bounded.
   Given a linear functional F on $\mathbb{R}^n$, we seek the solution $u \in \mathbb{R}^m$ satisfying $A(u,\varphi) = \langle F, \varphi \rangle, \quad \forall \varphi \in \mathbb{R}^n$.
   This is the same as
   $$Mu = b, \quad b_i = \langle F, e_i \rangle .$$

   For uniqueness, we need
   $$M : \mathbb{R}^m \xrightarrow{1:1} \mathbb{R}^n$$
   $$M : \mathbb{R}^m \xrightarrow{\text{1:1onto}} \text{Range}(M) \subseteq \mathbb{R}^n$$

$\exists M^{-1} = \text{ Range } M \longmapsto \mathbb{R}^m$.

M^{-1} is linear and hence bounded. There exists a C satisfying

$$\left|M^{-1} y\right| \leqslant C|y|, \quad y \in \text{Range } M.$$

This can be rewritten as $x = M^{-1}y$

$$|x| \leqslant C|Mx| \text{ fon all } x \in \mathbb{R}^m$$

$$|Mx| = \sup_{\substack{y \in \mathbb{R}^n \\ y \neq 0}} \frac{(Mx, y)}{|y|}$$

So.

$$|x| \leqslant C \sup_{y \in \mathbb{R}^n} \frac{(Mx, y)}{|y|}, \forall x \in \mathbb{R}^m$$

This is called the inf-sup condition. The minimal C satisfies

$$\frac{1}{c} = \inf_{x\in\mathbb{R}^m} \sup_{y\in\mathbb{R}^n} \frac{(Mx, y)}{|x||y|}$$

We are looking for solution to $Mu = b$. This has a solution only if $b \in$ Range (M).

$$\mathbb{R}^n = \text{ Range } M \oplus \text{Ker } M^\top \quad \text{(linear algebra).}$$

$b \in$ Range M iff $b \perp$ Ker $M^\top$, i.e. $b \in \{y \in \mathbb{R}^n, (y, \xi) = 0, \forall \xi \in \text{ ker } M^\top\}$.
$\xi \in \ker M^\top$ iff $M^\top \xi = 0$, ift $\left(M^\top \xi, x\right) = 0$ for all $x \in \mathbb{R}^m$ iff $(\xi, Mx) = 0 \quad \forall x \in \mathbb{R}^m$.

Compatibility condition $Mu = b$ has a solution iff $(b, \xi) = 0$, for all $\xi \in W := \{y \in \mathbb{R}^n$ and that $A(x, y) = 0 \quad \forall x \in \mathbb{R}^n\}$.

5.3.2 LBB condition: Extension to infinite dimensional case

Assumption

1. U, V are Banach spaces;
2. $A(u, v)$ is a bounded bilinear form on $U \times V$, namely
$$|A(u, v)| \leqslant c\|u\|_U \|v\|_V, u \in U, v \in V$$
3. Inf-sup condition, $\exists \beta > 0$. satisfying
$$\|u\|_U \leqslant \beta \cdot \sup_{v\in V} \frac{A(u, v)}{\|v\|_V}, \text{ for all } u \in U.$$
4. F is a bounded linear functional on V.
5. F satisfies the compactibility condition. iff $\langle F, \xi\rangle = 0$. for all $\xi \in W := \{\xi \in V, A(u, \xi) = 0$ for all $u \in U\} \subseteq V$ closed subspace.

Theorem 5.3

Assume that $(A_1) - (A_4)$ hold and let V is reflexive. Then $\exists$ unique $u \in U$ satisfying

$$A(u, \varphi) = \langle F, \varphi\rangle, \quad \forall \varphi \in V.$$

iff F satisfies the compactibility condition. ♡

Remark All Hilbert spaces are reflexive.

Lax-Milgram Theorem: consider $U = V =$ a Hilbert space. Assume $A(\cdot, \cdot)$ and F satisfy $(A_2), (A_4)$, respectively and that A is coercive on $U \times U$ i.e.

$$\|u\|^2 \leqslant \alpha A(u, u). \quad \forall u \in U.$$

Then $\exists$ unique $u \in U$ satisfying. $A(u, \varphi) = F(\varphi), \quad \forall \varphi \in U$.

Proof of the Remark: To apply the previous theorem we need only check (A_3) and compactibility condition.

Compactibility condition. If $w \in W$ then $A(u, w) = 0 \quad \forall u \in U$.

$$\Rightarrow \text{ Since } w \in U \Rightarrow \quad 0 = \alpha \cdot A(w \cdot w) = \|w\|_U^2 \Rightarrow w = 0.$$
$$\Rightarrow \quad \langle F, w\rangle = 0 \quad \forall w \in W \quad (\text{ since } w = 0).$$

For (A3), the coercivity implies that

$$\|u\|_v \leq \alpha \cdot \frac{A(u, u)}{\|u\|_U} \leq \alpha \sup \frac{A(u, v)}{\|v\|_V}.$$

5.3.2.1 Reflexivity

Let W be a Banach space, $W^* = \{$ set of bounded linear functional $\}$.

Given $F \in W^*$, we define $\|F\|_{W^*}$ to be the minimal constant $C > 0$ satisfying

$$|\langle F, \varphi\rangle| \leqslant c\|\varphi\|_W, \quad \forall \varphi \in W$$

So W^* is a Banach space under this norm.

Let $W^{**} = \{$set of bounded linear functionals on $W^*\}$ (also a Banach space).

We identify $w \in W$ with $F_w \in W^{**}$ satisfying $\langle F_w, G\rangle = \langle G, w\rangle$. So. $W \subseteq W^{**}$.

Definition 5.2

W is reflexive if $W = W^{**}$. ♣

Remark Hilbert spaces are reflexive by the Riesz Represendation Theorem.

$L^p(\Omega)$ is reflexive for $p \in (1, \infty)$. $\Rightarrow W^{k,p}(\Omega)$ is reflexive.

Theorem 5.4 (Hahn-Banach theorem)

If F is a bounded linear functional on a subspace W_0 of normed linear space W. Then F can be extended to $\widetilde{F}$ on With

$$\|F\|_{W_0} = \|\widetilde{F}\|_W.$$

♡

Theorem 5.5 (Open mapping theorem)

If a linear bounded. $A : U \xrightarrow{1:1\ onto} V$ (U, V are Banach spaces). Then A^{-1} is bounded i.e. $\exists C$ with

$$\left\|A^{-1} f\right\|_U \leq C\|f\|_V \quad \forall f \in V.$$

♡

Proof Proof of the theorem:

Define $\widetilde{A} : U \to V^*$ by.

$$\langle \widetilde{A}\theta, v\rangle = A(\theta, v). \quad \forall \theta \in U, v \in V$$

Inf- sup condition.

$$\|\theta\|_U \leqslant \alpha \sup_{v \in V} \frac{\left\langle \widetilde{A}\theta, v \right\rangle}{\|v\|_V} \tag{5.4}$$

Check Range $(\widetilde{A})$ is closed. Let $F_n \to F$ in Range$(\widetilde{A})$.

Let $\widetilde{A}\theta_n = F_n$. By (5.4) $\theta_n \to \theta$ in U and it is easy to check that $\widetilde{A}\theta = F$.

So Range $(\widetilde{A})$ is closed. (5.4) also implies that $\widetilde{A}$ is one-to-one.

Clearly by the open mapping theorem, $\widetilde{A}u = F \in V^*$ has a solution iff $F \in$ Range $\widetilde{A}$ in which case. the solution is unique.

Claim: $F \in$ Range $\widetilde{A}$ if $\Leftrightarrow < F, \varphi >= 0, \quad \forall \varphi \in W$.

If $F \in$ Range $\widetilde{A}$, then $F = \widetilde{A}\theta$ for some $\theta \in U$.

For $\varphi \in W$, $\langle F, \varphi\rangle = \langle \widetilde{A}\theta, \varphi\rangle = A(\theta, \varphi) = 0$. (Since $\varphi \in W$).

Conversely $F \notin$ Range $\widetilde{A}$, define ℓ on Range $(\widetilde{A}) \oplus$ span F by.

$$\ell\left(\omega^* + \beta F\right) = \beta. \quad \omega^* \in \text{ Range } \widetilde{A}.$$

Since Range $(\widetilde{A})$ is closed and $F \notin$ Range$(\widetilde{A})$, $\exists \varepsilon > 0$, such that

$$\|F - w^*\|_{V^*} > \varepsilon \quad \text{for all} \quad w^* \in \text{Range } \widetilde{A}$$

If $p \neq 0$.

$$\|w^* - \beta F\|_{V^*} = |\beta| \cdot \left\| \frac{1}{\beta} w^* - F \right\|_{V^*} > \varepsilon |\beta|$$

$$|\beta| = \left|\ell^* \left(\omega^* + \beta F\right)\right| \leq \frac{1}{\varepsilon} \left\|\omega^* + \beta F\right\|_{V^*} \Rightarrow \ell \text{ is bounded.}$$

By the Hahn-Banach theorem, we can extended ℓ to a bounded linear functional $\tilde{\ell}$ on V^*. Since V is reflexive, there is a $v \in V$ satisfying

$$\left\langle \hat{l}, w^* \right\rangle = \langle w^*, v \rangle . \quad \forall w^* \in V^*.$$

Consider $w^* = F$.

$$\langle F, v \rangle = \langle \tilde{\ell}, F \rangle = 1.$$

Also, for $u \in U$,

$$0 = \ell\left(\widetilde{A}u\right) = \tilde{\ell}\left(\widetilde{A}u\right) = \langle \widetilde{A}u, v \rangle = A(u, v) \Rightarrow v \in W$$

i.e. F does not satisfy the compactibility condition.

Since the problem: Find $u \in U$ satisfying $A(u, \varphi) = F(\varphi)$ is the same as $\widetilde{A}u = F$. Our proof is complete.

Remark Moreover

$$\|u\|_U \leq \sup_{\varphi \in V} \frac{A(u, \varphi)}{\|\varphi\|_V} = \alpha \cdot \sup_{\varphi \in V} \frac{\langle F, \varphi \rangle}{\|\varphi\|_V} = \alpha \cdot \|F\|_{V^*}$$

5.4 Galerkin finite element approximation

5.4.1 Variational Formulation

Assume we have a variational formulation which satisfies the properties of Lax-Milgram:

1. V is a Hilbert space.
2. $A(\cdot, \cdot)$ is a bounded bilinear form on V. That is, there exists a constant $\|A\|$ such that:
$$|A(u, v)| \leq \|A\| \|u\|_V \|v\|_V, \quad \forall u, v \in V.$$
3. F is a bounded functional on V.
4. A is coercive, i.e., there exists $\alpha > 0$ satisfying:
$$\alpha \|v\|_V^2 \leq A(v, v), \quad \forall v \in V.$$

Let $V_h \subseteq V$, then V_h satisfies (1)-(4) and hence there is a unique solution to: Find $u_h \in V_h$ s.t.

$$A(u_h, \varphi) = \langle F, \varphi \rangle, \quad \forall \varphi \in V_h.$$

So u_h is the Galerkin approximation to the solution $u \in V$ of:

$$A(u, \varphi) = \langle F, \varphi \rangle, \quad \forall \varphi \in V.$$

Subtracting the above equation gives (Galerkin orthogonality):

$$A(u - u_h, \varphi) = 0, \quad \forall \varphi \in V_h. \tag{5.5}$$

5.4.2 Céa's Lemma

Lemma 5.1 (Céa's Lemma)

$$\|u - u_h\|_V \leq \frac{1}{\alpha}\|A\| \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{5.6}$$

Here $\|A\|$ is the smallest constant satisfying

$$|A(\omega, v)| \leq \|A\|\|\omega\|_V \|u\|_V.$$

This is a quasi-optimal estimate. ♡

Proof By coercivity:

$$\begin{aligned}
\|u - u_h\|_V^2 &\leq \frac{1}{\alpha} A(u - u_h, u - u_h) \\
&= \frac{1}{\alpha} A(u - u_h, u - v_h + v_h - u_h) \quad (\text{for any } v_h \in V_h) \\
&= \frac{1}{\alpha} A(u - u_h, u - v_h) \quad (\text{since } A(u - u_h, v_h - u_h) = 0) \\
&\leq \frac{1}{\alpha}\|A\|\|u - u_h\|_V \|u - v_h\|_V.
\end{aligned}$$

Dividing by $\|u - u_h\|_V$, we get the result.

$$\|u - u_h\|_V \leq \frac{1}{\alpha}\|A\|\|u - v_h\|_V \quad \forall v_h \in V_h \implies \|u - u_h\|_V \leq \frac{\|A\|}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V$$

5.4.3 L^2 Error Estimates (Nitsche's Trick/Duality)

Now we let $V_h = S_h = \{v_h \in H^1 : v_h|_\tau \in \mathbb{P}^k\}$. According to our finite element interpolation error estimates, we have:

$$\|u - u_h\|_V \leq C\|u - \mathcal{I}_h u\|_V \leq C\|u - \mathcal{I}_h u\|_{H^1} \leq Ch^{\ell-1}\|u\|_{H^\ell(\Omega)},$$

where $2 \leq \ell \leq k + 1$.

Assumption: Let Ω be convex and $w \in V \subseteq H^1(\Omega)$ be the unique solution to the dual problem:

$$A(\theta, w) = (\theta, g), \quad \forall \theta \in V.$$

Under the assumptions of Lax-Milgram, w is well defined, thus $w \in H^2(\Omega)$. We assume full elliptic regularity:

$$\|w\|_{H^2(\Omega)} \leq C\|g\|_{L^2(\Omega)}, \quad \forall g \in L^2(\Omega).$$

Theorem 5.6 (Nitsche's trick / Duality)

Assume the setup for Lax-Milgram, $V \subseteq H^1(\Omega)$, and full elliptic regularity. Then:

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^\ell \|u\|_{H^\ell(\Omega)}.$$

♡

Proof Let $e_h = u - u_h$. Let $w \in V$ solve:

$$A(\theta, w) = (\theta, e_h), \quad \forall \theta \in V.$$

So $e_h = u - u_h \in V \subset H^1(\Omega) \equiv L^2(\Omega)$. Set $\theta = e_h$ to get:

$$\begin{aligned}
\|e_h\|_{L^2(\Omega)}^2 &= (e_h, e_h) = A(e_h, w) \\
&= A(e_h, w - \chi), \quad \forall \chi \in S_h \quad (\text{by Galerkin orthogonality}).
\end{aligned}$$

Thus,

$$\|e_h\|^2_{L^2(\Omega)} \leq \|A\|\|e_h\|_{H^1(\Omega)}\|w-\chi\|_{H^1(\Omega)}.$$

Taking $\chi = \mathcal{I}_h w$ gives:

$$\|w-\mathcal{I}_h w\|_{H^1(\Omega)} \leq Ch\|w\|_{H^2(\Omega)} \leq Ch\|e_h\|_{L^2(\Omega)}.$$

Combining the above estimates:

$$\|e_h\|^2_{L^2(\Omega)} \leq \|A\|\|e_h\|_{H^1} \cdot Ch\|e_h\|_{L^2(\Omega)}.$$

Therefore:

$$\|e_h\|_{L^2(\Omega)} \leq Ch\|e_h\|_{H^1(\Omega)}.$$

Finally, using the energy norm estimate:

$$\|u-u_h\|_{L^2(\Omega)} \leq Ch \cdot Ch^{\ell-1}\|u\|_{H^\ell} = Ch^\ell\|u\|_{H^\ell}. \quad 2 \leq \ell \leq k+1$$

5.5 Poincare Inequality

Consider the problem:

$$\begin{aligned} -\Delta u &= f \quad \text{in} \quad \Omega = (0,1)^2 \\ u(x) &= 0 \quad \text{on} \quad \Gamma_D \text{ (essential BC, imposed strongly in } V) \\ \frac{\partial u}{\partial \mathbf{n}}(x) &= g \quad \text{on} \quad \Gamma_N \text{ (natural BC, imposed weakly)} \end{aligned} \tag{5.7}$$

Derivation of the weak form:

$$\begin{aligned} \int_\Omega f\varphi &= \int_\Omega -\Delta u\varphi \\ &= -\int_{\partial\Omega} \frac{\partial u}{\partial \mathbf{n}}\varphi + \int_\Omega \nabla u \cdot \nabla\varphi \\ &= -\int_{\Gamma_N} \underbrace{\frac{\partial u}{\partial \mathbf{n}}}_{=g}\varphi - \int_{\Gamma_D} \frac{\partial u}{\partial \mathbf{n}} \underbrace{\varphi}_{=0} + \int_\Omega \nabla u \cdot \nabla\varphi \\ &= -\int_{\Gamma_N} g\varphi + \int_\Omega \nabla u \cdot \nabla\varphi \end{aligned}$$

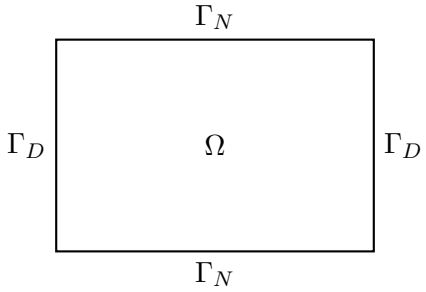


Figure 5.1: Domain Ω with mixed boundary conditions.

Define the Test Space (where we take test functions):

$$V = \left\{\varphi \in H^1(\Omega) : \varphi = 0 \text{ on } \Gamma_D\right\}.$$

Since the trace operator is continuous, i.e., $\|v\|_{L^2(\Gamma_D)} \leqslant C\|v\|_{H^1(\Omega)}$, V is a closed subspace of $H^1(\Omega)$.

Weak formulation: Find $u \in V$ satisfying

$$A(u, \varphi) = L(\varphi), \quad \forall \varphi \in V.$$

Here $A(u, \varphi) = \int_\Omega \nabla u \cdot \nabla \varphi$ and $L(\varphi) = \int_\Omega f\varphi + \int_{\Gamma_N} g\varphi$.

Theorem 5.7 (Poincare inequality)

Let Ω be a domain with Lipschitz continuous boundary $\partial\Omega \subseteq \mathbb{R}^d$. Let $V = \{\varphi \in H^1(\Omega) : \varphi = 0 \text{ on } \Gamma_D\}$. If the $(d-1)$-dimensional measure of Γ_D is positive ($|\Gamma_D| > 0$), there exists a constant $C_P > 0$ such that

$$\|u\|_{L^2(\Omega)} \leq C_P \|\nabla u\|_{L^2(\Omega)}, \quad \forall u \in V.$$

♡

Fact: If Ω has a Lipschitz continuous boundary, $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$ (Rellich-Kondrachov Theorem).

Proof Assume the estimate does not hold. Then for any $n \in \mathbb{N}$, there exists $u_n \in V$ such that

$$\|u_n\|_{L^2(\Omega)} > n\|\nabla u_n\|_{L^2(\Omega)}.$$

Normalize the sequence by setting $\|u_n\|_{L^2(\Omega)} = 1$. Then we have:

$$\|\nabla u_n\|_{L^2(\Omega)} < \frac{1}{n}.$$

Thus, $\|u_n\|^2_{H^1(\Omega)} = \|u_n\|^2_{L^2} + \|\nabla u_n\|^2_{L^2} \leq 1 + \frac{1}{n^2}$, so $\{u_n\}$ is bounded in $H^1(\Omega)$.

By the compact embedding result, there exists a subsequence (still denoted u_n) and a function $u \in L^2(\Omega)$ such that $u_n \to u$ strongly in $L^2(\Omega)$. Since $\|\nabla u_n\|_{L^2} \to 0$, $\{u_n\}$ is actually a Cauchy sequence in $H^1(\Omega)$ (because both L^2 parts and gradient parts converge).

Thus $u_n \to u$ strongly in $H^1(\Omega)$, which implies:

1. $\|\nabla u\|_{L^2} = \lim_{n\to\infty} \|\nabla u_n\|_{L^2} = 0 \implies u = \text{const.}$
2. Since $u_n \in V$, the trace $u|_{\Gamma_D} = 0$.

A constant function that is zero on a set of positive measure must be zero everywhere. Thus $u \equiv 0$. However, we assumed $\|u_n\|_{L^2} = 1$, so $\|u\|_{L^2} = \lim \|u_n\|_{L^2} = 1$. This contradicts $u \equiv 0$.

Remark The Poincare inequality implies that on the subspace V, the semi-norm $\|\nabla u\|_{L^2}$ is equivalent to the full H^1 norm. This ensures coercivity of $A(u, \varphi)$, allowing the application of the Lax-Milgram theorem.

Example 5.2

$$\begin{aligned} -\nabla \cdot (a\nabla u) &= f \quad \text{in} \quad \Omega \\ u &= g \quad \text{on} \quad \partial\Omega \end{aligned} \tag{5.8}$$

where $0 < a_0 \leqslant a(x) \leqslant a_1$.

Derivation: Multiply by test function φ (where $\varphi = 0$ on $\partial\Omega$):

$$-\int_\Omega \nabla \cdot (a\nabla u)\varphi = \int_\Omega a\nabla u \cdot \nabla\varphi - \int_{\partial\Omega} a(\nabla u \cdot \mathbf{n}) \underbrace{\varphi}_{0} = \int_\Omega a\nabla u \cdot \nabla\varphi.$$

Define the affine space for the trial solution:

$$\widetilde{V} = \{v \in H^1(\Omega) \mid v = g \text{ on } \partial\Omega\}.$$

Note: Usually we assume $g \in H^{1/2}(\partial\Omega)$ is the trace of some function $u_g \in H^1(\Omega)$.

Weak formulation: Find $u \in \widetilde{V}$ satisfying

$$\int_\Omega a\nabla u \cdot \nabla\varphi = \int_\Omega f\varphi, \quad \forall \varphi \in H^1_0(\Omega).$$

Lifting (Homogenization): Let $u = w + u_g$, where $u_g \in H^1(\Omega)$ is a known extension of boundary data

g, and $w \in H_0^1(\Omega)$ is the unknown. Plugging this in:

$$\int_\Omega a\nabla(w + u_g) \cdot \nabla\varphi = \int_\Omega f\varphi$$
$$\Rightarrow \int_\Omega a\nabla w \cdot \nabla\varphi = \int_\Omega f\varphi - \int_\Omega a\nabla u_g \cdot \nabla\varphi = \langle \tilde{F}, \varphi \rangle.$$

Now we can use Lax-Milgram on $H_0^1(\Omega)$ to find w, and then reconstruct $u = w + u_g$.

Chapter 6 Error of the Discrete System

6.1 Numerical Quadrature

We discussion the quadrature effects on the finite element approximation. Here we shall restrict our discussion to the two-dimensional space using triangular meshes.

6.1.1 Example Quadrature Schemes

Consider a triangle element τ with vertices V_1, V_2, V_3 and midpoints m_1, m_2, m_3.
Several common quadrature rule (e.g., for the reference triangle) are:

$$\int_\tau f(x)dx \approx Q_\tau f = \frac{|\tau|}{3}\sum_{i=1}^{3} f(v_i)\,.$$

$$\int_\tau f(x)dx \approx \frac{|\tau|}{3}\sum_{i=1}^{3} f(m_i)\,.$$

$$\int_\tau f(x)dx \approx |\tau|\cdot f(b_\tau)\,.$$

Definition 6.1

*Q_τ is **exact** on $\mathcal{P}^k$ if:*

$$Q_\tau\varphi = \int_\tau \varphi\,dx, \quad \forall\varphi\in\mathcal{P}^k$$

♣

All the above schemes are exact for $\mathcal{P}^0$. What about $\mathcal{P}^1, \mathcal{P}^2$?

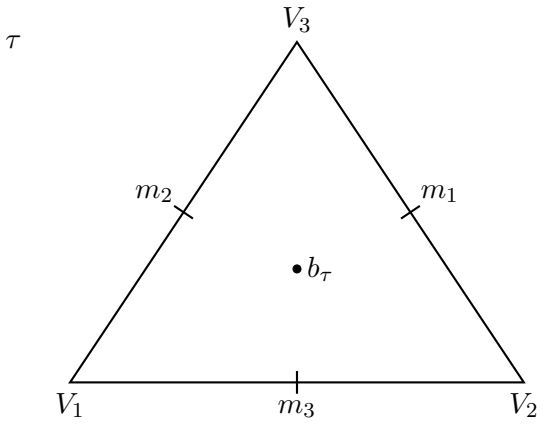
6.1.2 Discrete System Computation

In Finite Element Methods, we need to compute the stiffness matrix elements:

$$(A)_{ij} = \int_\Omega a(x)\nabla\varphi_j\cdot\nabla\varphi_i\,dx$$

We approximate it using quadrature:

$$(\widetilde{A}_h)_{ij} = \sum_{\tau\in\mathcal{T}_h} Q_\tau(a(x)\nabla\varphi_j\cdot\nabla\varphi_i)$$



Similarly for the load vector:

$$F_j = \int_\Omega f(x)\varphi_j(x)\,dx \implies (\widetilde{F}_h)_j = \sum_{\tau\in\mathcal{T}_h} Q_\tau(f(x)\varphi_j(x))$$

6.2 Weak Formulation with Quadrature

In general, we have the following weak formulation:

Find $u \in V$ such that:

$$A(u,\varphi) = \ell(\varphi), \quad \forall\varphi \in V \tag{6.1}$$

The standard Finite Element formulation reads:

Find $u_h \in V_h \subset V$ such that:

$$A(u_h,\varphi_h) = \ell(\varphi_h), \quad \forall\varphi_h \in V_h \tag{6.2}$$

But in practice, we cannot compute the integrals exactly. This leads to the **perturbed problem**:

Find $\tilde{u}_h \in V_h \subset V$ such that:

$$A_h(\widetilde{u}_h,\varphi_h) = \ell_h(\varphi_h), \quad \forall\varphi_h \in V_h \quad (*) \tag{6.3}$$

For example:

$$A_h(w_h,\varphi) = \sum_{\tau\in\mathcal{T}_h} Q_\tau\left(a(x)\nabla w_h\cdot\nabla\varphi\right)$$
$$\ell_h(\varphi) = \sum_{\tau\in\mathcal{T}_h} Q_\tau(f\cdot\varphi)$$

We consider two main issues:

1. The well-posedness of problem (*).
2. The error estimates between u and $\widetilde{u}_h$.

Definition 6.2 (Uniform V_h-ellipticity)

The form $A_h(\cdot,\cdot)$ is uniformly V_h-elliptic if there exists $\alpha > 0$, independent of h, satisfying:

$$\alpha\|v\|_V^2 \le A_h(v,v), \quad \forall v \in V_h \tag{6.4}$$

Remark Condition (6.4) implies the existence and uniqueness of the discrete solution $\widetilde{u}_h$ (Exercise).

6.3 Strang's First Lemma

Lemma 6.1 (Strang's First Lemma)

Assume that A_h is uniformly V_h-elliptic. Then there holds:

$$\|u-\widetilde{u}_h\|_V \le C \inf_{v_h\in V_h}\left\{\|u-v_h\|_V + \sup_{w_h\in V_h}\frac{|A(v_h,w_h)-A_h(v_h,w_h)|}{\|w_h\|_V}\right\} + \sup_{w_h\in V_h}\frac{|l(w_h)-l_h(w_h)|}{\|w_h\|_V} \tag{6.5}$$

(Note: The constant C typically involves α and the continuity of A.)

Proof By the triangle inequality, we have for any $v_h \in V_h$:

$$\|u - \widetilde{u}_h\|_V \leq \|u - v_h\|_V + \|v_h - \widetilde{u}_h\|_V \tag{6.6}$$

Next, we estimate the second term $\|v_h - \widetilde{u}_h\|_V$. Let $w_h = \widetilde{u}_h - v_h$. Using the uniform ellipticity of A_h:

$$\alpha \|w_h\|_V^2 \leq A_h(w_h, w_h) = A_h(\widetilde{u}_h - v_h, w_h) \tag{6.7}$$

Using the discrete equation $A_h(\tilde{u}_h, w_h) = l_h(w_h)$:

$$A_h(\tilde{u}_h - v_h, w_h) = l_h(w_h) - A_h(v_h, w_h) \tag{6.8}$$

We add and subtract terms to introduce the continuous forms $A(\cdot,\cdot)$ and $l(\cdot)$. Note that $A(u, w_h) = l(w_h)$ (exact solution consistency). Thus:

$$\begin{aligned}\alpha \|w_h\|_V^2 &\leq l_h(w_h) - A_h(v_h, w_h) \\ &= \underbrace{A(u, w_h) - A(v_h, w_h)}_{A(u - v_h, w_h)} + \underbrace{A(v_h, w_h) - A_h(v_h, w_h)}_{\text{Consistency Error in } A} + \underbrace{l_h(w_h) - l(w_h)}_{\text{Consistency Error in } l}\end{aligned}$$

Using the boundedness of A ($|A(u - v_h, w_h)| \leq \|A\| \|u - v_h\|_V \|w_h\|_V$):

$$\alpha \|w_h\|_V^2 \leq \|A\| \|u - v_h\|_V \|w_h\|_V + |A(v_h, w_h) - A_h(v_h, w_h)| + |l(w_h) - l_h(w_h)| \tag{6.9}$$

Dividing by $\|w_h\|_V$ (assuming $w_h \neq 0$):

$$\alpha \|\tilde{u}_h - v_h\|_V \leq \|A\| \|u - v_h\|_V + \frac{|A(v_h, w_h) - A_h(v_h, w_h)|}{\|w_h\|_V} + \frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_V} \tag{6.10}$$

Combining this with the initial triangle inequality:

$$\|u - \tilde{u}_h\|_V \leq \left(1 + \frac{\|A\|}{\alpha}\right) \|u - v_h\|_V + \frac{1}{\alpha} \frac{|A(v_h, w_h) - A_h(v_h, w_h)|}{\|w_h\|_V} + \frac{1}{\alpha} \frac{|l(w_h) - l_h(w_h)|}{\|w_h\|_V} \tag{6.11}$$

Taking the infimum over $v_h \in V_h$ and the supremum over $w_h \in V_h$ yields the lemma.

6.4 Linear Iterative Methods

Let us consider solving the linear system:

$$Ax = f \tag{6.12}$$

where A is a symmetric positive-definite (SPD) matrix.

6.4.1 The Picard Iteration

We start with the simplest iterative method for solving (6.12). Given an initial iterate x^0, we set:

$$x^{n+1} = x^n + \tau(f - Ax^n), \quad n = 0, 1, \ldots \tag{6.13}$$

Let the error be defined as $e^n = x - x^n$. Then it is immediate that:

$$e^{n+1} = (I - \tau A) e^n.$$

Let $\|I - \tau A\|$ denote the operator norm induced by the Euclidean norm $\|\cdot\|_2$, namely:

$$\|I - \tau A\| = \sup_{v \in \mathbb{R}^N} \frac{\|(I - \tau A)v\|}{\|v\|} = \sup_{\lambda \in \sigma(A)} |1 - \tau\lambda|.$$

Here $\sigma(A)$ denotes the spectrum of A, and λ_1, λ_N are respectively the smallest and largest eigenvalues of A. It follows that:

$$\|e^n\| \leq \|I - \tau A\| \|e^{n-1}\| \leq \cdots \leq \|I - \tau A\|^n \|e^0\|.$$

The convergence rate is determined by:

$$\|I - \tau A\| = 1 - \frac{\lambda_1}{\lambda_N} = 1 - \frac{1}{\kappa(A)}, \tag{6.14}$$

where $\kappa(A) = \frac{\lambda_N}{\lambda_1}$ is known as the spectral condition number of A.

Clearly, the bound is minimized when we choose $\tau = \frac{2}{\lambda_1 + \lambda_N}$, which gives:

$$\|I - \tau A\| = \frac{\lambda_N - \lambda_1}{\lambda_N + \lambda_1} = \frac{\kappa(A) - 1}{\kappa(A) + 1}. \tag{6.15}$$

Note that this is not much of an improvement over the simple choice of $\tau = \lambda_N^{-1}$ when $\kappa(A)$ is large.

6.4.2 Preconditioning

From the above discussion, it is clear that even with the best choice of τ, the scheme will converge slowly if the condition number of A is large.

The idea of preconditioning is to apply a "preconditioner" to the system and then iterate. The rate of convergence can be improved by the use of an appropriately defined preconditioner.

More precisely, let B be another symmetric and positive definite operator. Consider applying the iteration to the preconditioned equation:

$$BAx = Bf.$$

The iteration becomes:

$$x^{n+1} = x^n + \tau B(f - Ax^n). \tag{6.16}$$

Let $e^n = x - x^n$. Then $e^n = (I - \tau BA)e^{n-1}$. We analyze this using the A-inner product:

$$(u, v)_A = (Au, v), \quad \forall u, v \in \mathbb{R}^N.$$

The associated norm is $\|\cdot\|_A = (\cdot,\cdot)_A^{1/2}$. Note that BA is symmetric with respect to $(\cdot,\cdot)_A$ and is positive definite. The analysis is syntactically similar to the previous discussion but using the norm $\|\cdot\|_A$:

$$\|I - \tau BA\|_A = \sup_{v \in \mathbb{R}^N} \frac{\|(I - \tau BA)v\|_A}{\|v\|_A} = \sup_{\hat{\lambda} \in [\hat{\lambda}_1, \hat{\lambda}_N]} |1 - \tau\hat{\lambda}|,$$

where $\hat{\lambda}_1, \hat{\lambda}_N$ are respectively the smallest and largest eigenvalues of BA. The convergence rate can be estimated in terms of $\kappa(BA)$ as before.

If τ is chosen optimally, the spectral radius is $\rho(I - \tau BA) = \max_i |1 - \tau\hat{\lambda}_i|$. The iteration converges if and only if $|1 - \tau\hat{\lambda}_i| < 1$, which implies $0 < \tau\hat{\lambda}_i < 2$.

6.5 The Conjugate Gradient Method

6.5.1 Krylov Space and Minimization Problem

We define the Krylov space $\mathcal{K}_n(A, r^0) = \text{span}\{r^0, Ar^0, \ldots, A^{n-1}r^0\}$. Consider the minimization problem:

$$J(x) = \min_{y \in \mathbb{R}^N} J(y),$$

where the functional is defined as:

$$J(y) = \frac{1}{2}(Ay, y) - (f, y) = \frac{1}{2}(A(x - y), (x - y)) - \frac{1}{2}(Ax, x).$$

Successive approximations to the solution can be constructed by minimizing $J(y)$ over the Krylov space. Given x^0, let $r^0 = f - Ax^0$. Then $x^n = x^0 + \delta^n$ with $\delta^n \in \mathcal{K}_n$ being the minimizer of $J(y)$ over the affine space $\{x^0\} + \mathcal{K}_n$.

6.5.2 Algorithms

6.5.2.1 Algorithm (Standard CG)

Given x^0 arbitrary. Set $r^0 = f - Ax^0$ and $p^0 = r^0$. For $n = 0, 1, \ldots$ until convergence, compute:

1. $x^{n+1} = x^n + \alpha_n p^n$, where $\alpha_n = \frac{(r^n, p^n)}{(Ap^n, p^n)}$
2. $r^{n+1} = r^n - \alpha_n A p^n$
3. $p^{n+1} = r^{n+1} + \beta_n p^n$, where $\beta_n = -\frac{(Ap^n, r^{n+1})}{(Ap^n, p^n)}$

6.5.2.2 Algorithm (PCG)

Given x^0 arbitrary. Set $r^0 = f - Ax^0$ and $z^0 = Br^0$, $p^0 = z^0$. For $n = 0, 1, \ldots$ until convergence, compute:

1. $x^{n+1} = x^n + \alpha_n p^n$, where $\alpha_n = \frac{(r^n, z^n)}{(Ap^n, p^n)}$
2. $r^{n+1} = r^n - \alpha_n A p^n$
3. $z^{n+1} = Br^{n+1}$
4. $p^{n+1} = z^{n+1} + \beta_n p^n$, where $\beta_n = \frac{(r^{n+1}, z^{n+1})}{(r^n, z^n)}$

6.6 Convergence Analysis

Theorem: Let $e^n = x - x^n$. For CG:

$$\|e^n\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \|e^0\|_A, \quad \text{where } \kappa = \kappa(A). \tag{6.17}$$

For PCG:

$$\|e^n\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \|e^0\|_A, \quad \text{where } \kappa = \kappa(BA). \tag{6.18}$$

Remark: Note that $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} < \frac{\kappa-1}{\kappa+1}$, so CG converges faster than the simple iterative method.

6.7 Model Problem: Finite Element Application

Consider the problem:

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \tag{6.19}$$

The weak formulation is: Find $u \in H_0^1(\Omega)$ such that:

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv, \quad \forall v \in H_0^1(\Omega).$$

6.7.1 Multilevel Finite Element Constructions

Let $\{\mathcal{T}_k\}_{k=1}^{\infty}$ be nested triangulations obtained by subdividing each triangle into four pieces by connecting midpoints of edges. Let V_k be the continuous linear Finite Element space of $\mathcal{T}_k$. So,

$$V_1 \subset V_2 \subset V_3 \subset \cdots \subset V_J \subset H_0^1(\Omega).$$

Let h_k be the mesh size of $\mathcal{T}_k$. We denote the corresponding linear system $A_k U = F$, where:

$$(A_k)_{ij} = \int_\Omega \nabla\phi_j \cdot \nabla\phi_i, \quad F_j = \int_\Omega f\phi_j.$$

Here $U = (u_1, \ldots, u_N)$ so that $u_h = \sum_{i=1}^N u_i\phi_i$.

For each $T \in \mathcal{T}_k$, we note that for v_h linear in T:

$$\|v_h\|^2_{L^2(T)} \approx h_k^2 \sum_{i=1}^{3} v_h^2(v_i),$$

where v_i are the vertices of T. Summing over triangles:

$$\|v_h\|^2_{L^2(\Omega)} \sim h^2 \sum_{i=1}^{N} v_i^2.$$

By Poincaré inequality:

$$h^2 \sum_{i=1}^{N} v_i^2 \leq C\|v_h\|^2_{L^2(\Omega)} \leq C \int_\Omega |\nabla v_h|^2 = CV^T A_k V.$$

Thus, $V^T A_k V \geq Ch_k^2 |V|^2$. Since A_k is SPD, this means A_k's smallest eigenvalue is bounded below by Ch^2.

On the other hand, it can be shown that the largest eigenvalue of A_k is uniformly bounded above. Thus, the condition number behaves as:

$$\kappa(A_k) = \frac{\lambda_N}{\lambda_1} \leq Ch_k^{-2}. \tag{6.20}$$

One can also show that $\kappa(A_k) \geq ch_k^{-2}$ (Exercise). This means the problem is rather ill-conditioned, and iterative methods will converge slowly.