Numerical Analysis and Partial Differential Equations

March 12, 2010

Contents

Part I

Background

Chapter 1

Model Problems

Here is a small list of boundary and initial value problems for model partial differential equations.

1.1 Boundary Value Problems

Let $\Omega \subset \mathbb{R}^d$ be a bounded open set. Let $q, f \in C(\bar{\Omega}, \mathbb{R})$, $b \in (C(\bar{\Omega}, \mathbb{R}))^d$ and $g \in C(\partial\Omega, \mathbb{R})$. We study the **elliptic equation**

$$\begin{aligned} -\Delta u + b.\nabla u + qu &= f, \quad x \in \Omega, \\ u &= g, \quad x \in \partial\Omega. \end{aligned}$$

In our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in Ω and the equation holds pointwise.

When studying finite element methods we will study the weak or variational form of this problem which is defined from (1.1) as follows, in the case $g = 0$. Let $\mathcal{V} = H_0^1(\Omega)$ and define the bilinear form and L^2 inner-product, respectively, by

$$\begin{aligned} a(u, v) &= \int_\Omega [\nabla u \cdot \nabla v + (b \cdot \nabla u)v + quv]\, dx, \\ \langle u, v \rangle &= \int_\Omega uv dx. \end{aligned} \tag{1.1}$$

Then the weak formulation of (1.1) is to find

$$u \in \mathcal{V} : \; a(u, v) = \langle f, v \rangle \quad \forall v \in \mathcal{V}. \tag{1.2}$$

Under certain regularity assumptions on f this variational formulation is equivalent to the original strong form of the problem for classical solutions.

1.2 Transport/Advection

Let $c \in C(\mathbb{R}^+, \mathbb{R}^+)$. The **transport problem** is then:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} & = & 0 & (x,t) \in (0,\infty) \times (0,\infty), \\ u & = & g & (x,t) \in [0,\infty) \times \{0\}, \\ u & = & h & (x,t) \in \{0\} \times (0,\infty). \end{array} \tag{1.3}$$

In our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in $(0,\infty)\times(0,\infty)$ and the equation holds pointwise.

Let $c \in C([-,1,1], \mathbb{R}^+)$ be $2-$periodic. The **periodic transport problem** is then:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} & = & 0 & (x,t) \in (-1,1) \times (0,\infty), \\ u(-1,t) & = & u(1,t) & t \in (0,\infty), \\ u & = & g & (x,t) \in [-1,1] \times \{0\}. \end{array} \tag{1.4}$$

Again, in our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in $(-1,1) \times (0,\infty)$ and the equation holds pointwise.

Let $c \in C([-1,1], \mathbb{R}^+)$ be $2-$periodic. The (second order) **wave equation** is

$$\begin{array}{rcll} \frac{\partial^2 u}{\partial t^2} & = & \frac{\partial}{\partial x}[c^2\frac{\partial u}{\partial x}], & (x,t) \in (-1,1) \times (0,\infty), \\ u(-1,t) & = & u(1,t) & t \in (0,\infty), \\ \frac{\partial u}{\partial x}(-1,t) & = & \frac{\partial u}{\partial x}(1,t) & t \in (0,\infty), \\ u & = & g & (x,t) \in [-1,1] \times \{0\}, \\ \frac{\partial u}{\partial t} & = & h & (x,t) \in [-1,1] \times \{0\}. \end{array} \tag{1.5}$$

Again, in our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in $(-1,1) \times (0,\infty)$ and the equation holds pointwise.

1.3 Diffusion

Let $f, g \in C(\bar{\Omega}, \mathbb{R})$. The **heat equation** on an open bounded set $\Omega \subset \mathbb{R}^d$ is:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & \Delta u + f & (x,t) \in \Omega \times (0,\infty), \\ u & = & 0 & (x,t) \in \partial\Omega \times (0,\infty), \\ u & = & g & (x,t) \in \bar{\Omega} \times \{0\}. \end{array} \tag{1.6}$$

In our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in $\Omega \times (0,\infty)$ and the equation holds pointwise.

When studying finite element methods we will study the weak or variational form of this problem which is defined as follows. Define the bilinear form $a(u,v)$ and L^2 inner-product as in (1.1) with $p = q = 0$ and define $\mathcal{V} = H_0^1(\Omega)$ as before. Then we seek

$$\begin{array}{rcl} u \in C^1([0,\infty), \mathcal{V}) : \ \langle \frac{\partial u}{\partial t}, v \rangle + a(u,v) & = & \langle f, v \rangle \ \forall v \in \mathcal{V} \\ u & = & g, \ t = 0. \end{array} \tag{1.7}$$

Under certain regularity assumptions on f and g this variational formulation is equivalent to the original strong form of the problem for classical solutions.

On occasion it will be of interest to study the **periodic heat equation**, namely

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & \frac{\partial^2 u}{\partial x^2} + f & (x,t) \in (-1,1) \times (0,\infty), \\ u(-1,t) & = & u(1,t) & t \in (0,\infty), \\ \frac{\partial u}{\partial x}(-1,t) & = & \frac{\partial u}{\partial x}(1,t) & t \in (0,\infty), \\ u & = & g & (x,t) \in [-1,1] \times \{0\}. \end{array} \tag{1.8}$$

Again, in our study of finite difference methods we will study classical solutions of this problem, where all derivatives appearing in the equation exist everywhere in $(-1,1) \times (0,\infty)$ and the equation holds pointwise.

1.4 Dispersion

Let $g \in C([-1,1], \mathbb{R})$ be $2-$periodic. The **periodic Schrödinger equation** on $(-1,1)$ is:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & i\frac{\partial^2 u}{\partial x^2} & (x,t) \in (-1,1) \times (0,\infty), \\ u(-1,t) & = & u(1,t) & t \in (0,\infty), \\ \frac{\partial u}{\partial x}(-1,t) & = & \frac{\partial u}{\partial x}(1,t) & t \in (0,\infty), \\ u & = & g & (x,t) \in [-1,1] \times \{0\}. \end{array} \tag{1.9}$$

Chapter 2

Overview

2.1 A Boundary Value Problem

Consider the 1D boundary value problem: (D) given $f \in C(0,1)$, find $u \in C^2(0,1)$ such that

$$-u'' = f, \tag{2.1}$$
$$u(0) = u(1) = 0. \tag{2.2}$$

(D) indicates that this is a Dirichlet problem, i.e. the boundary values of u are prescribed. "Clearly" a solution exists to this boundary value problem:

$$u(x) = \int_0^x \int_0^s f(t)\ dt\, ds + ax + b. \tag{2.3}$$

Is this solution unique? Suppose u_1, u_2 solve the boundary value problem (D). Set $v(x) = u_1 - u_2$. Then $v(0) = v(1) = 0$ and $v''(x) = u_1''(x) - u_2''(x) = 0$. Thus

$$v'' = 0 \qquad \text{in } (0,1), \tag{2.4}$$
$$v(0) = v(1) = 0. \tag{2.5}$$

Solving this yields $v(x) = cx+d$ and using the boundary conditions gives us $v(x) = 0$, hence the solution to our original boundary value problem is unique.

Remark. 1. Existence of solutions to (BVP) for a (PDE) generally require some analysis. Function spaces and functional analysis.

2. Uniqueness can usually be proved by contradiction along the lines above.

3. Well-posedness is important, especially when we are interested in approximation. Well-posedness requires (i) existence; (ii) uniqueness; (iii) continuity with respect to data ($x_j \to x^* \Rightarrow f(x_j) \to f(x^*)$.

Some applications of (D):

1. Consider an elastic bar fixed at both ends subjected to a tangential load (i.e. parallel to the rod). Let $u(x)$ be the displacement tangential to the rod, $\sigma(x)$ be the traction. By linear elasticity

$$\sigma = Eu' \qquad \text{Hooke's law,} \tag{2.6}$$
$$-\sigma' = f \qquad \text{force balance/equilibrium equation,} \tag{2.7}$$
$$u(0) = u(1) = 0. \tag{2.8}$$

   E is the material constant, Young's modulus. Then $Eu'' = f, f \in I := (0,1)$, $u(0) = u(1) = 0$. If $E = 1$ we retrieve (D).

2. Consider the one dimensional heat equation

$$u_t = ku_{xx} + f, x \in I, \tag{2.9}$$
$$u(0) = u(1) = 0, \tag{2.10}$$

   where u s the temperature, k a material constant (thermal diffusivity) and f is a body heat source, t is 'time'. Then if we look for a steady state problem $u_t = 0$ and

$$-ku_{xx} = f, \qquad x \in I, \tag{2.11}$$
$$u(0) = u(1) = 0. \tag{2.12}$$

2.2 A Finite Element Method for the Two-Point BVP (D)

Let $\{x_j\}_{j=0}^{M+1}$ be a partition of the unit interval $\overline{I} = [0,1]$ so that $0 = x_0 < x_1 < x_2 < \ldots < x_j < x_{j+1} < \ldots < x_M < x_{M+1} = 1$. Let $I_j = (x_{j-1}, x_j), j = 1, \ldots, M+1$ and define $|I_j| = h_j$ and $h = \max_{j=1,\ldots,M+1} h_j$. Then h is said to be the "mesh" size and indicates how fine the mesh is. This mesh may be non-uniform, i.e. the h_j are non-uniform.

Actually each subinterval I_j is an element and h_j is the element size. We have divided I up into elements.

Now define V_h to be the finite dimensional space of functions

$$V_h := \{v_h \in C[0,1] : v_h(0) = v_h(1) = 0, v_h|_{I_j} \text{ is linear } \forall j \in \{1, \ldots, M+1\}, \quad (2.13)$$

i.e. V_h consists of continuous piecewise linear functions. V_h is determined by its value at the internal mesh points (nodes) $x_j, j = 1, \ldots M$ and $\dim(V_h) = M$. Clearly V_h is a linear space ($v_h, w_h \in V_h \Rightarrow \alpha v_h + \beta w_h \in V_h \; \forall \, \alpha, \beta \in \mathbb{R}$).

V_h is the simplest finite element space.

Since V_h is a linear space it must have a basis. Introduce the basis functions $\{\phi_j(x)\}_{j=1}^M, \phi_j \in V_h$ such that for $i = 0, 1, 2, ..., M, M+1$,

$$\phi_j(x_i) = \delta_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \quad (2.14)$$

These ϕ_j are 'hat' functions. Note we can write these as

$$\phi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h_j} & x \in I_j, \\ \frac{x_{j+1} - x}{h_{j+1}} & x \in I_{j+1}, \\ 0 & \text{otherwise}, \end{cases} \quad (2.15)$$

because ϕ_j is a continuous piecewise linear function. The $\{\phi_j(x)\}_{j=1}^M$ are linearly independent and any $v_h \in V_h$ can be written uniquely as $v_h(x) = \sum_{j=1}^M \alpha_j \phi_j(x)$ and

$$v_h(x_i) = \sum_{j=1}^M \alpha_j \phi_j(x_i) = \alpha_i.$$

Proof: Consider $w(x) = \sum_{j=1}^M w_j \phi_j(x)$. Then $w(x) = 0 \, \forall \, x \in I \Leftrightarrow w_j = 0, j = 1, \ldots, M$ (since $w(x_j) = \sum_{j=1}^M w_j \phi_j(x_j) = w_j$).

■

Since these coefficients are the values of v_h at the nodes we say that $\{\phi_j\}$ is a 'nodal' basis.

V_h is a finite dimensional function space. We seek solutions of PDE's in infinite dimensional function spaces and approximate the problem by seeking our approximation in a finite dimensional space. We wish to find an approximation $u_h \in V_h$ to the solution u of (D); of course u will not be piecewise linear!! There will be an approximation error $e_h(x) = u(x) - u_h(x)$. We would like e_h to be small and indeed as $h \to 0$ we hope that $e_h(x) \to 0$ in some sense.

How to find an approximation $u_h \in V_h$?

2.3 Finite Difference Method for (D)

Finite differences can be used – replace derivatives by finite differences. Consider problem (D) discretised with a uniform grid so $x_j = jh$. U_j is the approximation to $u(x_j)$, in general $U_j \neq u(x_j)$. The discrete problem (D_h) is

$$-\frac{U_{j-1} - 2U_j + U_{j+1}}{h^2} = f(x_j), \tag{2.16}$$

$$U_0 = U_M = 0. \tag{2.17}$$

This can be derived formally by Taylor series:

$$u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2}u''(x) \pm \frac{h^3}{3!}u'''(x) + \frac{h^4}{4!}u^{(4)}(x) \pm \dots \tag{2.18}$$

so

$$\frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + \frac{h^2}{24}(u^{(4)}(\eta^+) + u^{(4)}(\eta^-)) \tag{2.19}$$

$$= u''(x) + \frac{h^2}{12}u^{(4)}(\eta(x)) \tag{2.20}$$

$$\frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2} = u''(x_j) + \frac{h^2}{12}u^{(4)}(\eta_j) \tag{2.21}$$

$$= -f(x_j) + \frac{h^2}{12}u^{(4)}(\eta_j) \tag{2.22}$$

Rearranging, the solution of u satisfies

$$-\left(\frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2}\right) - f(x_j) = -\frac{h^2}{12}u^{(4)}(\eta_j) \tag{2.23}$$

This suggest replacing $u(x_j)$ by U_j and neglecting the right hand side to obtain (D_h).

2.4 Finite Element Method

Since V_h does not have functions that can be differentiated twice we need a different formulation of (D). Introduce the L^2 inner product:

$$(u, v) := \int_I u(x)v(x)\ dx. \tag{2.24}$$

For the moment suppose that functions are continuous and the differential equations hold for every $x \in I$.

$$-u'' - f = 0 \qquad x \in I \tag{2.25}$$

$$\Rightarrow \quad (-u'' - f, v) = 0 \qquad \forall\, v \tag{2.26}$$

$$\Rightarrow \quad (-u'', v) = (f, v). \tag{2.27}$$

Use integration by parts

$$(u', v') = (f, v) - [u'v]_0^1. \tag{2.28}$$

Set $V_C := \{v \in C[0,1] : v(0) = v(1) = 0\}$. If u solves (D) then $u \in V_C$ and the problem (P) is to find u such that:

$$(u', v') = (f, v) \qquad \forall\, v \in V_C. \tag{2.29}$$

We do not know that this problem has a solution or if it is unique. We need to impose a structure which has existence and uniqueness (this is a question in functional analysis). In order to do this we look for a solution in a larger space. We also need the integrals to make sense – this is a question of function spaces.

The finite element approximation of (D) is defined by approximating the variational problem (2.29). This leads us to pose the problem (P_h): find $u_h \in V_h$ such that

$$(u_h', v_h') = (f, v_h) \qquad \forall\, v_h \in V_h. \tag{2.30}$$

Note that $u_h(x) = \sum_{i=1}^M u_i\phi_i(x)$ and the problem of finding u_h is equivalent to finding the M numbers $\{u_i\}_{i=1}^M$.

Now choose $v_h = \phi_j, j = 1, \ldots, M$ and use the definition of u_h:

$$\left(\left(\sum_{i=1}^M u_i\phi_i\right)', \phi_j'\right) = (f, \phi_j) \qquad j = 1, \ldots, M \tag{2.31}$$

$$\Rightarrow \sum_{i=1}^M u_i(\phi_i', \phi_j') = (f, \phi_j) \qquad j = 1, \ldots, M. \tag{2.32}$$

Set $A_{ij} = (\phi_i', \phi_j') = (\phi_j', \phi_i'),\ i, j = 1, \ldots, M$. Thus A is a symmetric matrix. Set $b_j = (f, \phi_j), j = 1, \ldots, M$. Then we are trying to solve

$$\sum_{i=1}^{M} u_i(\phi_i', \phi_j') = (f, \phi_j) \qquad j = 1, \ldots, M \tag{2.33}$$

$$\Rightarrow \sum_{i=1}^{M} A_{ij} u_i = b_j \qquad j = 1, \ldots, M \tag{2.34}$$

$$\Leftrightarrow A\mathbf{u} = \mathbf{b}, \tag{2.35}$$

where $\mathbf{u} = (u_1, \ldots, u_M)^T$ and $\mathbf{b} = (b_1, \ldots, b_M)^T \in \mathbb{R}^M$.

Thus the finite element method applied to (D) leads to a system of linear equations for the coefficients $(u_1, \ldots, u_M)$ in $u_h = \sum_{i=1}^{M} u_i \phi_i$ where for V_h being piecewise linear $u_i = u(x_i)$.

We can easily discover more information about A:

1. We already know that

$$\phi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h_j} & x \in I_j, \\ \frac{x_{j+1} - x}{h_{j+1}} & x \in I_{j+1}, \\ 0 & \text{otherwise}, \end{cases} \tag{2.36}$$

so

$$\phi_j'(x) = \begin{cases} \frac{1}{h_j} & x \in I_j, \\ \frac{-1}{h_{j+1}} & x \in I_{j+1}, \\ 0 & \text{otherwise}. \end{cases} \tag{2.37}$$

Thus $\phi_j \in C(I), \phi_j' \notin C^1(I)$ but $\phi_j' \in L^2(I)$, i.e. $\int_I (\phi_j')^2\ dx\ < \infty$, and

$$\int_I (\phi_j')^2\ dx\ = \int_{x_{j-1}}^{x_{j+1}} (\phi_j')^2\ dx\ = \int_{x_{j-1}}^{x_j} \frac{1}{h_j^2}\ dx\ + \int_{x_j}^{x_{j+1}} \frac{1}{h_{j+1}^2}\ dx\ = \frac{1}{h_j} + \frac{1}{h_{j+1}}, \tag{2.38}$$

$$(\phi_i', \phi_j') = \int_I \phi_i'(x)\phi_j'(x)\ dx\ = \int_{x_{j-1}}^{x_{j+1}} \phi_i'(x)\phi_j'(x)\ dx\ = 0 \qquad |i - j| > 1, \tag{2.39}$$

$$(\phi_j', \phi_{j-1}') = \int_{x_{j-1}}^{x_j} \phi_j'(x)\phi_{j-1}'(x)\ dx\ = \int_{x_{j-1}}^{x_j} \frac{1}{h_j} \cdot \frac{-1}{h_j}\ dx\ = \frac{-1}{h_j}. \tag{2.40}$$

By the symmetry of A we have found a formula $\forall\, A_{ij}$.

2. In general

$$A = \begin{bmatrix} \frac{1}{h_1}+\frac{1}{h_2} & \frac{-1}{h_2} & 0 & \cdots & & 0 \\ \frac{-1}{h_2} & \frac{1}{h_2}+\frac{1}{h_3} & \frac{-1}{h_3} & 0 & \cdots & 0 \\ 0 & \frac{-1}{h_3} & \frac{1}{h_3}+\frac{1}{h_4} & \frac{-1}{h_4} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ & & & \text{etc} & & & \end{bmatrix} \tag{2.41}$$

which we note has a positive diagonal and a negative off-diagonal.

3. A is positive definite! Let $\mathbf{v} \in \mathbb{R}^M$. Set $v_k = \sum_{i=1}^M v_i \phi_i(x)$. Then

$$(v_k', v_k') = \left(\sum_{i=1}^M v_i \phi_i', \sum_{j=1}^M v_j \phi_j' \right) = \sum_{i=1}^M \sum_{j=1}^M v_i v_j (\phi_i', \phi_j') = \mathbf{v}^T A \mathbf{v}. \tag{2.42}$$

Observe that $\mathbf{v}^T A \mathbf{v} = \int_I (v_k')^2 \, dx \geq 0 \, \forall \, \mathbf{v} \in \mathbb{R}^M$ and $\mathbf{v}^T A \mathbf{v} = 0 \Leftrightarrow \int_I (v_k')^2 \, dx = 0 \Leftrightarrow (v_k')^2 = 0 \; \forall \, x \Leftrightarrow v_k' = 0 \Rightarrow v_k(x) = v_k(0) + \int_0^x v_k'(s) \, ds = v_k(0) = 0 \; \forall \, x \Rightarrow v_i = v_k(x_i) = 0 \; \forall \, i = 1, \ldots, M \Rightarrow \mathbf{v} = 0$.

4. In the special case when we have a uniform mesh (so that $h_j = h = 1/(M+1)$):

$$A = \frac{1}{h} \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \tag{2.43}$$

This matrix occurs in finite difference approximations.

5. The matrix A is sparse, i.e. most of its elements are 0! This is an important property which has to be exploited in computations. In fact it is an important motivation for using finite element spaces.

The computational problem is (i) to find the matrix A and right hand side $\mathbf{b}$, (ii) solve $A\mathbf{u} = \mathbf{b}$. (Remember that in fact we are trying to solve PDE's, and the finite element method is an approximation method for doing this!)

Part II

Finite Element Methods

Chapter 3

Abstract variational problems

3.1 Variational Problems

Let V be a normed real vector space with norm $\|\cdot\|_V$. Let $\{u_m\}_{m=1}^{\infty} \subset V$ be a sequence.

If the sequence satisfies

$$\lim_{n,m\to\infty} \|u_m - u_n\|_V = 0,$$

then the sequence is said to be *Cauchy*.

If V is such that for any Cauchy sequence $\{u_m\}_{m=1}^{\infty}$ there exists $u \in V$ satisfying

$$\lim_{n\to\infty} \|u_n - u\|_V = 0 \text{ (i.e, } \lim_{n\to\infty} u_n = u),$$

then V is said to be a *Banach space*.

Let $l(\cdot) : V \to \mathbb{R}$ be a linear functional or linear form[1], i.e,

$$l(\alpha v + \beta w) = \alpha l(v) + \beta l(w),\ \forall \alpha, \beta \in \mathbb{R},\ v, w \in V,$$

then $l(\cdot)$ is said to be *bounded* if

$$\exists c_l \in \mathbb{R} \text{ s.t } |l(v)| \leq c_l \|v\|_V,\ \forall v \in V.$$

Remark. We also say that a linear functional $l(\cdot)$ is *continuous* when it is bounded.

[1] Note that if $l(\cdot) : V \to \mathbb{R}$ is linear, taking $\alpha = \beta = 0$ yields $l(0) = 0$.

Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$. Then $a(\cdot,\cdot)$ is *bilinear* if for $\forall \alpha, \beta \in \mathbb{R},\ u, v, w \in V$

$$a(\alpha v + \beta w, u) = \alpha a(v, u) + \beta a(w, u),$$
$$a(u, \alpha v + \beta w) = \alpha a(u, v) + \beta a(u, w).$$

We say the bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is *symmetric* if

$$a(u, w) = a(w, u)\ \forall u, w \in V.$$

Let $\langle \cdot,\cdot \rangle : V \times V \to \mathbb{R}$ be a symmetric bilinear form on V satisfying

$$\langle v, v \rangle \geq 0\ \forall v \in V,$$
$$\langle v, v \rangle = 0 \Leftrightarrow v = 0,$$

then $\langle \cdot,\cdot \rangle$ is said to be an *inner product* on V.

Lemma 3.1.1 (Cauchy-Schwarz inequality)**.** *Let $\langle \cdot,\cdot \rangle$ be an inner product on V. Then*

$$|\langle u, v \rangle| \leq \sqrt{\langle u, u \rangle}\sqrt{\langle v, v \rangle}\ \forall u, v \in V.$$

Proof.

$$\begin{aligned}
0 &\leq \langle u + \lambda v, u + \lambda v \rangle \\
&= \langle u, u + \lambda v \rangle + \lambda \langle v, u + \lambda v \rangle \text{ (by linearity w.r.t 1st variable)} \\
&= \langle u, u \rangle + \lambda \langle u, v \rangle + \lambda \langle v, u \rangle + \lambda^2 \langle v, v \rangle \text{ (by linearity w.r.t 2nd variable)} \\
&= \langle u, u \rangle + 2\lambda \langle u, v \rangle + \lambda^2 \langle v, v \rangle \text{ (by symmetry)} \\
&:= Q(\lambda).
\end{aligned}$$

If $v = 0$, the required inequality is obvious since the both sides are zero. Suppose $v \neq 0$. Set

$$a = \langle v, v \rangle, b = \langle u, v \rangle, c = \langle u, u \rangle.$$

Noting

$$Q(\lambda) = a\lambda^2 + 2b\lambda + c \geq 0,\ a > 0,$$

we require a non-positive discriminant $b^2 \leq ac$, which is

$$\langle u, v \rangle^2 \leq \langle u, u \rangle \langle v, v \rangle.$$

□

Given a vector space V with an inner product $\langle \cdot, \cdot \rangle$ we can set

$$\|v\|_V := \sqrt{\langle v, v \rangle} \ \forall v \in V$$

to be a norm.

Lemma 3.1.2. *$\| \cdot \|_V$ defines a norm on V.*

Proof. i) $\|v\|_V \geq 0 \ \forall v \in V$.

ii) $\|v\|_V = 0 \Leftrightarrow v = 0$.

iii) $\|\lambda v\|_V = \sqrt{\langle \lambda v, \lambda v \rangle} = \sqrt{\lambda^2 \langle v, v \rangle} = |\lambda| \|v\|_V$.

iv) Triangle inequality;

$$\begin{aligned}
\|u + v\|_V &= \sqrt{\langle u + v, u + v \rangle} \\
&= \sqrt{\langle u, u \rangle + 2\langle u, v \rangle + \langle v, v \rangle} \\
&\leq \sqrt{\|u\|_V^2 + 2\|u\|_V \|v\|_V + \|v\|_V^2} \\
&\qquad \text{(by Cauchy-Schwarz inequality)} \\
&= \sqrt{(\|u\|_V + \|v\|_V)^2} \\
&= \|u\|_V + \|v\|_V.
\end{aligned}$$

□

Suppose V is an inner product space with $\langle v, v \rangle \ \forall v \in V$ and suppose V is a Banach space with the norm $\|v\|_V = \sqrt{\langle v, v \rangle} \ \forall v \in V$. Then V is a *Hilbert space*.

Remark. We say that the normed vector space V is *complete* if all Cauchy sequences converge in V, i.e, a Banach space is a complete normed vector space, a Hilbert space is a complete inner product space.

Let V^* be the space of all bounded linear functionals on V. Then V^* is a linear space. Indeed, for $l_1, l_2 \in V^*$ define

$$l(v) := \alpha l_1(v) + \beta l_2(v) \ \forall v \in V, \alpha, \beta \in \mathbb{R}.$$

Then l is a bounded linear functional on V, thus $l \in V^*$. This implies that V^* is a linear space.

For V^* we use the following norm;

$$\|l\|_{V^*} := \sup_{v \in V, v \neq 0} \frac{l(v)}{\|v\|_V} \ \forall v \in V.$$

Clearly

$$\begin{aligned} \|l\|_{V^*} &\geq \frac{l(v)}{\|v\|_V}, \\ \Rightarrow |l(v)| &\leq \|l\|_{V^*} \|v\|_V. \end{aligned}$$

Hence $\|l\|_{V^*}$ is the least upper bound of all c_l such that $|l(v)| \leq c_l \|v\|_V \ \forall v \in V$.

Consider $\{v_m\}_{m=0}^{\infty} \subset V$ satisfying $\lim_{m \to \infty} v_m = v \in V$ (i.e. $\lim_{m \to \infty} \|v_m - v\|_V = 0$). Then,

$$|l(v) - l(v_m)| = |l(v - v_m)| \leq \|l\|_{V^*} \|v - v_m\|_V \to 0 \text{ as } m \to \infty.$$

Thus, a bounded linear functional is a continuous linear functional.

The Abstract Variational Problem

Let V be a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and $\|v\|_V := \sqrt{\langle v, v \rangle} \ \forall v \in V$. Let $l(\cdot) : V \to \mathbb{R}$ be a bounded linear functional. Let $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a bilinear form.

We consider the following variational problem.

(**P**) Find $u \in V$ such that $a(u, v) = l(v) \ \forall v \in V$.

Theorem 3.1.3 (Lax-Milgram)**.** *Let $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a bilinear form such that*

1) $a(\cdot, \cdot)$ is bounded, i.e.,

$$\exists \gamma > 0 \ s.t. \ |a(v, w)| \leq \gamma \|v\|_V \|w\|_V \ \forall v, w \in V.$$

2) $a(\cdot, \cdot)$ is coercive (= V-elliptic), i.e.,

$$\exists \alpha > 0 \ s.t. \ a(v, v) \geq \alpha \|v\|_V^2 \ \forall v \in V.$$

Let $l(\cdot) : V \to \mathbb{R}$ be a bounded linear functional, i.e.,

$$\exists M > 0 \text{ s.t. } |l(v)| \leq M\|v\|_V \ \forall v \in V.$$

*Then the variational problem (**P**) has a unique solution.*

Proof. (Uniqueness)

Suppose u_1, u_2 are two solutions of (**P**).

$$\begin{aligned} &\Rightarrow a(u_1, v) = l(v), \\ &\quad a(u_2, v) = l(v) \ \forall v \in V. \end{aligned}$$

Subtraction and linearity give

$$a(u_2, v) - a(u_1, v) = a(u_2 - u_1, v) = 0 \ \forall v \in V.$$

In particular, choose $v = u_2 - u_1$, then we see

$$a(u_2 - u_1, u_2 - u_1) = 0.$$

By coercivety

$$\begin{aligned} 0 &\geq \alpha\|u_2 - u_1\|_V^2, \\ \Rightarrow\ 0 &\geq \|u_2 - u_1\|_V, \\ \Rightarrow\ u_1 &= u_2. \end{aligned}$$

(Existence)

This is proved by a direct consequence of the Riesz representation theorem (see Functional Analysis). Since we will now fix $v \in V$, then $a(v, \cdot) : V \to \mathbb{R}$ is a bounded linear functional. The Riesz representation theorem says that there uniquely exists a point (which we will call Av) in V such that

$$a(v, w) = \langle Av, w \rangle \ \forall w \in V.$$

Clearly this defines a map $A : V \to V$, which is linear. Furthermore it holds that

$$|\langle Av, w \rangle| = |a(v, w)| \leq \gamma\|v\|_V\|w\|_V \ \forall v, w \in V.$$

Now, take $w = Av$, then

$$\begin{aligned} \text{LHS} = \|Av\|_V^2 &\leq \gamma\|v\|_V\|Av\|_V = \text{RHS} \\ \Rightarrow \|Av\|_V &\leq \gamma\|v\|_V. \end{aligned}$$

This implies that A is a bounded linear operator

$$A : V \to V \text{ with } |A| \leq \gamma,$$

where

$$|A| := \sup_{v \in V, v \neq 0} \frac{\|Av\|_V}{\|v\|} \leq \sup_{v \in V, v \neq 0} \frac{\gamma \|v\|_V}{\|v\|_V} = \gamma.$$

It follows that

$$\begin{aligned} & a(u,v) = l(v) \ \forall v \in V \\ \Leftrightarrow \ & \langle Au, v \rangle = \langle L, v \rangle \ \forall v \in V \\ \Leftrightarrow \ & \langle Au - L, v \rangle = 0 \ \forall v \in V \\ \Leftrightarrow \ & Au = L : \text{ a linear operator equation .} \end{aligned}$$

Now let u be a solution of $(\mathbf{P})$.
$\Leftrightarrow$ u solves $Au = L$.
$\Leftrightarrow$ u solves $u = u - \rho(Au - L), \ \rho > 0$.

Consider

$$\begin{cases} u_{k+1} = u_k - \rho(Au_k - L), \\ u_0 \text{: given.} \end{cases}$$

Then we have an infinite sequence $\{u_k\}_{k=0}^{\infty}$.

Let w_ρ be a map defined by $w_\rho(v) = v - \rho(Av - L)$. Then we see $u_{k+1} = w_\rho(u_k)$ and

$$\begin{aligned} & \|w_\rho(v_2) - w_\rho(v_1)\|_V^2 \\ & \qquad = \|v_2 - v_1\|_V^2 - 2\rho a(v_1 - v_2, v_1 - v_2) + \rho^2 \|A(v_2 - v_1)\|_V^2. \\ & \quad \Rightarrow \|w_\rho(v_2) - w_\rho(v_1)\|_V^2 \leq (1 - 2\rho\alpha + \rho^2 |A|^2) \|v_2 - v_1\|^2. \end{aligned}$$

Choose $\rho \in (0, 2\alpha/|A|^2)$, then w_ρ is a strict contraction. Now the contraction mapping theorem for Hilbert spaces assures that w_ρ has a unique fixed point. $\square$

3.2 Remarks on the Lax-Milgram Result

1. Uniqueness: by our usual methods this follows from the linearity of $l(\cdot)$, the bilinearity of $a(\cdot, \cdot)$ and the coercivity of $a(\cdot, \cdot)$.

2. Stability estimate: we know that

$$c_0\|u\|_V^2 \leq a(u,u) = l(u) \leq c_2\|u\|_V$$

so we can deduce that the solution to our BVP satisfies

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{c_0}\|f\|_{L^2(\Omega)}. \tag{3.1}$$

3. Continuity with repsect to $l(\cdot)$. Consider the two problems

$$\begin{aligned} u_1 \in V \ s.t. \qquad & a(u_1,v) = l_1(v) \qquad \forall\, v \in V \\ u_2 \in V \ s.t. \qquad & a(u_2,v) = l_2(v) \qquad \forall\, v \in V. \end{aligned}$$

Then

$$a(u_1 - u_2, v) = l_1(v) - l_2(v) = \hat{l}(v). \tag{3.2}$$

Choosing $v = u_1 - u_2$:

$$\begin{aligned} c_0\|u_1 - u_2\|_V^2 \leq \hat{l}(u_1 - u_2) &\leq \|l_1 - l_2\|_{V^*}\|u_1 - u_2\| \\ \Rightarrow \|u_1 - u_2\|_V &\leq \frac{\|l_1 - l_2\|_{V^*}}{c_0} \end{aligned}$$

In terms of our original elliptic bvp's we have that

$$\|u_1 - u_2\|_{H^1(\Omega)} \leq \frac{1}{c_0}\|f_1 - f_2\|_{L^2(\Omega)}. \tag{3.3}$$

4. If l is the zero element of V^* (i.e. $l(v) = 0 \ \forall\, v \in V$) then $0 = a(u,u) \Rightarrow \|u\|_V = 0$ by coercivity and $u = 0$.

3.3 Calculus of Variations

Suppose $a(\cdot,\cdot)$ is also symmetric, i.e,

$$a(u,v) = a(v,u) \ \forall u, v \in V.$$

Define $J(\cdot) : V \to \mathbb{R}$ by

$$J(v) = \frac{1}{2}a(v,v) - l(v) \ \forall v \in V.$$

We say that $J(\cdot)$ is a *quadratic functional*.

Now consider the minimization problem;

(**M**) Find $u \in V$ such that

$$J(u) \leq J(v) \ \forall v \in V.$$

Theorem 3.3.1. *The problem (***P***) is equivalent to the problem (***M***).*

Proof. ((**P**) $\Rightarrow$(**M**)): Let u be a solution of (**P**).

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

Since u solves (**P**),

$$a(u, v - u) - l(v - u) = 0 \ \forall v \in V.$$

Therefore, noting that $a(\cdot, \cdot)$ is coercive,

$$\begin{aligned} J(v) &= J(u) + (a(u, v - u) - l(v - u)) + \frac{1}{2}a(v - u, v - u) \\ &= J(u) + \frac{1}{2}a(v - u, v - u) \\ &\geq J(u) + \frac{\alpha}{2}\|v - u\|_V^2 \\ &\geq J(u), \end{aligned}$$

for all $v \in V$. This means that u solves (**M**).

((**M**) $\Rightarrow$(**P**)): Let u denote a solution of (**M**). Since we have $J(u) \leq J(v)$ for all $v \in V$, for all $t \in \mathbb{R}$ we see

$$J(u) \leq J(u + tv). \tag{3.4}$$

Let us fix v and define

$$G(t) := J(u + tv).$$

Calculations yield

$$\begin{aligned} G(t) &= \frac{1}{2}a(u + tv, u + tv) - l(u + tv) \\ &= \frac{1}{2}a(u, u) - l(u) + t(a(u, v) - l(v)) + \frac{1}{2}t^2(a(v, v)), \end{aligned}$$

which means that G is quadratic in t.

Then, by (3.4)

$$G(0) = J(u) \leq J(u + tv) = G(t)$$

for all $t \in \mathbb{R}$. Note that $G(t)$ has a critical point (a minimum) at $t = 0$. Thus, $G'(0) = 0$. Since

$$G'(t) = a(u, v) - l(v) + ta(v, v),$$

we obtain

$$0 = a(u, v) - l(v),$$

which implies

$$a(u, v) = l(v) \ \forall v \in V.$$

Hence, u solves (**P**). □

Note that we say (**P**) is the variational *Euler-Lagrange* equation for (**M**).

Chapter 4

Abstract Finite Element Method

4.1 Abstract FEM

Let h be a parameter that we shall send to zero. We have a family V_h of finite dimensional subspaces of V, i.e,

- V_h is a linear space.
- $V_h \subset V$.
- $\dim V_h = N_h$ (integer depending on h), $N_h \to +\infty$ as $h \to 0$.

$(\mathbf{P}_h)$ Find $u_h \in V_h$ such that $a(u_h, v_h) = l(v_h)\ \forall v_h \in V_h$.

Theorem 4.1.1. *If $a(\cdot,\cdot)$ is a coercive bilinear form and $l(\cdot)$ is linear, there uniquely exists $u_h \in V_h$ solving $(\mathbf{P}_h)$.*

Proof. (Uniqueness): Suppose u_h^1, u_h^2 solve $(\mathbf{P}_h)$. Then we see

$$\begin{aligned} & a(u_h^i, v_h) = l(v_h)\ \forall v_h \in V_h,\ i = 1, 2. \\ \Longrightarrow\ & a(u_h^1 - u_h^2, v_h) = 0\ \forall v_h \in V_h. \\ \Longrightarrow\ & \alpha \|u_h^1 - u_h^2\|_V^2 \le a(u_h^1 - u_h^2, u_h^1 - u_h^2) = 0. \\ \Longrightarrow\ & u_h^1 - u_h^2 = 0. \end{aligned}$$

(Existence): Now, V_h is finite dimensional, so it has a basis $\{\phi_j^h\}_{j=1}^{N_h}$. For all u_h there uniquely exists a vector $(\alpha_1, \cdots, \alpha_{N_h}) \in \mathbb{R}^{N_h}$ such that

$$u_h = \sum_{j=1}^{N_h} \alpha_j \phi_j^h.$$

If u_h is a solution of $(\mathbf{P}_h)$,

$$\begin{aligned} & a(u_h, \phi_k^h) = l(\phi_k^h) \ k = 1, 2, \cdots, N_h. \\ \iff & \sum_{j=1}^{N_h} \alpha_j a(\phi_j^h, \phi_k^h) = l(\phi_k^h) \ k = 1, 2, \cdots, N_h. \\ \iff & A\alpha = \mathbf{b}, \end{aligned}$$

where $A = (a(\phi_j^h, \phi_k^h))_{j,k=1,\cdots,N_h}$, $\alpha = (\alpha_{\mathbf{1}}, \cdots, \alpha_{\mathbf{N_h}})^{\mathbf{T}}$,
$\mathbf{b} = (l(\phi_1^h), \cdots, l(\phi_{N_h}^h))^T$. Note that the matrix A is non-singular since $a(\cdot, \cdot)$ is coercive. Therefore, the vector $\mathbf{alpha} = \mathbf{A^{-1}b}$ gives a solution u_h of $(\mathbf{P}_h)$. □

Remark.

1) Similarly we can propose a problem
   $(\mathbf{P}^*)$ Find $u^* \in V$ such that $a(v, u^*) = l(v)$ for all $v \in V$.
   By setting $a^*(w, v) := a(v, w)$ and using Lax-Milgram's theorem we can show the unique existence of a solution of $(\mathbf{P}^*)$.

2) If $a(\cdot, \cdot)$ is symmetric, then $(\mathbf{P}^*)$ is equivalent to $(\mathbf{P})$.

3) If $a(\cdot, \cdot)$ is symmetric, then A is symmetric. If $a(\cdot, \cdot)$ is coercive, A is positive definite and satisfies $\beta \mathbf{A} \beta^{\mathbf{T}} > \mathbf{0} \ \forall \beta \in \mathbb{R}^{\mathbf{N_h}} \backslash \{\mathbf{0}\}$.

4) For symmetric $a(\cdot, \cdot)$ we propose a minimization problem:-
   $(\mathbf{M}_h)$ Find $u_h \in V_h$ such that $J(u_h) \leq J(v_h)$ for all $v_h \in V_h$.
   The equivalence between $(\mathbf{M}_h)$ and $(\mathbf{P}_h)$ can be proved in the same way as the equivalence between $(\mathbf{M})$ and $(\mathbf{P})$.

Exercise Show that $(\mathbf{M}_h)$ can be formulated as:-
Find $\alpha \mathbf{a}$ such that

$$\frac{1}{2} \alpha^{\mathbf{T}} \mathbf{A} \alpha - \mathbf{b}^{\mathbf{T}} \alpha \leq \frac{\mathbf{1}}{\mathbf{2}} \beta^{\mathbf{T}} \mathbf{A} \beta - \mathbf{b}^{\mathbf{T}} \beta$$

for all $\beta \in \mathbb{R}^{\mathbf{N_h}}$.

4.2 Galerkin Orthogonality

The idea of Galerkin orthogonality is the main ingredient for proving an error bound in the 'energy' norm, i.e. an error bound in the V norm.

Clearly $e_h = u - u_h$ is the error, which in general is not 0. In order to evaluate the quality of our approximation we try to estimate the error. In order to do this we find an equation that the error satisfies. Observe that since $V_h \subset V$ we have from (**??**) that by choosing $v = v_h \in V_h \subset V$ we have for $u \in V$

$$a(u, v_h) = l(v_h) \qquad \forall\, v_h \in V_h. \tag{4.1}$$

Now since u and u_h essentially solve the same variational equation subtracting (**??**) from (4.1) we have

$$a(u, v_h) - a(u_h, v_h) = l(v_h) - l(v_h) \tag{4.2}$$

$$\Rightarrow a(u - u_h, v_h) = 0 \qquad \forall\, v_h \in V_h. \tag{4.3}$$

This is called Galerkin orthogonality.

We wish to estimate a norm of e_h. So consider

$$\begin{aligned} a(e_h, e_h) &= a(u - u_h, u - u_h) = a(u - u_h, u - v_h + v_h - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h). \end{aligned}$$

We use Galerkin orthogonality to deduce that the last term here is zero beacuse $v_h - u_h \in V_h$! Thus

$$a(e_h, e_h) = a(e_h, u - v_h) \qquad \forall\, v_h \in V_h.$$

Lemma 4.2.1 (Cea's Lemma)**.** *The Galerkin approximation u_h of (**??**) satisifies the error bound*

$$\|u - u_h\|_V \leq \frac{c_1}{c_0} \min_{v_h \in V_h} \|u - v_h\|_V. \tag{4.4}$$

Proof: For $e_h = u - u_h$ we have seen that by Galerkin orthogonality $a(e_h, e_h) = a(e_h, u - v_h)\ \forall\, v_h \in V_h$. By the coercivity and boundedness of $a(\cdot, \cdot)$

$$\begin{aligned} &c_0 \|e_h\|_V^2 \leq a(e_h, e_h) = a(e_h, u - v_h) \leq c_1 \|e_h\|_V \|u - v_h\|_V \qquad \forall\, v_h \in V_h \\ &\Rightarrow \|e_h\|_V \leq \frac{c_1}{c_0} \min_{v_h \in V_h} \|u - v_h\|_V. \end{aligned}$$

■

Remark 4.2.2.

1. This shows us that the discretisation error e_h measured in the V norm is of the 'same size' as the best approximation to u in V_h. The best approximation to u in V_h is defined as u_h^B where $\|u - u_h^B\|_V = \min_{v_h \in V_h} \|u - v_h\|_V$.
2. We use Cea's lemma by using an element of V_h which is a natural approximation to u and for which we can estimate the error.
3. In general finite element spaces V_h have approximation properties like $\min_{v_h \in V} \|u - v_h\|_V \leq c(u)h^s$ where h is the element size and s in an integer depending on the polynomials used to define V_h. The higher s is the more derivatives in u that are required in order to define $c(u)$.

4.3 Abstract Aubin-Nitsche Lemma

Suppose that the assumptions of the Lax-Milgram theorem hold with respect to the bilinear form $a(\cdot, \cdot)$ and the Hilbert space V. Suppose that there exists a Hilbert space H with inner product $< \cdot, \cdot >_H$ into which V can be continuously embedded. It follows that for any $g \in H$ we may define a bounded linear functional $l_g(\cdot)$ by

$$l_g(v) :=< g, v >_H \quad \text{and} \quad |l_g(v)| = | < g, v >_H | \leq ||g||_H ||v||_H \leq c_H ||g||_H ||v||_V.$$

- **Dual regularity**

  There exists a positive constant c_s such that for any $g \in H$, the *adjoint problem* find $w(g) \in V$ such that

  $$a(v, w(g)) = l_g(v) \forall v \in V$$

  has a unique solution satisfying the regularity result

  $$||w(g)||_Z \leq c_s ||g||_H.$$

- **Approximation**

  Let V_h be a subspace of V which satisfies the following approximation property for a subspace $Z \subset V$: There exists a linear operator $\mathcal{I}_h : Z \to V_h$ for which there is a constant $\mathcal{K}$ independent of v and h such that for all $v \in Z$,

  $$||v - \mathcal{I}_h v||_V \leq \mathcal{K} h ||v||_Z.$$

Lemma 4.3.1. *(Aubin-Nitsche lemma) Under the above assumptions,*

$$||u - u_h||_H \leq \mathcal{K} c_s \gamma h ||u - u_h||_V.$$

Proof

Setting $e := u - u_h \in V$ we have that

$$||e||_H^2 =< e, e >_H = a(e, w(e))$$

and using Galerkin orthogonality,

$$\begin{aligned} < e, e >_H = a(e, w(e)) = a(u - u_h, w(e)) = a(u - u_h, w(e) - w(e)_h^*) \\ \leq \gamma ||u - u_h||_V ||w(e) - w(e)_h^*||_V \end{aligned}$$

so that, by approximation,

$$||e||_H^2 \leq \gamma ||u - u_h||_V \mathcal{K} h ||w(e)||_Z$$

and applying the dual regularity result,

$$||e||_H^2 \leq \gamma c_s \mathcal{K} h ||e||_H$$

from which we have the desired result. □

4.4 Abstract Error Bound

- **Regularity**

  There exists a positive constant c_r such that for any $f \in H$, the *problem* find $w(f) \in V$ such that

  $$a(w(f), v) = l_f(v) \forall v \in V$$

  has a unique solution satisfying the regularity result

  $$||w(f)||_Z \leq c_r ||f||_H.$$

From Cea's lemma we have

$$||u - u_h||_V \leq \frac{\gamma}{\alpha} ||u - v||_V \quad \forall v \in V_h$$

and using the approximation assumption together with the regularity assumption $u \in Z$ we have

$$||u - u_h||_V \leq \frac{\gamma}{\alpha}\mathcal{K}h||u||_Z$$

and

$$||u - u_h||_H \leq Ch^2||u||_Z.$$

Higher order approximation of u

Suppose that

- u is sufficently smooth in a space $Z(k)$, say, and the space V_h has a higher power of approximation so that

$$||v - \mathcal{I}_h v||_V \leq \mathcal{K}h^k||v||_{Z(k)}.$$

then we have the a priori error bound

$$||u - u_h||_H + h||u - u_h||_V \leq Ch^{k+1}||u||_{Z(k)}$$

Chapter 5

Variational Formulation of Boundary Value Problems

5.1 Elements of Function Spaces

5.1.1 Space of Continuous Functions

- $\mathbb{N}$ is a set of non-negative integers.
- 1) An n-tuple $\alpha = (\alpha_{\mathbf{1}}, \cdots, \alpha_{\mathbf{n}}) \in \mathbb{N}^{\mathbf{n}}$ is called a *multi-index*.

  2) The length of α is

  $$|\alpha| := \sum_{\mathbf{j=1}}^{\mathbf{n}} \alpha_{\mathbf{j}}.$$

  3) $\underline{0} = (0, \cdots, 0)$.
- Set $D^{\alpha} := (\partial/\partial x_1)^{\alpha_1} \cdots (\partial/\partial x_n)^{\alpha_n}$.

Example 5.1.1. *Assume* $n = 3$, $\alpha = (\alpha_{\mathbf{1}}, \alpha_{\mathbf{2}}, \alpha_{\mathbf{3}}) \in \mathbb{N}^{\mathbf{3}}$, $u(x_1, x_2, x_3) : \mathbb{R}^3 \to \mathbb{R}$.

What is $\sum_{|\alpha|=\mathbf{3}} D^\alpha u$ *?*

$$|\alpha| = \mathbf{3} \implies \sum_{\mathbf{j=1}}^{\mathbf{3}} \alpha_{\mathbf{j}} = \mathbf{3}.$$

$$\implies \alpha = (\mathbf{3},\mathbf{0},\mathbf{0}), (\mathbf{0},\mathbf{3},\mathbf{0}), (\mathbf{0},\mathbf{0},\mathbf{3}), (\mathbf{2},\mathbf{1},\mathbf{0}), (\mathbf{2},\mathbf{0},\mathbf{1}), (\mathbf{0},\mathbf{2},\mathbf{1}),$$
$$(1,2,0), (1,0,2), (0,1,2), (1,1,1).$$

$$\implies \sum_{|\alpha|=\mathbf{3}} D^\alpha u = \frac{\partial^3 u}{\partial x_1^3} + \frac{\partial^3 u}{\partial x_2^3} + \frac{\partial^3 u}{\partial x_3^3} + \frac{\partial^3 u}{\partial x_1^2 \partial x_2} + \frac{\partial^3 u}{\partial x_1^2 \partial x_3} + \frac{\partial^3 u}{\partial x_2^2 \partial x_3}$$
$$+ \frac{\partial^3 u}{\partial x_1 \partial x_2^2} + \frac{\partial^3 u}{\partial x_1 \partial x_3^2} + \frac{\partial^3 u}{\partial x_2 \partial x_3^2} + \frac{\partial^3 u}{\partial x_1 \partial x_2 \partial x_3}.$$

This sort of list can get very long. Hence D^α *is useful notation.*

Definition 5.1.2. Let Ω be an open set in $\mathbb{R}^n$. Let $k \in \mathbb{N}$. Define spaces $C^k(\Omega)$, $C^k(\overline{\Omega})$ and $C^\infty(\Omega)$ by

$$C^k(\Omega) := \{u : \Omega \to \mathbb{R} \mid D^\alpha u \text{ is continuous in } \Omega \text{ for all } |\alpha| \le \mathbf{k}\},$$
$$C^k(\overline{\Omega}) := \{u : \overline{\Omega} \to \mathbb{R} \mid D^\alpha u \text{ is continuous in } \overline{\Omega} \text{ for all } |\alpha| \le \mathbf{k}\},$$
$$C^\infty(\Omega) := \{u : \Omega \to \mathbb{R} \mid D^\alpha u \text{ is continuous in } \Omega \text{ for all } \alpha \in \mathbb{N}^{\mathbf{n}}\},$$

where $\overline{\Omega}$ is the closure of Ω. If Ω is bounded, $\overline{\Omega} = \Omega \cup \partial\Omega$, where $\partial\Omega$ is the boundary of Ω. We denote $C(\Omega) = C^0(\Omega)$ and $C(\overline{\Omega}) = C^0(\overline{\Omega})$.

Example 5.1.3. *Set* $I := (0,1)$ *and* $u(x) := 1/x^2 \ \forall x \in I$. *Then clearly for all* $k \ge 0$ $u \in C^k(I)$. *However, in* $\overline{I} = [0,1]$ u *is not continuous at* 0. *Thus,* $u \notin C(\overline{I})$.

Definition 5.1.4. For a bounded open set $\Omega \subset \mathbb{R}^n$, $k \in \mathbb{N}$ and $u \in C^k(\overline{\Omega})$, the norm $\|u\|_{C^k(\overline{\Omega})}$ is defined by

$$\|u\|_{C^k(\overline{\Omega})} := \sum_{|\alpha| \le \mathbf{k}} \sup_{x \in \overline{\Omega}} |D^\alpha u(x)|.$$

Example 5.1.5. *Let* $I = (0,1)$, $u(x) := x$, $u \in C(\overline{I})$. *Then,* $\sup_{x \in \overline{I}} |u(x)| = 1$.

Definition 5.1.6. For an open set $\Omega \subset \mathbb{R}^n$ and $\eta \in C(\Omega)$ the support of η denoted by *support* η ($\subset \mathbb{R}^n$) is defined by

$$\textit{support } \eta := \text{ the closure of } \{x \in \Omega \mid \eta(x) \neq 0\}.$$

Remark.
- The support of η is the smallest closed subset of $\overline{\Omega}$ such that $\eta = 0$ in $\Omega \backslash$ *support* u.

- If *support* η is bounded then we say that η has compact support.

Definition 5.1.7. Define $C_0^k(\Omega)$ $(\subset C^k(\Omega))$ by

$$C_0^k(\Omega) := \{u \in C^k(\Omega) \mid \textit{support } u \text{ is a bounded subset of } \Omega\}.$$

Lemma 5.1.8. *$C_0^\infty(\Omega)$ is dense in $L^p(\Omega)$ for all $1 \le p < \infty$.*

This means that every function in $L^p(\Omega)$ can be arbitrarily closely approximated by a function from $C_0^\infty(\Omega)$ (with the error measured in the $L^p(\Omega)$ norm).

Example 5.1.9. *1) Let $0 = x_0 < x_1 < \cdots < x_n = 1$ be a patition of $[0,1]$. Define $\phi_j(x) : [0,1] \to \mathbb{R}$ by*

$$\phi_j(x) := \begin{cases} \dfrac{x - x_{j-1}}{h} & x \in (x_{j-1}, x_j), \\ \dfrac{x_{j+1} - x}{h} & x \in (x_j, x_{j+1}), \\ 0 & \textit{elsewhere.} \end{cases}$$

Then we see $\phi_j \in C(\bar{I})$ and support $\phi_j = [x_{j-1}, x_{j+1}]$.

2) Define $w(x) : \mathbb{R}^n \to \mathbb{R}$ by

$$w(x) := \begin{cases} e^{-\dfrac{1}{1-|x|^2}} & |x| < 1, \\ 0 & \textit{otherwise.} \end{cases}$$

Then we see support $w = \{x \in \mathbb{R}^n \mid |x| \le 1\}$ and $w \in C_0^\infty(\Omega)$ for any Ω containing $\bar{B}(0,1) := \{x \in \mathbb{R}^n : |x| \le 1\}$.

Proof For $|x| < 1$, write $w(x) = e^{-t}$ with $t = (1 - |x|^2)^{-1}$ and show that $w_{x_j} = -2e^{-t}t^2 x_j$. Prove by induction that for all multi-indices α, that there exists a polynomial P_α such that $D^\alpha w(x) = P_\alpha(x)e^{-t}t^{2|\alpha|}, |x| < 1$.

5.1.2 Spaces of Integrable Functions

Definition 5.1.10. Let Ω denote an open subset of $\mathbb{R}^n$ and assume $1 \le p < \infty$. We define a space of integrable functions $L^p(\Omega)$ by

$$L^p(\Omega) := \left\{ v : \Omega \to \mathbb{R} \mid \int_\Omega |v(x)|^p dx < +\infty \right\}.$$

The space $L^p(\Omega)$ is a Banach space with norm $\|\cdot\|_{L^p(\Omega)}$ defined by

$$\|v\|_{L^p(\Omega)} = \left(\int_\Omega |v(x)|^p dx \right)^{1/p}.$$

Especially the space $L^2(\Omega)$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{L^2(\Omega)}$ defined by

$$\langle u, v \rangle_{L^2(\Omega)} := \int_\Omega u(x)v(x)dx$$

and norm $\|\cdot\|_{L^2(\Omega)}$ defined by $\|u\|_{L^2(\Omega)} := \sqrt{\langle u, v \rangle_{L^2(\Omega)}}$.

We have *Minkowski's inequality* as follows. For $u, v \in L^p(\Omega)$, $1 \leq p < \infty$

$$\|u + v\|_{L^p(\Omega)} \leq \|u\|_{L^p(\Omega)} + \|v\|_{L^p(\Omega)}.$$

We also have *Hölder's inequality.* For $u \in L^p(\Omega)$ and $v \in L^q(\Omega)$, $1 \leq p, q < \infty$ with $1/p + 1/q = 1$

$$\Big| \int_\Omega u(x)v(x)dx \Big| \leq \|u\|_{L^p(\Omega)} \|v\|_{L^q(\Omega)}.$$

Now, any two integrable functions are equivalent if they are equal *almost everywhere*, that is, they are equal except on a set of zero measure. Strictly speaking, $L^p(\Omega)$ consists of equivalent classes of functions.

Example 5.1.11. *Let* $u,\ v : (-1, 1) \to \mathbb{R}$ *be*

$$u(x) = \begin{cases} 1 & x \in (0, 1), \\ 0 & x \in (-1, 0], \end{cases} \quad v(x) = \begin{cases} 1 & x \in [0, 1), \\ 0 & x \in (-1, 0), \end{cases}$$

The functions u *and* v *are equal almost everywhere, since the set* $\{0\}$ *where* $u(0) \neq v(0)$ *has zero measure in the interval* $(-1, 1)$*. So* u *and* v *are equal as integrable functions in* $(-1, 1)$*.*

Suppose that $u \in C^k(\Omega)$, where Ω is an open set of $\mathbb{R}^n$. Let $v \in C_0^\infty(\Omega)$. Then we see by integration by parts

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

where $|\alpha| \leq \mathbf{k}$.

Definition 5.1.12. A function $\eta : \Omega \to \mathbb{R}$ is locally integrable if $\eta \in L^1(K)$ for every bounded open set K such that $\overline{K} \subset \Omega$. The space $L^1_{loc}(\Omega)$ consists of locally integrable functions.

Definition 5.1.13. Weak derivative Suppose $\eta : \Omega \to \mathbb{R} \in L^1_{loc}(\Omega)$ and there is a locally integrable function $w_\alpha : \Omega \to \mathbb{R}$ such that

$$\int_\Omega w_\alpha(x)\phi(x)dx = (-1)^{|\alpha|} \int_\Omega \eta(x) D^\alpha \phi(x)dx$$

for all $\phi \in C_0^\infty(\Omega)$.

Then the weak derivative of η of order α denoted by $D^\alpha u$ is defined by $D^\alpha u = w_\alpha$.

Note that at most only one w_α satisfies (5.1.13) so the weak derivative of u is well-defined. Indeed, the following *DuBois-Raymond* lemma shows such w_α is unique.

Lemma 5.1.14. *(DuBois-Raymond) Suppose Ω is an open set in $\mathbb{R}^n$ and $w : \Omega \to \mathbb{R}$ is locally integrable. If*

$$\int_\Omega w(x)\phi(x)dx = 0$$

for all $\phi \in C_0^\infty(\Omega)$, then $w(x) = 0$ for a.e $x \in \Omega$.

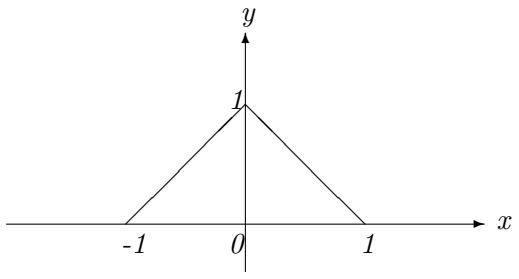
We will use D for both classical and weak derivatives.

Example 5.1.15. *Let $\Omega = \mathbb{R}$. Set $u(x) = (1 - |x|)_+$, $x \in \Omega$, where*

$$(x)_+ := \begin{cases} x & x > 0, \\ 0 & x \leq 0. \end{cases}$$

Thus,

$$u(x) := \begin{cases} 0 & x \leq -1, \\ 1 + x & -1 \leq x \leq 0, \\ 1 - x & 0 \leq x \leq 1, \\ 0 & 1 \leq x. \end{cases}$$



Clearly we see that u is locally integrable, $u \in C(\Omega)$ and $u \notin C^1(\Omega)$. However, it may have a weak derivative. Take any $\phi \in C_0^\infty(\Omega)$ and $\alpha = 1$. Then,

$$\begin{aligned}
(-1)^{|\alpha|} \int_\Omega u(x) D^\alpha \phi(x) dx &= -\int_{-\infty}^{\infty} u(x)\phi'(x)dx \\
&= -\int_{-1}^{1} (1-|x|)\phi'(x)dx \\
&= -\int_{-1}^{0} (1+x)\phi'(x)dx - \int_{0}^{1} (1-x)\phi'(x)dx \\
&= \int_{-1}^{0} 1 \cdot \phi(x)dx + \int_{0}^{1} (-1)\phi(x)dx \\
&= \int_\Omega w(x)\phi(x)dx,
\end{aligned}$$

where

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

Here we do not worry about the points $x = -1, 0, 1$, since they have zero measure. Thus, u has its weak derivative $Du = w$.

Definition 5.1.16. Let k be a non-negative integer and $p \in [0, \infty)$. The space $W^{k,p}(\Omega)$ defined by

$$W^{k,p}(\Omega) := \{u \in L^p(\Omega) \mid D^\alpha u \in L^p(\Omega)\ \forall |\alpha| \leq \mathbf{k}\}$$

is called a *Sobolev space*. It is a Banach space with the norm

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

Especially, when $p = 2$, we denote $H^k(\Omega)$ as $W^{k,2}(\Omega)$. It is a Hilbert space with the inner product

$$\langle u, v \rangle_{H^k(\Omega)} := \sum_{|\alpha| \leq \mathbf{k}} \langle D^\alpha u, D^\alpha v \rangle_{L^2(\Omega)}.$$

Of special interest are $H^1(\Omega)$ and $H^2(\Omega)$. If $\Omega = (a,b), \subset \mathbb{R}$, we see that

$$\begin{aligned}
\langle u, v\rangle_{H^1(\Omega)} &= \langle u, v\rangle_{L^2(\Omega)} + \langle Du, Dv\rangle_{L^2(\Omega)} \\
&= \int_a^b u(x)v(x)dx + \int_a^b Du(x)Dv(x)dx. \\
\langle u, v\rangle_{H^2(\Omega)} &= \langle u, v\rangle_{L^2(\Omega)} + \langle Du, Dv\rangle_{L^2(\Omega)} + \langle D^2u, D^2v\rangle_{L^2(\Omega)} \\
&= \int_a^b u(x)v(x)dx + \int_a^b Du(x)Dv(x)dx + \int_a^b D^2u(x)D^2v(x)dx.
\end{aligned}$$

Remark. 1) By using Hölder's inequality, we can prove *Cauchy-Schwarz* inequality for the inner product of $H^k(\Omega)$ as follows.

$$\begin{aligned}
|\langle u, v\rangle_{H^k(\Omega)}| &\leq \sum_{|\alpha|\leq \mathbf{k}} |\langle D^\alpha u, D^\alpha v\rangle_{L^2(\Omega)}| \\
&\leq \sum_{|\alpha|\leq \mathbf{k}} \|D^\alpha u\|_{L^2(\Omega)} \|D^\alpha v\|_{L^2(\Omega)} \\
&\leq \sqrt{\sum_{|\alpha|\leq \mathbf{k}} \|D^\alpha u\|^2_{L^2(\Omega)}} \sqrt{\sum_{|\alpha|\leq \mathbf{k}} \|D^\alpha v\|^2_{L^2(\Omega)}} \\
&= \|u\|_{H^k(\Omega)} \|v\|_{H^k(\Omega)}.
\end{aligned}$$

2) Let $\Omega = (a,b) \subset \mathbb{R}$ and $u \in H^1(\Omega)$. Then $u \in C(\overline{\Omega})$. In higher space dimensions this statement is no longer true.

Definition 5.1.17. $H_0^k(\Omega)$ is defined to be the closure of $C_0^\infty(\Omega)$ with respect to $||\cdot||_{H^k(\Omega)}$.

Loosely speaking $H_0^k(\Omega)$ consists of those functions in $H^k(\Omega)$ whose derivatives up to order $|\alpha| \leq k-1$ vanish on the boundary

Theorem 5.1.18. Poincare Inequality

If Ω is a bounded domain then then there exists a constant $C = C(\Omega)$ (depending on Ω) such that

$$||v||_{L^2(\Omega)} \leq C|v|_{H^1(\Omega)} \qquad \forall v \in H_0^1(\Omega)$$

Theorem 5.1.19. Trace theorem

Let Ω be bounded with a Lipschitz boundary $\partial\Omega$. Then there exists a bounded linear operator $tr : H^1(\Omega) \to L^2(\partial\Omega)$ with the property that $tr(v)$ and $v|_{\partial\Omega}$ coincide on $\partial\Omega$ for all $v \in C(\bar{\Omega}) \cap H^1(\Omega)$. It follows that there exists a constant $c = C(\Omega)$ such that

$$||tr(v)||_{L^2(\partial\Omega)} \leq C||v||_{H^1(\Omega)} \quad \forall v \in H^1(\Omega)$$

The following inequality is also true:

$$||v||_{L^2(\partial\Omega)} \leq C||v||^{1/2}_{L^2(\Omega)}||v||^{1/2}_{H^1(\Omega)}.$$

Theorem 5.1.1. Divergence theorem *Let Ω be a bounded Lipschitz domain in $\mathbb{R}^n$ and $Q : \overline{\Omega} \to \mathbb{R}^n$ be a vector field whose components are in $H^1(\Omega)$. The following equality holds.*

$$\int_\Omega \nabla Q dx = \int_{\partial\Omega} Q \cdot \nu ds$$

where ν is the unit outward pointing normal to $\partial\Omega$.

Remark.

- Suppose $Q = f\mathbf{e_i}$ with the coordinate vector $\mathbf{e}_i = (0, \cdots, 1, \cdots, 0)^T$, i.e, the jth component is $\{\mathbf{e}_i\}_j = \delta_{i,j}$. Then we see

$$\nabla Q = \frac{\partial}{\partial x_i} f.$$

So by the Divergence theorem

$$\int_\Omega \frac{\partial f}{\partial x_i} dx = \int_{\partial\Omega} f\nu_i ds.$$

In one dimensional case where $\Omega = (a, b)$, $\partial\Omega = \{a, b\}$, the Divergence theorem becomes

$$\int_a^b \frac{\partial f}{\partial x} dx = f(b) - f(a).$$

- Similarly $Q = wv\mathbf{e_i}$ yields

$$\int_\Omega \frac{\partial w}{\partial x_i} v dx = -\int_\Omega w \frac{\partial v}{\partial x_i} dx + \int_{\partial\Omega} wv\nu_i dS.$$

Let us derive the integration by parts formula.

Proposition 5.1.20. Integration by parts
For $Q \in H^1(\overline{\Omega}; \mathbb{R}^n)$, $g \in H^1(\Omega)$,

$$\int_\Omega Q \cdot \nabla g dx = \int_{\partial\Omega} gQ \cdot \nu ds - \int_\Omega g\nabla \cdot Q dx.$$

Proof. By the divergence theorem we see that

$$\int_\Omega \nabla(Qg)dx = \int_{\partial\Omega} Q \cdot \nu g ds.$$

Alternatively,

$$\nabla(Qg) = g\nabla Q + Q \cdot \nabla g.$$

By combining these equality we get the desired formula. □

For example, if $Q = \nabla u$ and $g = v$, we have by integration by parts formula that

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_{\partial\Omega} v\nabla u \cdot \nu ds - \int_\Omega v\Delta u dx.$$

where Δ is the Laplacian and noting that

$$\nabla u \cdot \nu = \frac{\partial \mathbf{u}}{\partial \nu}$$

we obtain

$$\int_\Omega v\Delta u dx = \int_{\partial\Omega} v\frac{\partial u}{\partial \nu} ds - \int_\Omega \nabla u \cdot \nabla v dx.$$

Similarly we have that

$$-\int_\Omega \nabla(p\nabla u)v dx = \int_\Omega p\nabla u \cdot \nabla v dx - \int_{\partial\Omega} p\frac{\partial u}{\partial \nu} v ds.$$

5.1.3 The space $H_0^1(\Omega)$ in variational problems

Because of the Poincare inequality there exists a constant c such that

$$c||v||_{H^1(\Omega)} \leq |v|_{H_0^1(\Omega)} \leq ||v||_{H^1(\Omega)}.$$

It follows that we may take in a variational problem $V = H_0^1(\Omega)$ with norm either $||v||_V = ||v||_{H^1(\Omega)}$ or $||v||_V = |v|_{H^1(\Omega)}$.

Note that for

$$a(w, v) = \int_\Omega (p\nabla w \cdot \nabla v + qwv)dx$$

with $p(x) \geq p_m > 0$ and $q(x) \geq q_m \geq 0$ we can use the inequalities

$$a(v, v) \geq p_m |v|^2_{H^1(\Omega)} \quad \text{if} \quad q_m = 0$$

or

$$a(v, v) \geq \min(p_m, q_m)||v||^2_{H^1(\Omega)} \quad \text{if} \quad q_m > 0$$

to deduce that $a(\cdot, \cdot)$ is coercive.

5.2 One Dimensional Problem

5.2.1 One Dimensional H^1 inequalities

Here we derive some inequalities in the one dimensional case $\Omega = (a, b)$. We define a function space $H^1_{\epsilon_0}(\Omega)$ by

$$H^1_{\epsilon_0}(\Omega) := \{\phi \in H^1(\Omega) \mid \phi(a) = 0\}.$$

The following inequality is one example of *Poincaré-Friedrichs inequality.*

Proposition 5.2.1. *For all $\phi \in H^1_{\epsilon_0}(\Omega)$,*

$$\|\phi\|_{L^2(\Omega)} \leq \frac{1}{\sqrt{2}}(b-a)\|D\phi\|_{L^2(\Omega)}.$$

Proof. We can write that for $a \leq \forall x \leq b$

$$\phi(x) = \int_a^x D\phi(\eta)d\eta.$$

Then we see that

$$\begin{aligned}
\|\phi\|^2_{L^2(\Omega)} &= \int_a^b \phi(x)^2 dx \\
&= \int_a^b \left(\int_a^x D\phi(\eta)d\eta\right)^2 dx \\
&\leq \int_a^b \left(\int_a^x 1^2 dx\right)\left(\int_a^x D\phi(\eta)^2 d\eta\right) dx \\
&= \int_a^b (x-a)\int_a^x D\phi(\eta)^2 d\eta dx \\
&\leq \int_a^b (x-a)\int_a^b D\phi(\eta)^2 d\eta dx \\
&= \frac{1}{2}(b-a)^2\|D\phi\|^2_{L^2(a,b)}.
\end{aligned}$$

□

Example 5.2.2. *We can apply this inequality to prove the unique solvability of the Dirichlet problem with a functional space V, a bilinear form $a(\cdot,\cdot)$ and a linear*

functional $l(\cdot)$ defined by

$$\begin{aligned} V &:= \{\phi \in H^1(0,1) \mid \phi(0) = \phi(1) = 0\}, \\ a(u,v) &:= \int_0^1 pDuDvdx \text{ for } \forall u, v \in V, \\ l(u) &:= \int_0^1 fudx \text{ for } \forall u \in V. \end{aligned}$$

We only check that the bilinear form a is bounded and coercive. We see that

$$|a(u,v)| \leq \sup_{x\in(0,1)} |p(x)| \|Du\|_{L^2(\Omega)} \|Dv\|_{L^2(\Omega)} \leq \sup_{x\in(0,1)} |p(x)| \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}$$

and

$$\begin{aligned} a(v,v) &\geq p_0 \|Dv\|^2_{L^2(\Omega)} \\ &= \frac{p_0}{2}\|Dv\|^2_{L^2(\Omega)} + \frac{p_0}{2}\|Dv\|^2_{L^2(\Omega)} \\ &\geq \frac{p_0}{2}\left(\|Dv\|^2_{L^2(\Omega)} + \frac{2\|v\|^2_{L^2(\Omega)}}{(b-a)^2}\right) \\ &\geq \alpha \|v\|^2_V, \end{aligned}$$

where $\alpha := p_0 \min(1, 2/(b-a)^2)/2$.

Proposition 5.2.3. *The following Agmon's inequality holds. For all $\phi \in H^1_{\epsilon_0}(\Omega)$*

$$\max_{x\in\overline{\Omega}} |\phi(x)|^2 \leq 2\|\phi\|_{L^2(\Omega)} \|D\phi\|_{L^2(\Omega)}.$$

Proof.

$$\begin{aligned} \phi(x)^2 &= \int_a^x \frac{d\phi(\eta)^2}{d\eta} d\eta \\ &= 2\int_a^x \phi(\eta) D\phi(\eta) d\eta \\ &\leq 2\left(\int_a^x \phi(\eta)^2 d\eta\right)^{1/2} \left(\int_a^x D\phi(\eta)^2 d\eta\right)^{1/2} \\ &\leq 2\left(\int_a^b \phi(\eta)^2 d\eta\right)^{1/2} \left(\int_a^b D\phi(\eta)^2 d\eta\right)^{1/2} \\ &\leq 2\|\phi\|_{L^2(\Omega)} \|D\phi\|_{L^2(\Omega)}, \end{aligned}$$

which gives the inequality. □

Noting that

$$\|\phi\|_{L^2(\Omega)} \leq \sqrt{b-a} \max_{x\in[a,b]} |\phi(x)|,$$

this Agmon's inequality yields

$$\max_{x\in\overline{\Omega}} |\phi(x)|^2 \leq 2\sqrt{b-a} \max_{x\in[a,b]} |\phi(x)| \|D\phi\|_{L^2(\Omega)},$$

or

$$\max_{x\in\overline{\Omega}} |\phi(x)| \leq 2\sqrt{b-a} \|D\phi\|_{L^2(\Omega)}$$

for any $\phi \in H^1_{\epsilon_0}(\Omega)$.

5.2.2 Dirichlet condition

Let $\Omega = (0,1)$, $p(\cdot), q(\cdot) \in C(\overline{\Omega})$ and $f(\cdot) \in L^2(\Omega)$. Note that $\partial\Omega = \{x = 0\} \cup \{x = 1\}$. We consider the following problem.

Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} -\dfrac{d}{dx}(p\dfrac{du}{dx}) + qu = f, & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega. \end{cases} \tag{BVP}$$

Specifying the value of u at boundary points is said to be a *Dirichlet boundary condition*. Now the methodology is

1) multiply the equation by a test function, integrate by parts and use boundary conditions appropriately,

2) identify V, $a(\cdot,\cdot)$ and $l(\cdot)$,

3) verify, if possible, the assumptions of Lax-Milgram.

$\Longrightarrow$ Unique existence to the variational formulation of the BVP.

Let $\phi : \overline{\Omega} \to \mathbb{R}$ be sufficiently smooth. We will call ϕ our test function. Let us follow the methodology.

1)

$$\int_\Omega f(x)\phi(x)dx = \int_\Omega \left(-\frac{d}{dx}(p(x)\frac{du(x)}{dx})\phi(x) + q(x)u(x)\phi(x)\right)dx$$
$$= \left[-p(x)\frac{du(x)}{dx}\phi(x)\right]_{x=0}^{x=1} + \int_\Omega \left(p(x)\frac{du(x)}{dx}\frac{d\phi(x)}{dx} + q(x)u(x)\phi(x)\right)dx.$$

We want to eliminate the term $[p(x)du(x)/dx\phi(x)]_{x=0}^{x=1}$, so we suppose that the test function ϕ satisfies the same Dirichlet conditions as u, i.e, $\phi(0) = \phi(1) = 0$. Then we have that

$$\int_\Omega \left(p(x)\frac{du(x)}{dx}\frac{d\phi(x)}{dx} + q(x)u(x)\phi(x)\right)dx = \int_\Omega f(x)\phi(x)dx$$

for any test function ϕ. We want u, ϕ to be from the same space. For the term $\int_\Omega u\phi dx$ to make sense, we need $u, \phi \in L^2(\Omega)$. For the derivatives $du/dx, d\phi/dx$ to make sense, we take this further, so $u, \phi \in H^1(\Omega)$.

2) Let us choose

$$V := \left\{\phi \in H^1(\Omega) \mid \phi(0) = \phi(1) = 0\right\},$$

where

$$H^1(\Omega) = \left\{\phi \in L^2(\Omega) \mid D\phi \in L^2(\Omega)\right\}.$$

We equip V with the inner product $\langle\cdot,\cdot\rangle_V := \langle\cdot,\cdot\rangle_{H^1(\Omega)}$. Let us define

$$a(u,v) := \int_\Omega (pDuDv + quv)dx,$$
$$l(v) := \int_\Omega fv\,dx.$$

Moreover, assume that $p(x) \geq p_0 > 0$, $q(x) \geq q_0 > 0$ for all $x \in \overline{\Omega}$.

3) We will verify the assumptions of Lax-Milgram's theorem.

i) For $\phi \in V$ and $f \in L^2(\Omega)$, we see by Cauchy-Schwarz inequality that

$$\begin{aligned}|l(\phi)| &= |\int_\Omega f\phi dx| \\ &\leq \|f\|_{L^2(\Omega)}\|\phi\|_{L^2(\Omega)} \\ &\leq \|f\|_{L^2(\Omega)} \quad (\|\phi\|^2_{L^2(\Omega)} + \|D\phi\|^2_{L^2(\Omega)})^{1/2} \\ &= c_l\|\phi\|_V,\end{aligned}$$

where we have set $c_l := \|f\|_{L^2(\Omega)}$. Thus, $l : V \to \mathbb{R}$ is bounded. Clearly l is linear, i.e, $l(\alpha\phi + \beta\psi) = \alpha l(\phi) + \beta l(\psi)$ for any $\phi, \psi \in V$ and $\alpha, \beta \in \mathbb{R}$.

ii) Obviously $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is bilinear. Moreover $a(\cdot,\cdot)$ is bounded. Indeed,

$$\begin{aligned}
|a(\phi,\psi)| &\le |\int_\Omega pD\phi D\psi dx| + |\int_\Omega q\phi\psi dx| \\
&\le \max_{x\in\overline{\Omega}} |p(x)| \int_\Omega |D\phi D\psi| dx + \max_{x\in\overline{\Omega}} |q(x)| \int_\Omega |\phi\psi| dx \\
&\le \max_{x\in\overline{\Omega}} |p(x)| \|D\phi\|_{L^2(\Omega)} \|D\psi\|_{L^2(\Omega)} + \max_{x\in\overline{\Omega}} |q(x)| \|\phi\|_{L^2(\Omega)} \|\psi\|_{L^2(\Omega)} \\
&\le C(\|D\phi\|_{L^2(\Omega)} \|D\psi\|_{L^2(\Omega)} + \|\phi\|_{L^2(\Omega)} \|\psi\|_{L^2(\Omega)}) \\
&\le C\sqrt{\|D\phi\|^2_{L^2(\Omega)} + \|\phi\|^2_{L^2(\Omega)}} \sqrt{\|D\psi\|^2_{L^2(\Omega)} + \|\psi\|^2_{L^2(\Omega)}} \\
&= C\|\phi\|_{H^1(\Omega)} \|\psi\|_{H^1(\Omega)},
\end{aligned}$$

where we have set

$$C := \max\{\max_{x\in\overline{\Omega}} |p(x)|, \max_{x\in\overline{\Omega}} |q(x)|\}.$$

The bilinear form $a(\cdot,\cdot)$ is coercive, since for all $\phi \in V$

$$\begin{aligned}
a(\phi,\phi) &= \int_\Omega p|D\phi|^2 dx + \int_\Omega q|\phi|^2 dx \\
&\ge p_0 \int_\Omega |D\phi|^2 dx + q_0 \int_\Omega |\phi|^2 dx \\
&= \hat{C}\|\phi\|^2_V,
\end{aligned}$$

where we have set $\hat{C} := \min\{p_0, q_0\}$.

We can now apply Lax-Milgram's theorem to see that there uniquely exists a solution to the following problem

(**P**) Find $u \in V$ such that

$$\int_\Omega (pDuD\phi + qu\phi) dx = \int_\Omega f\phi dx,$$

for any $\phi \in V$.

Remark. For $V = H^1_0(\Omega)$ we can use the norm $\|\cdot\|_V$ defined by

$$\|\phi\|^2_V = \int_\Omega |D\phi|^2 dx,$$

since the following Poincare inequality holds:- there exists $C > 0$ such that

$$\int_\Omega |\phi|^2 dx \leq C \int_\Omega |D\phi|^2 dx \text{ for } \forall \phi \in V.$$

By using this inequality we can prove the unique existence of the solution solving (**P**) with $q \equiv 0$ in the same way as above.

5.2.3 One Dimensional Problem: Neumann condition

Let $\Omega = (0,1)$, $p(x), q(x) \in C(\overline{\Omega})$ and $f(x) \in L^2(\Omega)$. We consider the following problem.

Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} -\dfrac{d}{dx}\left(p\dfrac{du}{dx}\right) + qu = f, & x \in \Omega, \\ \dfrac{d}{dx}u(x) = 0, & x \in \partial\Omega. \end{cases} \tag{NBVP}$$

Specifying the value of du/dx at boundary points is said to be a *Neumann boundary condition*. We assume the same conditions for p, q as before, i.e, $p(x) \geq p_0 > 0$, $q(x) \geq q_0 > 0$ in Ω. Let us derive the variational form. Take a sufficiently smooth test function ϕ, multiply (NBVP) by ϕ and integrate.

$$\begin{aligned} &\int_\Omega f(x)\phi(x)dx = \int_\Omega \left(-\frac{d}{dx}\left(p(x)\frac{du(x)}{dx}\right)\phi(x) + q(x)u(x)\phi(x)\right)dx \\ &= \left[-p(x)\frac{du(x)}{dx}\phi(x)\right]_{x=0}^{x=1} + \int_\Omega \left(p(x)\frac{du(x)}{dx}\frac{d\phi(x)}{dx} + q(x)u(x)\phi(x)\right)dx \\ &= \int_\Omega \left(p(x)\frac{du(x)}{dx}\frac{d\phi(x)}{dx} + q(x)u(x)\phi(x)\right)dx. \end{aligned}$$

We have eliminated the term $[p(x)du(x)/dx\phi(x)]_{x=0}^{x=1}$ by taking into account the Neumann boundary conditions $du(x)/dx = 0$ for $x = 0, 1$. Let us choose the functional space $V := H^1(\Omega)$ in this case and define

$$\begin{aligned} a(u,v) &:= \int_\Omega (pDuDv + quv)dx, \\ l(v) &:= \int_\Omega fv\,dx. \end{aligned}$$

The corresponding variational problem is that:-

(**P**) Find $u \in V$ such that

$$\int_\Omega (pDuD\phi + qu\phi)dx = \int_\Omega f\phi dx,$$

for any $\phi \in V$.

Again the linear form $l(\cdot) : V \to \mathbb{R}$ and the bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ satisfy the assumptions of Lax-Milgram's theorem, hence the problem (**P**) has the unique solution.

Remark. Consider (NBVP) with $q \equiv 0$. Then we see

$$\begin{aligned}\int_\Omega f(x)dx &= -\int_\Omega \frac{d}{dx}(p(x)\frac{du(x)}{dx})dx \\ &= \left[p(x)\frac{du(x)}{dx}\right]_{x=0}^{x=1} = 0.\end{aligned}$$

Thus, we need to assume $\int_\Omega f(x)dx = 0$ as a compatibility condition in this case. In order to prove the unique existence of the solution, we need to modify the functional space. Let us define $H^1_m(\Omega)$ by

$$H^1_m(\Omega) := \{\phi \in H^1(\Omega) \mid \int_\Omega \phi dx = 0\},$$

and equip the same inner product as $H^1(\Omega)$. Again *Poincare's inequality* is available for this space $H^1_m(\Omega)$, i.e, for all $\phi \in H^1_m(\Omega)$

$$\int_\Omega |\phi|^2 dx \le C\int_\Omega |D\phi|^2 dx,$$

where $C > 0$ is a constant. By using this inequality we can prove the unique existence of the solution solving (**P**) with $q \equiv 0$ and $V = H^1_m(\Omega)$ in the same way as above on the assumption $\int_\Omega f(x)dx = 0$.

5.2.4 One Dimensional Problem: Robin/Newton Condition

Let $\Omega = (0,1)$, $p(x), q(x) \in C(\overline{\Omega})$, $f(x) \in L^2(\Omega)$, $\delta, g_0, g_1 \in \mathbb{R}$ be constants. We consider the following problem. Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} -\dfrac{d}{dx}\left(p\dfrac{du}{dx}\right) + qu = f, \quad x \in \Omega, \\ -p\dfrac{d}{dx}u(0) + \delta u(0) = g_0, \\ p\dfrac{d}{dx}u(1) + \delta u(1) = g_1. \end{cases} \tag{RNBVP}$$

Let us derive the variational form. Take sufficiently smooth ϕ. This kind of boundary condition is said to be a *Robin/Newton boundary condition*. We assume the same conditions for p, q as before, i.e, $p(x) \geq p_0 > 0$, $q(x) \geq q_0 > 0$ in Ω and that $\delta \geq 0$.

$$\begin{aligned}
&\int_\Omega f(x)\phi(x)dx = \int_\Omega \left(-\frac{d}{dx}\left(p(x)\frac{du(x)}{dx} \right) \phi(x) + q(x)u(x)\phi(x) \right) dx \\
&= \left[-p(x)\frac{du(x)}{dx}\phi(x) \right]_{x=0}^{x=1} + \int_\Omega \left(p(x)\frac{du(x)}{dx}\frac{d\phi(x)}{dx} + q(x)u(x)\phi(x) \right) dx \\
&= (-g_1 + \delta u(1))\phi(1) - (g_0 - \delta u(0))\phi(0) \\
&\quad + \int_\Omega \left(p(x)\frac{du(x)}{dx}\frac{d\phi(x)}{dx} + q(x)u(x)\phi(x) \right) dx,
\end{aligned}$$

which is equal to

$$\begin{aligned}
&\int_\Omega \left(p(x)\frac{du(x)}{dx}\frac{d\phi(x)}{dx} + q(x)u(x)\phi(x) \right) dx + \delta u(1)\phi(1) + \delta u(0)\phi(0) \\
&\quad = \int_\Omega f(x)\phi(x)dx - g_1\phi(1) + g_0\phi(0).
\end{aligned}$$

This suggests that we should define $a(\cdot,\cdot)$, $l(\cdot)$ and the functional space V as following.

$$\begin{aligned}
a(u,v) &:= \int_\Omega (pDuDv + quv)dx + \delta u(1)v(1) + \delta u(0)v(0), \\
l(v) &:= \int_\Omega fvdx + g_1 v(1) + g_0 v(0), \\
V &:= H^1(\Omega).
\end{aligned}$$

As usual we need to show that the (bi)linear forms a, l are bounded and a is coercive to establish the unique solvability of $(RNBVP)$. Let us assume the following inequality holds true for a while.

$$|\phi(x)| \leq C\|\phi\|_{H^1(\Omega)} \tag{5.1}$$

for all $\phi \in H^1(\Omega)$. Then it is easy to see that $a(\cdot,\cdot), l(\cdot)$ are bounded. Now

$$\begin{aligned}
a(\phi,\phi) &\geq \min(p_0,q_0)\|\phi\|_V^2 + \delta(\phi(1)^2 + \phi(0)^2) \\
&\geq \min(p_0,q_0)\|\phi\|_V^2 \\
&= \alpha\|\phi\|_V^2,
\end{aligned}$$

where $\alpha = \min(p_0, q_0)$. Hence the bilinear form a becomes coercive and Lax-Milgram's theorem assures the unique existence of the solution.

5.3 Variational formulation of elliptic equations

5.3.1 Weak Solutions to Elliptic Problems

The simplest elliptic equation is Laplaces' equation:

$$\Delta u = 0, \tag{5.2}$$

where $\Delta := \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2}$ is the Laplace operator. A general second order elliptic equation is: given a bounded open set $\Omega \subset \mathbb{R}^n$ find u such that:

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x) \qquad x \in \Omega, \tag{5.3}$$

where classically $a_i j \in C^1(\Omega), i, j = 1, \ldots, n; b_i \in C(\Omega), i = 1, \ldots, n; c \in C(\Omega); f \in C(\Omega)$. For the equation to be elliptic we require

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geq \tilde{C}\sum_{i=1}^{n}\xi_i^2 \qquad \forall\, \xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n, \tag{5.4}$$

where $\tilde{C} > 0$ is independent of x, ξ. Condition (5.4) is called uniform ellipticity.

The equation is usually supplemented with boundary conditions - Dirichlet, Neumann, Robin, or a mixed Dirichlet/Neumann boundary.

In the case of the homogeneous Dirichlet problem ($u = 0$ on $\partial\Omega$) u is said to be a classical solution provided $u \in C^2(\Omega) \cap C(\overline{\Omega})$. Elliptic theory tells us that there exists a unique classical solution provided a_{ij}, b_i, c, f and $\partial\Omega$ are sufficiently smooth. However we are only intersted in problems where the data is not smooth, for example $f = sign(1/2 - |x|), \Omega = (-1, 1)$. This problem can't have $u \in C^2(\Omega)$ because Δu has a jump discontinuity are $|x| = 1/2$. With the help of functional analysis the existence/uniqueness theory for 'weak', 'variational' solutions turn out to be easy and is good for FEM.

5.3.2 Variational Formulation of Elliptic Equation: Neumann Condition

Let Ω be a bounded domain in $\mathbb{R}^n$ with smooth boundary $\partial\Omega$. Let $p, q \in C(\overline{\Omega})$ such that

$$p(x) \geq p_0 > 0,\ q(x) \geq q_0 > 0\ \forall x \in \overline{\Omega},$$

and $f \in L^2(\Omega)$.

Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} -\nabla \cdot (p\nabla u) + qu = f, & x \in \Omega, \\ \dfrac{\partial u}{\partial \mathbf{n}} = 0, & x \in \partial\Omega, \end{cases} \tag{NBVP}$$

where $\mathbf{n}$ is the unit outward normal to the boundary $\partial\Omega$. Note that

$$\begin{aligned} \nabla \cdot (p\nabla u) &= \sum_{i=1}^{n} \frac{\partial}{\partial_{x_i}} \left(p \frac{\partial u}{\partial x_i} \right) \\ &= \sum_{i=1}^{n} \left(p \frac{\partial^2 u}{\partial x_i^2} + \frac{\partial p}{\partial x_i} \frac{\partial u}{\partial x_i} \right) \\ &= p\Delta u + \nabla p \cdot \nabla u, \\ \frac{\partial u}{\partial \mathbf{n}} &= \nabla u \cdot \mathbf{n}. \end{aligned}$$

So we have a second order PDE. In one dimensional problem, in order to derive the variational formulation we used integration by parts. Let us revise some formulae related to the integration by parts.

Notation:

$$\begin{aligned} &\nabla v = \left(\frac{\partial v}{\partial x_1}, \ldots, \frac{\partial v}{\partial x_1} \right)^T \\ &\nabla \cdot \nabla v = \nabla^2 v = \Delta v \\ &\nabla \cdot \mathbf{A} = \sum_{i=1}^{n} \frac{\partial A_i}{\partial x_i} \\ &(D^2 v)_{ij} = \frac{\partial^2 v}{\partial x_i \partial x_j} \\ &\mathrm{Tr}(D^2 v) = \Delta v. \end{aligned}$$

Let v be a sufficiently smooth test function. Multiply (NBVP) by v and integrate using Divergence theorem.

$$\begin{aligned} \int_\Omega f v dx &= \int_\Omega (-\nabla \cdot (p\nabla u) + qu) v dx \\ &= \int_\Omega p\nabla u \nabla v dx - \int_{\partial\Omega} p \frac{\partial u}{\partial \mathbf{n}} v ds + \int_\Omega quv dx. \end{aligned}$$

Since $\partial u/\partial \mathbf{n} = \mathbf{0}$ on $\partial\Omega$, we do not need to place a restriction on the test function v. So if u solve (BVP), then

$$\int_\Omega (p\nabla u \nabla v + quv)dx = \int_\Omega fvdx,$$

for any sufficiently smooth function v.

Now to use Lax-Milgram, we have to set up V, $a(\cdot,\cdot)$ and $l(\cdot)$. In order for the two inner products on the left hand side to make sense, we take

$$\begin{aligned} V &= H^1(\Omega), \\ a(u,v) &= \int_\Omega (p\nabla u \nabla v + quv)dx, \\ l(v) &= \int_\Omega fvdx, \end{aligned}$$

for all $u, v \in V$. Note that V is a real Hilbert space with the norm

$$\|v\|_V = \|v\|_{H^1(\Omega)} = \left(\int_\Omega |\nabla v|^2 dx + \int_\Omega v^2 dx\right)^{1/2}$$

and obviously $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is bilinear and $l(\cdot) : V \to \mathbb{R}$ is linear. Moreover we observe

$$\begin{aligned} a(v,v) &= \int_\Omega (p|\nabla v|^2 + qv^2)dx \\ &\geq p_0 \int_\Omega |\nabla v|^2 dx + q_0 \int_\Omega v^2 dx \\ &\geq \min\{p_0, q_0\}\|v\|^2_{H^1(\Omega)} \\ &= \alpha \|v\|^2_{H^1(\Omega)}, \end{aligned}$$

where we have put $\alpha := \min\{p_0, q_0\}$. Thus $a(\cdot,\cdot)$ is coercive.

$$\begin{aligned} |a(v,w)| &= \left| \int_\Omega (p\nabla v \cdot \nabla w + qvw)\, dx \right| \\ &\leq \int_\Omega (|p\nabla u \cdot \nabla w| + |qvw|)dx \\ &\leq C \int_\Omega (|\nabla v|^2 + v^2)^{1/2}(|\nabla w|^2 + w^2)^{1/2} dx \\ &\leq C \left(\int_\Omega (|\nabla v|^2 + v^2)dx\right)^{1/2} \left(\int_\Omega (|\nabla w|^2 + w^2)dx\right)^{1/2} \\ &= C\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)}. \end{aligned}$$

Therefore, $a(\cdot,\cdot)$ is bounded. Finally let us check the boundedness of $l(\cdot)$.

$$\begin{aligned}
|l(v)| &= \Big| \int_\Omega f v dx \Big| \\
&\leq \int_\Omega |f||v| dx \\
&\leq \left(\int_\Omega f^2 dx \right)^{1/2} \left(\int_\Omega v^2 dx \right)^{1/2} \\
&= \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\
&\leq \|f\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} \\
&= C_l \|v\|_{H^1(\Omega)}.
\end{aligned}$$

Thus, $l(\cdot)$ is bounded. Now we can apply Lax-Milgram to prove that there exists a unique solution $u \in V$ to the following problem (**P**).

(**P**) Find $u \in V$ such that

$$a(u,v) = l(v)$$

for all $v \in V$.

5.3.3 Variational Formulation of Elliptic Equation: Dirichlet Problem

On the same assumptions on Ω, p, q, f, we consider the following problem.

Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} -\nabla \cdot (p \nabla u) + qu = f, & x \in \Omega, \\ u = 0, & x \in \partial\Omega, \end{cases} \tag{DBVP}$$

Let us derive the variational form of (DBVP) as in the section 2.5. Multiply (DBVP) by a sufficiently smooth test function v and integrate. Then we see

$$\int_\Omega f v dx = \int_\Omega (p \nabla u \nabla v + quv) dx - \int_{\partial\Omega} p \frac{\partial u}{\partial n} v dx.$$

Since we have $u \equiv 0$ on $\partial\Omega$, we have to force our test function v to satisfy the same condition; $v \equiv 0$ on $\partial\Omega$. Then we obtain

$$\int_\Omega (p \nabla u \nabla v + quv) dx = \int_\Omega f v dx$$

for any sufficient smooth function v with $v \equiv 0$ on $\partial\Omega$. Set

$$\begin{aligned} V &:= \{v \in H^1(\Omega) \mid v = 0 \text{ on } \partial\Omega\} \\ &= H_0^1(\Omega). \end{aligned}$$

Note that V is a real Hilbert space with the inner product

$$\langle v, w\rangle_V := \langle \nabla v, \nabla w\rangle_{L^2(\Omega)} + \langle v, w\rangle_{L^2(\Omega)}$$

and $\|v\|_V = \|v\|_{H^1(\Omega)}$. As before we define

$$\begin{aligned} a(v, w) &:= \int_\Omega (p\nabla v \nabla w + qvw)\ dx, \\ l(v) &= \int_\Omega fv\ dx. \end{aligned}$$

The same argument as the previous section shows that $a(\cdot, \cdot)$ is a coercive and bounded bi-linear form and $l(\cdot)$ is a bounded linear functional on V. Therefore Lax-Milgram's theorem tells us that there uniquely exists a solution to the variational problem of (DBVP).

5.3.4 A general second order elliptic problem

Consider the problem

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_j}\left(a_{ij}(x)\frac{\partial u}{\partial x_i}\right) + \sum_{i=1}^{n} b_i(x)\frac{\partial u}{\partial x_i} + c(x)u = f(x) \qquad \forall\, x \in \Omega \tag{5.5}$$

with $u = 0$ on $\partial\Omega$. Multiply by a test funciton and integrate by parts in the second order term using the divergence theorem. The result is the weak (variational) form of the BVP: find $u \in V$ such that $a(u, v) = l(v)\ \forall\, v \in V$ where $V = H_0^1(\Omega)$ and

$$\begin{aligned} a(w, v) &:= \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x)\frac{\partial w}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{i=1}^{n} \int_\Omega b_i(x)\frac{\partial w}{\partial x_i}v(x) + \int_\Omega c(x)w(x), \\ l(v) &:= \int_\Omega f(x)v(x) = (f, v). \end{aligned}$$

We seek to apply the Lax-Milgram theorem. Recall $(v, w)_{H_0^1(\Omega)} = \int_\Omega vw + \nabla v \nabla w = (v, w) + (\nabla v, \nabla w)$. We have three conditions to check to satisfy the theorem.

(1) Is $l(\cdot)$ a bounded linear functional? Clearly

$$l(\alpha v + \beta w) = (f, \alpha v + \beta w) = \alpha(f, v) + \beta(f, w) = \alpha l(v) + \beta l(w)$$

so $l(\cdot)$ is a linear functional on V and

$$|l(v)| = \left| \int_\Omega f(x)v(x)\ dx \right| \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{H_0^1(\Omega)}$$

where we have used the Cauchy-Schwartz inequality and thus $l(\cdot)$ is bounded.

(2) Is $a(\cdot,\cdot)$ bounded? Assume that $\|a_{ij}\|_{L^\infty(\Omega)}, \|b_i\|_{L^\infty(\Omega)}, \|c\|_{L^\infty(\Omega)}$ are all bounded for all i, j and that $f \in L^2(\Omega)$. Then

$$\begin{aligned}
|a(w,v)| &\leq \left| \sum_{i,j=1}^n \int_\Omega a_{ij} w_{x_i} v_{x_j}\ dx \right| + \left| \sum_{i=1}^n \int_\Omega b_i w_{x_i} v\ dx \right| + \left| \int_\Omega cwv\ dx \right| \\
&\leq \sum_{i,j=1}^n \max_{x\in\Omega} |a_{ij}(x)| \int_\Omega |w_{x_i}||v_{x_j}|\ dx \\
&+ \sum_{i=1}^n \max_{x\in\Omega} |b_i(x)| \int_\Omega |w_{x_i}||v|\ dx + \max_{x\in\Omega} |c(x)| \int_\Omega |w||v|\ dx \\
&\leq \tilde{c} \left(\sum_{i,j=1}^n \int_\Omega |w_{x_i}||v_{x_j}|\ dx + \sum_{i=1}^n \int_\Omega |w_{x_i}||v|\ dx + \int_\Omega |w||v|\ dx \right) \\
&\leq \tilde{c} \left(\sum_{i,j=1}^n \|w_{x_i}\|\|v_{x_j}\| + \sum_{i=1}^n \|w_{x_i}\|\|v\| + \|w\|\|v\| \right) \\
&\leq \tilde{c} \left(\sum_{i,j=1}^n \|w\|_V\|v\|_V + \sum_{i=1}^n \|w\|_V\|v\|_V + \|w\|_V\|v\|_V \right) \\
&= c_1\|w\|_V\|v\|_V
\end{aligned}$$

where $\tilde{c} = \max\{\max_{ij} \max_\Omega |a_{ij}(x)|, \max_i \max_\Omega |b_i(x)|, \max_\Omega |c(x)|\}$ and $c_1 = \tilde{c}(n^2 + n + 1)$.

(3) Is $a(\cdot,\cdot)$ coercive? The crucial assumption is that the a_{ij} satisfies the ellipticity assumption

$$\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \geq \tilde{c} \sum_{i=1}^n \xi_i^2 \qquad \forall\, (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n,\ \forall\, x \in \overline{\Omega}, \tag{5.6}$$

i.e. for all $x \in \overline{\Omega}$ we must have

$$\xi^T A(x)\xi \geq \hat{c}\|\xi\|^2 = \hat{c}\xi^T\xi. \tag{5.7}$$

We also assume that

$$c(x) - \frac{1}{2}\sum_{i=1}^{n} \frac{\partial b_i(x)}{\partial x_i} \geq 0 \forall\, x \in \Omega. \tag{5.8}$$

Then

$$\begin{aligned} a(v,v) &= \sum_{i,j=1}^{n} \int_\Omega a_{ij}(x) v_{x_i} v_{x_j} + \sum_{i=1}^{n} \int_\Omega b_i(x) v_{x_i} v + \int_\Omega c(x) v(x)^2 \\ &\geq \tilde{c} \int_\Omega \sum_{i=1}^{n} v_{x_i}^2 + \sum_{i=1}^{n} \int_\Omega b_i(x) \frac{\partial v^2/2}{\partial x_i} + \int cv^2. \end{aligned}$$

The middle integral here is $\frac{1}{2}\int_\Omega b \cdot \nabla(v^2)$, which after integration by parts equals $-\frac{1}{2}\int_\Omega v^2 \nabla \cdot b$ so that

$$\begin{aligned} a(v,v) &\geq \tilde{c} \sum_{i=1}^{n} \int_\Omega v_{x_i}^2 + \int_\Omega v^2 (c(x) - \frac{1}{2}\nabla \cdot b(x)) \\ &\geq \tilde{c} \sum_{i=1}^{n} \int_\Omega v_{x_i}^2 \\ &= \tilde{c}\|\nabla v\|^2 \end{aligned}$$

Note that we need $\nabla \cdot b \in L^\infty(\Omega)$ for this to work. We wish to show that

$$a(v,v) \geq c_0 \|v\|_V^2 = c_0(\|v\| + \|\nabla v\|). \tag{5.9}$$

Recall the Poincare-Friedrichs inequalities

$$\|v\|^2 \leq c_* \|\nabla v\|^2 \qquad \forall\, v \in H_0^1(\Omega).$$

Hence

$$\begin{aligned} a(v,v) &\geq \tilde{c}\|\nabla v\|^2 \geq \frac{\tilde{c}}{c_*}\|v\|^2 \\ \frac{1}{2}a(v,v) + \frac{1}{2}a(v,v) &\geq \frac{\tilde{c}}{2}\|\nabla v\|^2 + \frac{\tilde{c}}{2c_*}\|v\|^2 \\ &\geq c_0\left(\|\nabla v\|^2 + \|v\|^2\right) \end{aligned}$$

5.4 Inhomogeneous Boundary Conditions

Consider the elliptic problem

$$-\nabla \cdot (p\nabla u) + qu = f \qquad x \in \Omega \tag{5.10}$$
$$u = g \qquad x \in \partial\Omega \tag{5.11}$$

where Ω is a bounded open subset of $\mathbb{R}^2$. We assume that the data p, q, f, g are sufficiently smooth and that

$$p_M \geq p(x) \geq p_0 > 0 \qquad \forall\, x \in \Omega$$
$$q_M \geq q(x) \geq q_0 > 0 \qquad \forall\, x \in \Omega$$

Let v be a test function. Multiply by v and integrate:

$$\begin{aligned} 0 &= -\int_\Omega v\nabla \cdot (p\nabla u) + \int_\Omega quv - \int_\Omega fv \\ &= I_1 + I_2 + I_3 \end{aligned}$$

Nov choosing $\varphi = v, \mathbf{f} = p\nabla u$ in:

$$\begin{aligned} \nabla \cdot (\varphi \mathbf{f}) &= \varphi\nabla \cdot \mathbf{f} + \nabla\varphi \cdot \mathbf{f} \\ \nabla \cdot (vp\nabla u) &= v\nabla \cdot (p\nabla u) + \nabla v \cdot p\nabla u \\ I_1 &= -\int_\Omega \nabla \cdot (vp\nabla u) + \nabla v \cdot p\nabla u \\ &= \int_\Omega p\nabla v \cdot \nabla u - \int_{\partial\Omega} vp\nabla u \cdot \nu \end{aligned}$$

Choosing $v = 0$ on $\partial\Omega$ we have $I_1 = \int_\Omega p\nabla v \cdot \nabla u$. Thus

$$0 = \int_\Omega p\nabla v \cdot \nabla u + quv - fv \qquad \forall\, v \in H_0^1(\Omega). \tag{5.12}$$

Set $V_0 = H_0^1(\Omega), a(u,v) = \int_\Omega p\nabla v \cdot \nabla u + quv, l(v) = \int_\Omega fv$. Note that $u \notin V_0$. However $g \in H^1(\Omega)$ so $u - g \in H^1(\Omega)$ and $u - g \in H_0^1(\Omega) = V_0$, i.e $u \in V_g := \{w \in V = H^1(\Omega) : w = g + v, v \in V_0\}$.

Thus our variational problem (P) is to find $u \in V_g$ such that $a(u,v) = l(v)\ \forall\, v \in V_0$. Observe that V_0 is a linear space but $V_g = g + V_0$ is an affine space. We can't apply Lax-Milgram directly. Consider $u^* = u - g \in V_0$:

$$a(u^* + g, v) = a(u,v) = l(v) \qquad \forall\, v \in V_0$$

so $u^* \in V_0$ solves

$$a(u^*, v) = l(v) - a(g, v) =: l^*(v) \qquad \forall\, v \in V_0$$

Now we just need to check Lax-Milgram for this problem. Clearly $a(\cdot, \cdot)$ is bilinear (and symmetric). Coercivity:

$$a(v, v) = \int_\Omega p|\nabla v|^2 + qv^2 \geq p_0 \int_\Omega |\nabla v|^2 + q_0 \int_\Omega v^2 \geq \min(p_0, q_0) \int_\Omega |\nabla v|^2 + v^2 \geq c_0 \|v\|_{H^1(\Omega)}.$$

Boundedness: using the Cauchy-Schwartz inequality we have

$$\begin{aligned} |a(w, v)| &= \left| \int_\Omega p \nabla w \nabla v + qwv \right| \leq p_M \int_\Omega |\nabla w||\nabla v| + q_M \int_\Omega |w||v| \\ &\leq \max(p_M, q_M)(\|\nabla w\|\|\nabla v\| + \|w\|\|v\|) \\ &\leq \tilde{c}\|w\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}. \end{aligned}$$

Clearly l^* is linear and

$$|l^*(v)| = |l(v) - a(g, v)| \leq |l(v)| + |a(g, v)| \leq \|f\|\|v\| + \tilde{c}\|g\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)} \leq L^*\|v\|_{H^1(\Omega)}.$$

Thus there exists a unique u^* and we conclude therefore that there exists a unique $u = u^* + g$.

The bilinear form is symmetric so there is an energy and associated minimisation problem:

$$\begin{aligned} &J(v) = \frac{1}{2}a(v, v) - l(v) \\ &\text{Find } u \in V_g \text{ } s.t. \qquad J(u) \leq J(v) \qquad \forall\, v \in V_g. \end{aligned}$$

Exercise: Prove that these two problems are equivalent.

Chapter 6

Finite Element Method

6.1 One Dimensional Problems

6.1.1 Dirichlet Problem

Let $\Omega = (0,1)$, $p(\cdot), q(\cdot) \in C(\overline{\Omega})$ with $p(x) \geq p_0 > 0$, $q(x) \geq q_0 > 0$ for all $x \in \overline{\Omega}$, $f(\cdot) \in L^2(\Omega)$ and set $V := H^1(\Omega)$. We consider the following Dirichlet problem.

Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{cases} -\dfrac{d}{dx}(p\dfrac{du}{dx}) + qu = f & \text{in } \Omega, \\ u = 0 & \text{in } \partial\Omega. \end{cases} \tag{BVP}$$

For simpler presentation we set

$$q := 1$$

in the following. Further, we set

$$a(u,v) := \int_0^1 \left(p\frac{du}{dx}\frac{dv}{dx} + uv \right) dx,$$
$$l(v) := \int_0^1 fv dx.$$

We formulated (BVP) as the following variational problem.

(**P**) Find $u \in V$ such that

$$a(u, \phi) = l(\phi) \text{ for all } \phi \in V. \tag{6.1}$$

The Lax-Milgram theorem shows the unique existence of the solution u to (**P**).

6.1.2 Abstract framework

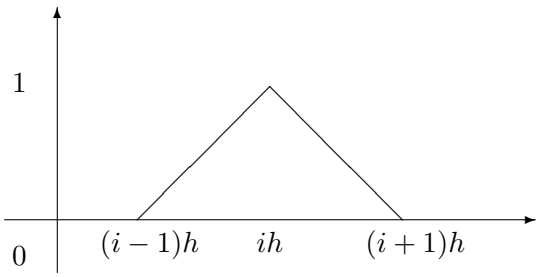
The goal is now to approximate solutions to (**P**) by piecewise linear functions. We use a uniform grid on Ω with $N+1$ nodes $x_j = jh$, $j = 0, 1, \ldots, N+1$ where $h = 1/N$. Now, let us define the finite dimensional space V_h by

$$V_h := \{v \in C[0,1] \mid v|_{[x_j, x_{j+1}]} \text{ is linear for any } j = 0, 1, \cdots, N, \\ \text{and } v(0) = v(1) = 0\}.$$

We have that $v_h \in V_h$ if and only if

$$v_h = \sum_{i=1}^{N} v_i \phi_i \quad \text{with } v_i = v_h(x_i)$$

where $\{\phi_i\}_{i=1}^N$ are the piecewise linear continuous functions defined by $\phi_i(x_j) = \delta_{i,j}$ $i, j = 1, \ldots, N$.



Moreover, the classical derivative of v_h is defined and constant on each interval (x_j, x_{j+1}) and given by

$$\frac{d}{dx} v_h = \frac{v_{j+1} - v_j}{h} \text{ in } (x_j, x_{j+1}).$$

Extending this to a piecewise constant function we obtain a function in $L^2(\Omega)$ which we denote by Dv_h and which is the weak derivative of v_h: for any $\phi \in C_0^\infty([0,1])$

$$\begin{aligned}\int_0^1 v_h \frac{d}{dx}\phi \, dx &= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} v_h \frac{d}{dx}\phi \, dx \\ &= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} -\frac{d}{dx} v_h \, \phi \, dx + v_h(x_{i+1})\phi(x_{i+1}) - v_h(x_i)\phi(x_i) \\ &= \int_0^1 -Dv_h \, \phi \, dx + \underbrace{v_h(x_{N+1})\phi(x_{N+1})}_{=v_h(1)\phi(1)=0} - \underbrace{v_h(x_0)\phi(x_0)}_{=v_h(0)\phi(0)=0}\end{aligned}$$

where we used the continuity of v_h in the second identity and the boundary values in the last line. A a consequence, $V_h \subset V$, and we may formulate the following discrete problem:

($\mathbf{P}_h$) Find $u_h \in V_h$ such that

$$a(u_h, v_h) = l(v_h) \text{ for all } v_h \in V_h. \tag{6.2}$$

Unique solvability follows again with the Lax-Milgram theorem.

6.1.3 Reformulation as system of linear equations

By substituting $u_h = \sum_{j=1}^N u_j \phi_j$ into (6.2) we obtain

$$\sum_{j=1}^{N} u_j a(\phi_j, \phi_i) = l(\phi_i), \;\; i = 1, 2, \cdots, N.$$

This can be written as $AU = F$ where

$$\begin{aligned}A &= (A_{i,j})_{i,j=1,\ldots,N} = (a(\phi_j, \phi_i))_{i,j=1,\ldots,N}, \\ F &= (l(\phi_1), \ldots, l(\phi_N))^T.\end{aligned}$$

Let us characterize the matrix $A = K + M$ where

$$K_{i,j} = \int_0^1 pD\phi_i D\phi_j dx \quad \text{and} \quad M_{i,j} = \int_0^1 \phi_i \phi_j dx.$$

If $j \notin \{i-1, i, i+1\}$, then $K_{i,j} = 0$ and $M_{i,j} = 0$ since the supports of ϕ_i and ϕ_j are disjoint. Clearly N and K are symmetric matrices. We have that

$$\begin{aligned} K_{i,i+1} &= \int_{x_{i-1}}^{x_{i+1}} pD\phi_i D\phi_{i+1} dx \\ &= \int_{x_i}^{x_{i+1}} p\left(-\frac{1}{h}\right)\left(\frac{1}{h}\right) dx \\ &= -\frac{1}{h^2}\int_{x_i}^{x_{i+1}} p dx \\ K_{i,i} &= \int_{x_{i-1}}^{x_{i+1}} pD\phi_i D\phi_i dx \\ &= \frac{1}{h^2}\int_{x_{i-1}}^{x_{i+1}} p dx. \end{aligned}$$

Set

$$\hat{p}_i := \frac{1}{h}\int_{x_i}^{x_{i+1}} p dx.$$

Then we have seen that

$$K_{i,i} = \frac{1}{h}(\hat{p}_{i-1} + \hat{p}_i), \quad K_{i,i+1} = -\frac{1}{h}\hat{p}_i, \quad K_{i,i-1} = -\frac{1}{h}\hat{p}_{i-1}.$$

Now

$$\int_0^1 \phi_j^2(x)dx = \int_{x_{j-1}}^{x_j} \frac{(x-x_{j-1})^2}{h^2}dx + \int_{x_j}^{x_{j+1}} \frac{(x-x_{j+1})^2}{h^2}dx = 2\int_0^1 y^2 h dy = \frac{2}{3}h$$

and

$$\int_0^1 \phi_j(x)\phi_{j-1}(x)dx = \int_{x_{j-1}}^{x_j} \frac{(x-x_{j-1})(x_j-x)}{h^2}dx = \int_0^1 y(1-y)hdy = \frac{h}{6}.$$

Then we have seen that

$$M_{i,i} = \frac{2}{3}h, \quad M_{i,i+1} = \frac{1}{6}h, \quad M_{i,i-1} = \frac{1}{6}h.$$

If we set

$$\hat{f}_i = \frac{1}{h}\int_{x_{i-1}}^{x_{i+1}} f\phi_i dx,$$

then we see $F_i = l(\phi_i) = h\hat{f}_i$.

After dividing the system $AU = F$ by h we end up with

$$\begin{bmatrix} \frac{\hat{p}_0+\hat{p}_1}{h^2}+\frac{2}{3} & \frac{-\hat{p}_1}{h^2}+\frac{1}{6} & 0 & \cdots & 0 \\ \frac{-\hat{p}_1}{h^2}+\frac{1}{6} & \frac{\hat{p}_1+\hat{p}_2}{h^2}+\frac{2}{3} & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \frac{-p_{N-1}}{h^2}+\frac{1}{6} \\ 0 & \cdots & 0 & \frac{-\hat{p}_{N-1}}{h^2}+\frac{1}{6} & \frac{\hat{p}_{N-1}+\hat{p}_N}{h^2}+\frac{2}{3} \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ u_N \end{bmatrix} = \begin{bmatrix} \hat{f}_1 \\ \vdots \\ \vdots \\ \vdots \\ \hat{f}_N \end{bmatrix}.$$

Remark 6.1.1.

1. Note that A is a tridiagonal symmetric and non-singular matrix since U is unique.

2. The next steps will be

   1) find u_h,

   2) bound the error $\|u - u_h\|_V$.

   The step 1) requires numerical linear algebra, the step 2) requires abstract error analysis and approximation theory.

3. We can also apply this approach to other boundary conditions.

4. Looking forward, we shall see that the discrete equations are very similar to a finite difference approximation of this boundary value problem: the matrix $\frac{1}{h}K$ is identical to the standard approximation of the negative of the second derivative, see also section 2.3. However, the overall approximation differs in two ways: firstly, the term involving μ contains the matrix M where the identity would naturally appear in a finite difference approximation (replacing hN by the identity is known as **mass-lumping**); secondly the vector F is found by testing the function f against basis functions, not against delta functions (numerical integration may be used to achieve point evaluations of f).

6.2 Finite Element Method in two dimensions

Take Ω to be a polygon. Let $\mathcal{T}_h$ be a triangulation of Ω, $\mathcal{T}_h = \{\kappa\}$ and set $h_\kappa = \operatorname{diam}\kappa$ (the length of the longest side), $h = \max \operatorname{diam}\kappa$. We assume that $|\mathcal{T}_h| < \infty$. Any triangles in $\mathcal{T}_h$ must intersect along a complete edge, at a vertex of not at all.

Note that any linear function on $\mathbb{R}^2$ is of the form $v(x,y) = a+bx+cy$ and is defined by three parameters. Thus any function

$$v_h \in V_h := \{\eta \in C(\overline{\Omega}) : v_h|_\kappa \text{ is linear}\} \tag{6.3}$$

is uniquely determined by it values at the vertices of the triangulation:

$$\phi_i(x_j) = \delta_{ij} \qquad i,j = 1,\ldots,N_h,\ x_j \text{ is a triangle vertex.} \tag{6.4}$$

The support of the basis functions is local, so A will again be sparse.

Example 6.2.1. *Find $u \in H^1(\Omega)$ such that*

$$\int_\Omega p\nabla u\nabla v + quv\ dx = \int_\Omega fv\ dx \qquad \forall\, v \in H^1(\Omega). \tag{6.5}$$

A finite element method applied to this yields the problem: find $u_h \in H^1(\Omega)$ such that

$$\int_\Omega p\nabla u_h\nabla v_h + qu_hv_h\ dx = \int_\Omega fv_h\ dx \qquad \forall\, v_h \in V_h. \tag{6.6}$$

i.e. $A\mathbf{u} = \mathbf{b}$, where

$$\begin{aligned} u_h &= \sum_{i=1}^{N_h} u_j\phi_j, \\ A_{jk} &= \int_\Omega p\nabla\phi_j\nabla\phi_k + q\phi_j\phi_k\ dx, \\ b_j &= \int_\Omega f\phi_j\ dx. \end{aligned}$$

Note that:

$$\begin{aligned} A_{jk} &= \int_\Omega p\nabla\phi_j\nabla\phi_k + q\phi_j\phi_k\ dx \\ &= \int_{Support_{\phi_j}\cap Support_{\phi_k}} p\nabla\phi_j\nabla\phi_k + q\phi_j\phi_k\ dx \\ &= \sum_{\kappa\ with\ common\ edge\ jk} \int_\kappa p\nabla\phi_j\nabla\phi_k + q\phi_j\phi_k\ dx. \end{aligned}$$

Let Ω be a polygon. A triangulation of Ω is a union of triangles, $\mathcal{T}_h$, such that no vertex of any triangle lies in the interior of any edge of any other triangle or lies in

the interior of any other triangle. A generic triangle is denoted by K. Set

$$\begin{aligned} h_K &:= \text{ the diameter of } K \\ &= \text{ the length of the longest edge (side) of } K. \\ \mathcal{T}_h &:= \text{ set of the triangles forming } \Omega, \\ h &:= \max_{K \in \mathcal{T}_h} h_K. \end{aligned}$$

We assume that $\mathcal{T}_h$ contains a finite number of triangles, each of which has a non-empty interior. Any triangles in $\mathcal{T}_h$ must intersect along a complete edge, or at a vertex, or not at all.

Note that a linear function v on $\mathbb{R}^2$ is of the form

$$v(x, y) = a + bx + cy.$$

It is defined by 3 parameters. Consider a triangle K with vertices p_1, p_2, p_3 which are not collinear. Consider ϕ_i^K $(i = 1, 2, 3)$ where each ϕ_i^K is a linear function satisfying

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

Thus, on K we have

$$v(x) = \sum_{i=1}^{3} v_i \phi_i^K(x),$$

where $v_j := v(p_j)$, since for any linear functions v_1, v_2 in K $v_1(p_j) = v_2(p_j)$ $(j = 1, 2, 3)$ implies $v_1 \equiv v_2$. Otherwise there would be a non-zero linear function vanishing at three points which are not collinear and this is impossible.

Now, let v and w be two linear functions on K.

$$v = \sum_{i=1}^{3} v_i \phi_i, \ w = \sum_{i=1}^{3} w_i \phi_i.$$

Since $v(x, y) = w(x, y)$ $\forall (x, y) \in K$ if and only if $v_i = w_i$ for $i = 1, 2, 3$. Hence the ϕ_i^K are linearly independent on K.

Now, let $\mathcal{T}_h$ be a triangulation of Ω. Suppose that the vertices are $\{x_j\}_{j=1,\cdots,N(h)}$. Consider the vertex x_k and all the triangles which have x_k as a vertex. Consider a continuous function ϕ_k which is linear on such a triangle and satisfies

$$\phi_k(x_j) = \delta_{j,k}$$

for any vertex x_j. Set

$$V_h := \{v \in C(\overline{\Omega}) \mid v|_K \text{ is linear}\},$$
$$= \left\{ v \in C(\overline{\Omega}) \;\middle|\; v = \sum_{j=1}^{N(h)} v_j \phi_j(x) \right\}.$$

Then, V_h is a subspace of $H^1(\Omega)$. Note that this is a consequence of calculating the weak derivatives $w_j := \partial_{x_j} v, j = 1, 2$ of $v \in V_h$ to be the $L^2(\Omega)$ function which is piecewise defined by $w_{j|K} = \partial_{x_j}(v_{|K})$.

If we have a problem with a Dirichlet condition, then we might use

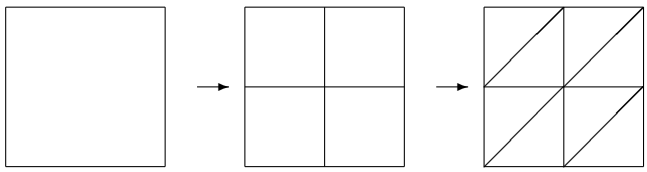
$$V_h = \{v \in C(\overline{\Omega}) \mid v|_K \text{ is linear}, v = 0 \text{ on } \partial\Omega\}$$
$$= \left\{ v \in C(\overline{\Omega}) \;\middle|\; v = \sum_{j=1}^{N^*(h)} v_j \phi_j(x) \right\},$$

where $N^*(h)$ is the number of interior nodes.

Example 6.2.2. *Let $\Omega = (0,1) \times (0,1)$, $h = 1/(N+1)$. Set $(x_i, y_j) = (ih, jh)$. Consider the triangulation $\mathcal{T}_h$ of Ω defined by the method below.*

i) Divide Ω into squares with vertices (x_i, y_j).

ii) Divide each square into two triangles using the diagonal in the north east direction.



Let the vertices i, j be the position (x_i, y_j). Let $\phi_{i,j}(x, y)$ be the continuous piecewise linear function defined by

1) $\phi_{i,j}$ is linear on each triangle.

2) $\phi_{i,j}$ is continuous.

3)

$$\phi_{i,j}(x_k, y_l) = \begin{cases} 1, & i = k, j = l, \\ 0, & \textit{otherwise}\ . \end{cases}$$

Take

$$V_h := \{v \mid v(x,y) = \sum_{i=1}^{N}\sum_{j=1}^{N} v_{i,j}\phi_{i,j}(x,y)\} \subset H_0^1(\Omega).$$

Typically, in implementing the Finite Element Method we have to calculate $A_{i,j} = a(\phi_{i,j}, \phi_{k,l})$. *In this instance we would like to evaluate* $a(\phi_{i,j}, \phi_{k,l})$.

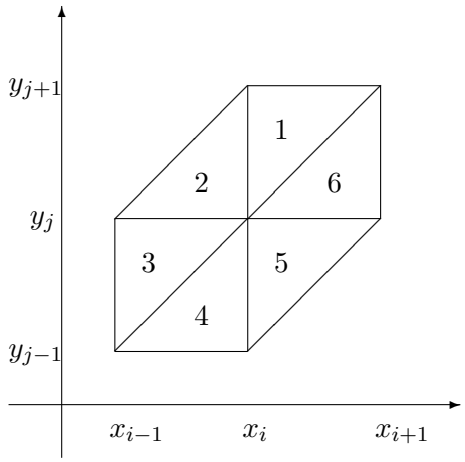
$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v dxdy.$$

We wish to calculate

$$\int_\Omega \nabla\phi_{i,j} \cdot \nabla\phi_{k,l} dxdy = \int_{Support\phi_{i,j}} \nabla\phi_{i,j} \cdot \nabla\phi_{k,l} dxdy,$$

since $\phi_{i,j} = 0$ *else where.*

We order the triangles surrounding our node, i.e, the vertix (i,j). Suppose that $\phi_{i,j}$ takes 1 at (i,j), 0 at the other vertices.



$$\frac{\partial \phi_{i,j}}{\partial x} = 0 \text{ in } \Delta 1, \Delta 4,$$
$$\frac{\partial \phi_{i,j}}{\partial x} = \frac{1}{h} \text{ in } \Delta 2, \Delta 3,$$
$$\frac{\partial \phi_{i,j}}{\partial x} = -\frac{1}{h} \text{ in } \Delta 5, \Delta 6.$$

Similarly,

$$\frac{\partial \phi_{i,j}}{\partial y} = 0 \text{ in } \Delta 3, \Delta 6,$$
$$\frac{\partial \phi_{i,j}}{\partial y} = \frac{1}{h} \text{ in } \Delta 4, \Delta 5,$$
$$\frac{\partial \phi_{i,j}}{\partial y} = -\frac{1}{h} \text{ in } \Delta 1, \Delta 2.$$

Now, we want

$$\epsilon(k,l) := \int_{\text{Support}\phi_{i,j}} \left\{ \frac{\partial \phi_{k,l}}{\partial x} \frac{\partial \phi_{i,j}}{\partial x} + \frac{\partial \phi_{i,j}}{\partial y} \frac{\partial \phi_{k,l}}{\partial y} \right\} dx,$$

for

$$(k,l) \in \{(i,j), (i-1,j), (i+1,j), (i,j+1), \\ (i,j-1), (i-1,j-1), (i+1,j+1)\}.$$

Then we see

$$\epsilon(i,j) = 4 \times \frac{h^2}{2} \frac{1}{h^2} + 4 \times \frac{h^2}{2} \frac{1}{h^2} = 4.$$

For $\epsilon(i+1,j)$ just 2 triangles $\Delta 5, \Delta 6$ to consider

$$\frac{\partial \phi_{i,j}}{\partial x} = -\frac{1}{h} \text{ in } \Delta 5, \Delta 6,$$
$$\frac{\partial \phi_{i,j}}{\partial y} = \begin{cases} 1/h, & \Delta 5, \\ 0, & \Delta 6. \end{cases}$$

$$\epsilon(i+1,j) = \left(-\frac{1}{h^2} \right) \cdot 2 \cdot \frac{h^2}{2} = -1.$$

Since A is symmetric, $\epsilon(i-1,j) = \epsilon(i+1,j)$. By the same procedure as for $\epsilon(i+1,j)$ and by symmetry we have $\epsilon(i,j+1) = \epsilon(i,j-1) = -1$. Also we see that $\epsilon(i+1,j+1) = \epsilon(i-1,j-1) = 0$.

It follows that the finite element approximation of the boundary value problem

$$-\Delta u = 1 \ \text{ in } \ \Omega, \ u = 0 \text{ on } \partial\Omega$$

on this uniform triangulation yields the five point formula

$$4U_{i,j} - 4U_{i-1,j} - U_{i+1,j} - U_{i,j-1} - U_{i,j+1} = h^2 \ , \ i,j = 1,2,..,N$$

for the interior nodal variables.

6.2.1 Element Stiffness Matrix in 2D for a general triangle

We may have to evaluate

$$\int_\Omega \nabla w \cdot \nabla v dx dy = \sum_{K \in \mathcal{T}_h} \int_K \nabla w \cdot \nabla v dx dy$$

for $v, w \in V_h$. Using the expressions

$$v = \sum_{j=1}^{3} v_j \phi_j^K, \ w = \sum_{j=1}^{3} w_j \phi_j^K \text{ on } K,$$

we have

$$\nabla v = \sum_{j=1}^{3} v_j \nabla \phi_j^K,$$

similar for ∇w. So we see that

$$\begin{aligned}
\int_K \nabla v \cdot \nabla w dx &= \int_K \left(\sum_{i=1}^{3} v_i \nabla \phi_i^K \right) \left(\sum_{j=1}^{3} w_j \nabla \phi_j^K \right) dx \\
&= \sum_{i=1}^{3} \sum_{j=1}^{3} v_i w_j \int_K \nabla \phi_i^K \cdot \nabla \phi_j^K dx \\
&= V A^K W^T,
\end{aligned}$$

where $V = (v_1, v_2, v_3)$, $W = (w_1, w_2, w_3)$ and $A^K = (A_{i,j}^K)_{i,j=1,2,3}$,

$$A_{i,j}^K = \int_K \nabla \phi_i^K \cdot \nabla \phi_j^K dx.$$

The matrix A^K is called element stiffness matrix.

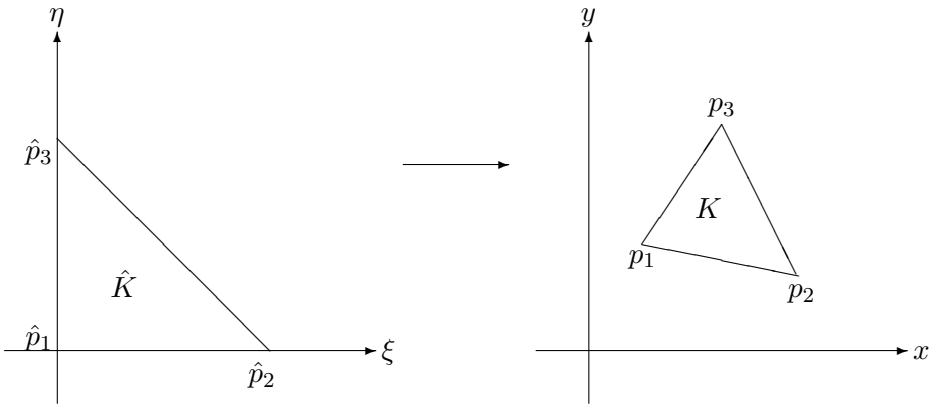
Note that

$$\int_K |\nabla v_h|^2 dxdy = V A^K V^T.$$

Consider the canonical triangle $\hat{K}$ with the vertices $\hat{p}_1 = (0,0), \hat{p}_2 = (1,0), \hat{p}_3 = (0,1)$ in $\xi - \eta$ plane.

$$x(\xi, \eta) = (1 - \xi - \eta)p_1 + \xi p_2 + \eta p_3$$

is the mapping which takes $\hat{p}_i$ to $p_i = (x_i, y_i)$ $(i = 1, 2, 3)$ respectively.



By expanding we see

$$\begin{aligned} x(\xi, \eta) &= p_1 + \xi(p_2 - p_1) + \eta(p_3 - p_1) \\ &= p_1 + [p_2 - p_1, p_3 - p_1] \begin{bmatrix} \xi \\ \eta \end{bmatrix} \\ &= p_1 + \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix}. \\ &= p_1 + J\xi, \end{aligned}$$

where J is the Jacobian mapping $(\xi, \eta) \to (x, y)$, thus, $dxdy = \det J d\xi d\eta$. Note that

J is invertible provided the $p_i, i = 1, 2, 3$ are not collinear.

$$\begin{aligned}\text{Area of } K &= \int_K 1 dx dy \\ &= \int_{\hat{K}} \det J d\xi d\eta \\ &= \frac{1}{2} \det J.\end{aligned}$$

We seek now an inverse mapping.

$$\begin{aligned}\begin{bmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{bmatrix} &= \begin{bmatrix} \frac{\partial x}{\partial \xi}\frac{\partial}{\partial x} + \frac{\partial y}{\partial \xi}\frac{\partial}{\partial y} \\ \frac{\partial x}{\partial \eta}\frac{\partial}{\partial x} + \frac{\partial y}{\partial \eta}\frac{\partial}{\partial y} \end{bmatrix} \\ &= \begin{bmatrix} x_2 - x_1 & y_2 - y_1 \\ x_3 - x_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix} \\ &= J^T \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix}.\end{aligned}$$

$\Longrightarrow$

$$\begin{aligned}\begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix} &= (J^T)^{-1} \begin{bmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{bmatrix} \\ &= \frac{1}{\det J} \begin{bmatrix} y_3 - y_1 & -(y_2 - y_1) \\ -(x_3 - x_1) & x_2 - x_1 \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{bmatrix}.\end{aligned}$$

Thus,

$$\begin{aligned}\hat{v}_h(\xi, \eta) &= v_h(x(\xi, \eta), y(\xi, \eta)) \\ &= \sum_{i=1}^{3} v_i \hat{\phi}_i(\xi, \eta),\end{aligned}$$

where $\hat{\phi}$ is linear and $\hat{\phi}_i(\hat{p}_j) = \delta_{i,j}$.

Note that

$$\frac{\partial \hat{v}_h}{\partial \xi} = v_2 - v_1, \ \frac{\partial \hat{v}_h}{\partial \eta} = v_3 - v_1.$$

Then,

$$
\begin{aligned}
|\nabla v_h|^2 &= (\frac{\partial v_h}{\partial x})^2 + (\frac{\partial v_h}{\partial y})^2 \\
&= \frac{1}{(\det J)^2}\Big\{ \left((y_3-y_1)\frac{\partial v_h}{\partial \xi} - (y_2-y_1)\frac{\partial v_h}{\partial \eta}\right)^2 \\
&\quad + \left(-(x_3-x_1)\frac{\partial v_h}{\partial \xi} + (x_2-x_1)\frac{\partial v_h}{\partial \eta}\right)^2 \Big\} \\
&= \frac{1}{(\det J)^2}\Big\{ |r_3-r_1|^2 \left(\frac{\partial v_h}{\partial \xi}\right)^2 \\
&\quad + |r_2-r_1|^2 \left(\frac{\partial v_h}{\partial \eta}\right)^2 + 2(r_1-r_2)\cdot(r_3-r_1)\frac{\partial v_h}{\partial \xi}\frac{\partial v_h}{\partial \eta} \Big\} \\
&= \frac{1}{(\det J)^2}\{|r_3-r_1|^2(v_2-v_1)^2 + |r_2-r_1|^2(v_3-v_1)^2 \\
&\quad + 2(r_1-r_2)\cdot(r_3-r_1)(v_2-v_1)(v_3-v_1)\} \\
&= \frac{1}{(\det J)^2}(v_1,v_2,v_3)B(v_1,v_2,v_3)^T,
\end{aligned}
$$

where B is given by

$$
B = \begin{pmatrix}
|r_3-r_2|^2 & (r_2-r_3)\cdot(r_3-r_1) & (r_1-r_2)\cdot(r_2-r_3) \\
(r_2-r_3)\cdot(r_3-r_1) & |r_3-r_1|^2 & (r_1-r_2)\cdot(r_3-r_1) \\
(r_1-r_2)\cdot(r_2-r_3) & (r_1-r_2)\cdot(r_3-r_1) & |r_1-r_2|^2
\end{pmatrix}.
$$

Exercise

Let $\mathcal{T}_h$ consist of equilateral triangles whose edges have length h. Calculate the element stiffness matrix for the bilinear form $a(w,v) := \int_\Omega p\nabla w \cdot \nabla v$. Consider the finite element apprioximation of the boundary value problem:-

$$-\nabla \cdot p\nabla u = 1 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega$$

where Ω can be triangulated by a union of uniform equilateral triangles. Hence or otherwise find the generic equation holding at each vertex i of the triangulation

$$a(u_h, \phi_i) = l(\phi_i).$$

Use the notation O for the vertex i and W, E, P, Q, R, S for the surrounding 6 vertices.

6.2.2 Integration formula of linear functions on triangles

Let v be a linear function on a triangle K with nodal values $v_i, i = 1, 2, 3$. It follows that

$$\int_K v dx dy = \sum_{i=1}^{3} v_i \int_K \phi_i = \frac{Area(K)}{3} \sum_{i=1}^{3} v_i$$

since

$$\int_K \phi_i = \frac{Area(K)}{3}.$$

Proof

Consider the reference triangle in the last section. We have that

$$\int_K \phi_i = \int_{\hat{K}} \hat{\phi}_i \, \det J d\xi d\eta$$

where $\hat{\phi}_j(\hat{p}_j) = \delta_{i,j}$ and $\hat{\phi}_j$ is linear. Since $\hat{\phi}_2 = \xi$ we have

$$\int_{\hat{K}} \hat{\phi}_2 = \int_0^1 \int_0^{1-\xi} \xi d\eta d\xi = \int_0^1 \xi(1-\xi) d\xi = \frac{1}{6}.$$

Since $\det J = 2Area(K)$ we have the desired result for ϕ_2. By symmetry the result follows for $\hat{\phi}_3$ and $\hat{\phi}_3$. Since $\hat{\phi}_1 = 1 - \hat{\phi}_2 - \hat{\phi}_3$ we have that

$$\int_{\hat{K}} \hat{\phi}_1 = \frac{1}{2} - \frac{1}{6} - \frac{1}{6} = \frac{1}{6}$$

which completes the proof.

Chapter 7

Finite element error analysis

7.1 One Dimensional Problems

Consider the problem (P) where $f \in L^2(\Omega) = (0,1)$ is given and we seek u such that

$$-u'' + u = f \qquad 0 < x < 1 \tag{7.1}$$

$$u(0) = u(1) = 0. \tag{7.2}$$

The variational formulation is to find $u \in H_0^1(\Omega)$ such that

$$\int_0^1 u'v' + uv \; dx = a(u,v) = l(v) = \int_0^1 fv \; dx \qquad \forall\, v \in H_0^1(\Omega). \tag{7.3}$$

Let V_h be the finite element space of continuous piecewise linear functions on a non-uniform grid. We are in the setting of the Lax-Milgram theorem and

$$a(v,v) = \int_0^1 (v')^2 + v^2 \; dx \equiv \|v\|^2_{H^1(\Omega)}. \tag{7.4}$$

Let $v \in C(\bar{I})$ where I is some bounded interval. Consider the interpolant of v defined by

$$I_h v = \sum_{j=0}^{M} v(x_j)\phi_j(x) \in V_h. \tag{7.5}$$

This is known as Lagrange interpolation. Clearly $I_h u$ is a candidate for Cea's lemma, which would then give us

$$\|u - u_h\|_{H^1(0,1)} \leq \|u - I_h u\|_{H^1(0,1)}. \tag{7.6}$$

Remark 7.1.1. **A simple interpolation estimate** Let $p_1(\cdot)$ be a linear polynomial which interpolates some function $f(\cdot)$ at $x = x_0$ and $x = x_1$, i.e. $p_1(x_0) = f(x_0)$ and $p_1(x_1) = f(x_1)$. Suppose that $f \in C^2(I)$. Then

$$\begin{aligned} f(x) - p_1(x) &= \frac{1}{2}(x - x_0)(x - x_1) f''(\xi_x) \qquad x_0 \leq \xi_x \leq x_1, \\ \|f(x) - p_1(x)\|_{L^\infty(I)} &\leq \frac{h^2}{8} \max_{\xi \in [x_0, x_1]} |f''(\xi)| \qquad x \in (x_0, x_1) \\ \Rightarrow \|v - I_h v\|_{L^\infty(I)} &\leq \frac{h^2}{8} \|v''\|_{L^\infty(I)}. \end{aligned}$$

However recalling that $V = H^1$, we note that this bound requires knowledge of $\|v''\|_{L^\infty(I)}$! But the interpolation bound clearly is an indication that the finite element method should converge.

Lemma 7.1.2. Interpolation error *Let V_h be the piece-wise linear finite element space on a unforom grid with mesh size h. Suppose that $v \in H^2(0,1)$ and $I_h v$ interpolates v to V_h as before. Then*

$$\|v - I_h v\|_{L^2(0,1)} \leq \left(\frac{h}{\pi}\right)^2 \|v''\|_{L^2(0,1)} \tag{7.7}$$

$$\|(v - I_h v)'\|_{L^2(0,1)} \leq \frac{h}{\pi} \|v''\|_{L^2(0,1)}. \tag{7.8}$$

Proof: Consider a subinterval (x_{i-1}, x_i), $1 \leq i \leq N$, $h = 1/N$. Define $\eta(x) = v(x) - I_h v(x)$ so $\eta \in H^2(x_{i-1}, x_i)$ and $\eta(x_i) = \eta(x_{i-1}) = 0$. Then η can be expanded as a Fourier series:

$$\begin{aligned} \eta(x) &= \sum_{k=1}^{\infty} a_k \sin\left(k\pi \frac{(x - x_i)}{h}\right) \\ \int_{x_{i-1}}^{x_i} \eta^2(x)\, dx &= \frac{h}{2} \sum_{k=1}^{\infty} |a_k|^2. \end{aligned}$$

Differentiating the Fourier series twice we see that the coefficients of η' are $(k\pi a_k)/h$

and the coefficients of η'' are $-(k\pi/h)^2 a_k$. Thus

$$\int_{x_{i-1}}^{x_i} \eta'^2(x)\ dx = \frac{h}{2}\sum_{k=1}^{\infty} |a_k|^2 \frac{k^2\pi^2}{h^2}$$
$$\int_{x_{i-1}}^{x_i} \eta''^2(x)\ dx = \frac{h}{2}\sum_{k=1}^{\infty} |a_k|^2 \frac{k^4\pi^4}{h^4}$$

Note that $\eta''(x) = v''(x) - I_h v''(x) = v''(x)$. Now

$$\int_{x_{i-1}}^{x_i} \eta^2(x)\ dx = \frac{h}{2}\sum_{k=1}^{\infty} |a_k|^2 \leq \frac{h}{2}\sum_{k=1}^{\infty} |a_k|^2 \frac{k^4\pi^4}{h^4}\frac{h^4}{\pi^4} = \frac{h^4}{\pi^4}\int_{x_{i-1}}^{x_i} \eta''^2(x)\ dx$$
$$\int_{x_{i-1}}^{x_i} \eta^2(x)\ dx \leq \frac{h^2}{\pi^2}\int_{x_{i-1}}^{x_i} \eta''^2(x)\ dx$$
$$\|\eta\|^2_{L^2(\Omega)} = \sum_{i=1}^{N}\int_{x_{i-1}}^{x_i} \eta^2(x) \leq \frac{h^4}{\pi^4}\|v''\|^2_{L^2(\Omega)}$$
$$\|\eta'\|^2_{L^2(\Omega)} \leq \frac{h^2}{\pi^2}\|v''\|^2_{L^2(\Omega)}.$$

■

Lemma 7.1.3. *Suppose that the solution of (P) satisfies $u \in H^2(I)$. Then*

$$\|u - u_h\|_{H^1(I)} \leq Ch\|u''\|_{L^2(I)}. \tag{7.9}$$

Proof:

$$\|u - u_h\|^2_{H^1(I)} \leq \|u - I_h u\|^2_{H^1(I)} = \|u - I_h u\|^2_{L^2(I)} + \|(u - I_h u)'\|^2_{L^2(I)}$$
$$\leq \frac{h^4}{\pi^4}\|u''\|^2_{L^2(I)} + \frac{h^2}{\pi^2}\|u''\|^2_{L^2(I)} \leq \frac{2h^2}{\pi^2}\|u''\|^2_{L^2(I)}$$

as $h < 1$.

■

The fact that although our problem is posed in H^1 and the solution is in H^2 is called elliptic regularity. In our example since $f \in L^2$ it is easy to show the desired

regularity:

$$\begin{aligned}
\|u\|_{H^1}^2 = a(u,u) = l(u) = \int_0^1 fu\ dx \leq \|f\|_{L^2}\|u\|_{L^2} \\
\Rightarrow \|u\|_{L^2}^2 \leq \|u\|_{H^1} \leq \|f\|_{L^2}\|u\|_{L^2} \\
\Rightarrow \|u\|_{L^2} \leq \|f\|_{L^2} \\
\Rightarrow \|u'\|_{L^2}^2 \leq \|u\|_{H^1}^2 \leq \|f\|_{L^2}\|u\|_{L^2} \leq \|f\|_{L^2}^2 \\
\Rightarrow \|u'\|_{L^2} \leq \|f\|_{L^2}
\end{aligned}$$

and since $u'' = u - f$

$$\|u''\|_{L^2} = \|u - f\|_{L^2} \leq \|u\|_{L^2} + \|f\|_{L^2} \leq 2\|f\|_{L^2}.$$

Thus we can conclude that $\|u - u_h\|_{H^1(0,1)} \leq Kh\|f\|_{L^2(0,1)}$ where K is independent of u, h, f.

Remark 7.1.4.

1. This in an a priori bound. The number on the right hand side is usually larger than the error. The order with respect to h is correct.
2. It says that the discretisation is convergent - $\lim_{h\to 0} \|u - u_h\|_{H^1} = 0$.
3. It says that if h is reduced by a factor of a half say, then one expects the error to be reduced by a factor of one half.
4. It is important for confidence in numerical calculations to have such error bounds.

7.1.1 Bounding the L^2 error

The above work has enabled us to get an H^1 bound on the error $e = u - u_h$ (and hence also an L^2 bound). But can the L^2 bound be improved? Yes - by using an Aubin-Nitsche duality argument. For example, consider the problem

$$\begin{aligned}
-(pu')' + qu = f \qquad x \in (a,b) \\
u(a) = u(b) = 0.
\end{aligned}$$

Regularity theory implies that $\|u''\|_{L^2(a,b)} \leq C\|f\|_{L^2(a,b)}$. Now consider the problem

$$\begin{aligned}
-(pw')' + qw = e \qquad x \in (a,b) \\
w(a) = w(b) = 0
\end{aligned}$$

where $e = u - u_h$. Then from our standard results there exists a unique w solving this problem and $\|w''\|_{L^2(a,b)} \leq C\|e\|_{L^2(a,b}$. Now we have that

$$\begin{aligned}
\|e\|^2_{L^2(a,b)} &= \int_a^b e^2 \\
&= (e,e) \\
&= a(w,e) \\
&= a(e,w) \\
&= a(u-u_h,w) \\
&= a(u-u_h, w - I_h w) \quad (a(u-u_h, I_h w) = 0 \text{ by Galerkin orthogonality}) \\
&\leq \gamma \|u-u_h\|_{H^1(a,b)} \|w - I_h w\|_{H^1(a,b)} \\
&\leq \gamma (Ch\|f\|_{L^2(a,b)})(Ch\|w''\|_{L^2(a,b)}) \\
&\leq \overline{C} h \|f\|_{L^2(a,b)} h \|e\|_{L^2(a,b)} \\
\Rightarrow \quad \|e\|_{L^2(a,b)} &\leq Ch^2 \|f\|_{L^2(a,b)}.
\end{aligned}$$

7.2 Two Dimensional Problem

Let $\mathcal{T}_h$ be a triangulation of Ω. Suppose that h is the maximum diameter of any triangle in the triangulation. We assume that the smallest angle of the triangulation is bounded below independently of h.

Theorem 7.2.1. Interpolation error *Let V_h be the space of continuous piece-wise linear functions on $\mathcal{T}_h$. Let $v \in H^2(\Omega)$. Then*

$$\|v - I_h v\|_{H^1(\Omega)} \leq k_1 h |v|_2 \tag{7.10}$$

$$\|v - I_h v\|_{L^2(\Omega)} \leq k_2 h^2 |v|_2 \tag{7.11}$$

where k_1, k_2 are independent of h and u and $|u|_2$ is the L^2 norm of all second derivatives.

Here $|\cdot|_2$ is the semi-norm:

$$|v|_2^2 = \int_\Omega \left| \frac{\partial^2 v}{\partial {x_1}^2} \right| + 2 \left| \frac{\partial^2 v}{\partial x_1 \partial x_2} \right|^2 + \left| \frac{\partial^2 v}{\partial {x_2}^2} \right| \, dx_1 \, dx_2. \tag{7.12}$$

Remark 7.2.2. The order of approximation in the H^1 norm is 1, or linear, with respect to h, and in the L^2 norm it is 2, or quadratic, with respect to h. We hope to reproduce these rates of convergence using our finite element discretisation.

Our variational problems for second order elliptic equation are set in the Sobolev space $H^1(\Omega)$ (or $H_0^1(\Omega)$ etc.). From the variational formulation we do not know that $|u|_2$ is bounded! The fact that the solution of the equation lies in $H^2(\Omega)$ would be a regularity result. It is not true in general for polygons. It is true when the boundary is smooth or when the domain Ω is convex.

Example 7.2.3. *Let $\Omega = (0,1) \times (0,1)$ and consider the problem:*

$$\begin{aligned} -\Delta w &= g \qquad \text{in } \Omega \\ w &= 0 \qquad \text{on } \partial\Omega \end{aligned}$$

Then there exists a unique $w \in H_0^1(\Omega)$ when $g \in L^2(\Omega)$. Suppose that $w \in H^2(\Omega) \cap H_0^1(\Omega)$. Then

$$\begin{aligned} \|\Delta w\|^2_{L^2(\Omega)} &= \int_\Omega \left(\frac{\partial^2 w}{\partial {x_1}^2} + \frac{\partial^2 w}{\partial {x_2}^2} \right)^2 \, dx_1 \, dx_2 \\ &= \int_\Omega \left(\frac{\partial^2 w}{\partial {x_1}^2} \right)^2 + 2 \left(\frac{\partial^2 w}{\partial {x_1}^2} \frac{\partial^2 w}{\partial {x_2}^2} \right) + \left(\frac{\partial^2 w}{\partial {x_2}^2} \right)^2 \, dx_1 \, dx_2 \\ \int_\Omega \frac{\partial^2 w}{\partial {x_1}^2} \cdot \frac{\partial^2 w}{\partial {x_2}^2} \, dx_1 \, dx_2 &= \int_\Omega \left| \frac{\partial^2 w}{\partial x_1 \partial x_2} \right|^2 \, dx_1 \, dx_2 \end{aligned}$$

where we have used integration by parts and the boundary conditions for the last part. Thus for the square $\|\Delta w\|^2_{L^2(\Omega)} = |w|_2^2$ when $w \in H^2(\Omega) \cap H_0^1(\Omega)$. So for this elliptic boundary value problem

$$\|w\|_{H^2(\Omega)} \leq \|g\|_{L^2(\Omega)}. \tag{7.13}$$

Equation (7.13) is an example of an elliiptic regularity result. If we wish to apply interpolation error bounds to a finite element discretisation of an elliptic boundary value problem we need H^2 regularity.

Example 7.2.4. *Let $\Omega = (0,1) \times (0,1)$ and consider the problem:*

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

The weak formulation is the problem (P): find $u \in V$ such that

$$a(u,v) = l(v) \qquad \forall\, v \in V.$$

We set $V = H_0^1(\Omega), a(u,v) = \int_\Omega \nabla u \nabla v, l(v) = \int_\Omega fv$. *The finite element problem is* (P_h) *find* $u_h \in V_h$ *such that*

$$a(u_h, v_h) = l(v_h) \qquad \forall\, v_h \in V_h.$$

Since we have a convex domain in $\mathbb{R}^2$ *we have that* $u \in H^2(\Omega) \Rightarrow u \in C(\overline{\Omega})$ *(by the Sobolev embedding theorem). Thus it follows that* $I_h u$ *is well defined. Suppose that we have the interpolation error bound*

$$\|v - I_h v\|_{H^1(\Omega)} \leq k_1 h \|v\|_{H^2(\Omega)} \qquad \forall\, v \in H^2(\Omega).$$

Then using the Galerkin orthogonality $a(u - u_h, v_h) = 0 \forall\, v_h \in V_h$ *we have that*

$$\begin{aligned}
a(u - u_h, u - u_h) &= a(u - u_h, u - I_h u) + a(u - u_h, I_h u - u_h) \\
&= a(u - u_h, u - I_h u) \\
\|\nabla(u - u_h)\|^2_{L^2(\Omega)} &= \int_\Omega \nabla(u - u_h)\, \nabla(u - u_h) \\
&\leq \|\nabla(u - u_h)\|_{L^2(\Omega)} \|\nabla(u - I_h u)\|_{L^2(\Omega)} \\
\Rightarrow \quad \|\nabla(u - u_h)\|_{L^2(\Omega)} &\leq \|\nabla(u - I_h u)\|_{L^2(\Omega)}
\end{aligned}$$

Recalling the Poincaré inequality $\|v\|_{L^2(\Omega)} \leq C_* \|\nabla v\|_{L^2(\Omega)}\ \forall\, v \in H_0^1(\Omega)$, *we have that*

$$\begin{aligned}
\|u - u_h\|_{H^1(\Omega)} &\leq C' \|u - I_h u\|_{H^1(\Omega)} \\
&\leq C' k_1 h \|v\|_{H^2(\Omega)} \\
&\leq \overline{C} h \|f\|_{L^2(\Omega)}
\end{aligned}$$

where $\overline{C}$ *depends only on the domain* Ω *and the fact that we are using a piecewise linear finite element approximation on a triangulation. We note that* $\overline{C}$ *is independent of* h *and* f.

As in the one dimensional case we can apply the Aubin-Nitsche trick to show the the error in the L^2 *norm is of higher order than the error in the* H^1 *norm.*

Note that the dual problem is to find $w(g) \in H_0^1(\Omega)$ *such that*

$$(\nabla v, \nabla w(g)) = (g, v) \quad \forall v \in H_0^1(\Omega)$$

which, because the bilinear form is symmetric, is the variational formulation of

$$\begin{aligned}
-\Delta w &= g \qquad \text{in } \Omega \\
w &= 0 \qquad \text{on } \partial\Omega
\end{aligned}$$

for which we have a regularity result.

Let $e = u - u_h$. Consider $-\Delta w = e_h$ in $\Omega, w = 0$ on $\partial\Omega$. Elliptic regularity implies that $\|w\|_{H^2(\Omega)} \leq C_1 \|e_h\|_{L^2(\Omega)}$. Now

$$\begin{aligned}
\|u - u_h\|^2_{L^2(\Omega)} &= a(e_h, e_h) \\
&= (e_h, -\Delta w) \\
&= a(e_h, w) \\
&= a(u - u_h, w - I_h w) \\
&\leq \|\nabla(u - u_h)\|_{L^2(\Omega)} \|\nabla(w - I_h w)\|_{L^2(\Omega)} \\
&\leq \overline{C} h \|f\|_{L^2(\Omega)} k_1 h \|w\|_{H^2(\Omega)} \\
&\leq \overline{C} k_1 h^2 \|e_h\|_{L^2(\Omega)} \|f\|_{L^2(\Omega)} \\
\Rightarrow \|e_h\|_{L^2(\Omega)} &\leq C k_1 h^2 \|f\|_{L^2(\Omega)} \\
&= k_2 h^2 \|f\|_{L^2(\Omega)}
\end{aligned}$$

where we have used Galerkin orthogonality.

7.3 Summary

To summarise how we obtain these error bounds:

$$\text{Galerkin orthogonality + elliptic regularity} \Rightarrow H^1 \text{ error bound}$$
$$\text{Aubin-Nitsche + regularity} \Rightarrow L^2 \text{ error bound}$$

Chapter 8

FEM:-Miscellaneous

8.1 Inhomogeneous Dirichlet boundary conditions

Consider the elliptic problem

$$-\nabla \cdot (p\nabla u) + qu = f \qquad x \in \Omega \tag{8.1}$$

$$u = g \qquad x \in \partial\Omega \tag{8.2}$$

where Ω is a bounded open subset of $\mathbb{R}^2$. We assume that the data p, q, f, g are sufficiently smooth and that

$$p_M \geq p(x) \geq p_0 > 0 \qquad \forall\, x \in \Omega$$

$$q_M \geq q(x) \geq q_0 > 0 \qquad \forall\, x \in \Omega$$

Thus, see Chapter 4, our variational problem is:

(P_g) Find $u \in V_g$ such that

$$a(u, v) = l(v)\ \forall\, v \in V_0.$$

The bilinear form is symmetric so there is an energy and associated minimisation problem with $J(v) = \frac{1}{2}a(v, v) - l(v)$:-

(M_g) Find $u \in V_g$ *s.t.*

$$J(u) \leq J(v) \qquad \forall\, v \in V_g.$$

The finite element approach is then to construct $V_0^h := \{v \in C(\Omega) : v|_K \text{ is linear}, v|_{\partial\Omega} = 0\}$ and $V_g^h := \{v \in C(\Omega) : v|_K \text{ is linear}, v|_{\partial\Omega} = g_h\}$ and to consider the problems:

(P_g^h) Find $u_h \in V_g^h$ such that

$$a(u_h, v) = l(v) \quad \forall v \in V_0^h.$$

(M_g^h) Find $u_h \in V_g^h$ such that

$$J(u_h) \leq J(v) \qquad \forall\, v \in V_g^h.$$

8.2 Other finite elements

There is an enormous zoo of finite elements. One generalisation is to increase the order of the polynomial in each triangle. For example a quadratic function in two space dimensions has 6 parameters which can be fixed by the values of function at the three vertices and the three mid points. The finite element space is then defined as $V_h := \{v \in C(\bar{\Omega}) : v|_K \text{ is quadratic } \forall K \in \mathcal{T}_h\}$.

Another possibility is to use rectangular finite elements. For example a simple element is based on bilinear functions $v = a + bx + cy + dxy$ on a rectangle which is fixed by the four vertex values of v.

On the other hand one is also interested in higher space dimensions. The genralization to three space dimensions is to use a tetrahedron whose 4 vertices are not coplanar. Triangles and tetrahedra are examples of simplices. A simplex in $\mathbb{R}^n$ has $(n+1)$ vertices. A linear function is then fixed by the values at the vertices.

8.3 Programming

1. Mesh generation is fundamental. Sometimes a macro-triangulation with coarse triangles (large diameter) may be defined by hand. Then a refinement method may be employed. For example a triangle may be subdivided into 4 congruent triangles in a natural way by using the midpoints of each edge to define the

vertices of new triangles. On the other hand a triangle may be subdivided into two by adding an edge joining the midpoint of the largest edge to the opposite vertex.

2. Matrices and vectors defining the linear algebraic equations are assembled element by element. That is each element is visited once, the element stiffness matrix is calculated as well as the contributions to the right hand side and then added to the global matrix and righthand side vector.

8.4 Higher Order Equations

So far we have only considered the finite element approximation to second order partial differential equations. But what about higher order equations? In this section we will see how to tackle these types of problems.

4th Order Problem in 1D

Consider the problem

$$\begin{aligned} (u'')'' - (pu')' + qu &= f \qquad x \in (0,1) \\ u(0) = u(1) &= 0 \\ u'(0) = u'(1) &= 0 \end{aligned}$$

As usual we multiply by a test function v and integrate, using integration by parts on the terms with derivatives:

$$\begin{aligned} (f,v) &= \int_0^1 fv \; dx \\ &= \int_0^1 (u'')''v - (pu')'v + quv \; dx \\ &= [u'''v]_0^1 - \int_0^1 u'''v' \; dx - [pu'v]_0^1 + \int_0^1 pu'v' \; dx + \int_0^1 qv \; dx \\ &= [u'''v]_0^1 - [u''v']_0^1 + \int_0^1 u''v'' \; dx - [pu'v]_0^1 + \int_0^1 pu'v' \; dx + \int_0^1 qv \; dx \end{aligned}$$

Set $V = \{v \in H^2(0,1) : v(0) = v(1) = v'(0) = v'(1) = 0\}$. Then

$$(f,v) = \int_0^1 u''v'' \; dx + \int_0^1 pu'v' \; dx + \int_0^1 qv \; dx$$

so our variational problem is to find $u \in V$ such that

$$\begin{aligned} a(u,v) &= l(v) \\ a(u,v) &= \int_0^1 u''v'' + pu'v' + qv \; dx \\ l(v) &= \int_0^1 fv. \end{aligned}$$

We now seek to check the conditions of the Lax-Milgran theorem. Let us assume that $p, q \in C(0,1)$ and that $p, q \geq 1$. Then

$$\begin{aligned} a(v,v) &\geq \int_0^1 v''^2 + v'^2 + v^2 \; dx = \|v\|_{H^2(0,1)} \\ a(u,v) &\leq \max(|p|,|q|) \int_0^1 |u''||v''| + |u'||v'| + |u||v| \; dx \\ &\leq C\|u\|_{H^2(0,1)}\|v\|_{H^2(0,1)}. \end{aligned}$$

Similarily we could have consider the above problem with the following boundary conditions:

$$\begin{aligned} u(0) = u(1) = 0 \\ u''(0) = u''(1) = 0 \end{aligned}$$

where again we would set $V = \{v \in H^2(0,1) : v(0) = v(1) = v'(0) = v'(1) = 0\}$.

In these examples our test space V is a subset of H^2. To approximate this with our finite element method we would need to use higher order elements than we have done previous (until now we have only considered linear elements). However, we can tackle a modified problem using only linear elements. Consider the problem

$$\begin{aligned} -(u'')'' - u'' &= f \qquad x \in (0,1) \\ u(0) = u(1) &= 0 \\ u''(0) = u''(1) &= 0 \end{aligned}$$

Let $v = -u''$. Then we can rewrite our fourth order problem as a system of second order problems:

$$\begin{aligned} -u'' &= v \qquad x \in (0,1) \\ -v'' + v &= f \qquad x \in (0,1) \\ u(0) = u(1) &= 0 \\ v(0) = v(1) &= 0 \end{aligned}$$

We can now discretise these two problems using piecewise linear elements, remembering that we have this coupled system for u, v to solve.

Chapter 9

Finite element spaces

9.1 Definition of a finite element

Definition 9.1.1. A finite element is a triple (K, P_K, N_K), where

- (i) $K \subset \mathbb{R}^n$ is the closure of a bounded Lipshitz domain.
- (ii) P_K is a finite dimensional space of functions: $K \to R$.
- (iii) N_K is a basis for $(P_K)'$, the dual space of P_K.

Remark 9.1.2.

- dim $(P_K)' =$ dim P_K
- Let $d =$ dim P_K and $N_K = \{N_1, N_2, ..N_d\} \subset (P_K)'$ then the following statements are equivalent: (i) N_K is a basis for $(P_K)'$

  (ii)for all $p \in P_K, N_i(p) = 0$ for all $i = 1, 2..d$ implies $p = 0$.

  We say that if these hold then N_K determines P_K.

Example 9.1.3. *In* $\mathbb{R}^2$ *we set* K *to be the triangle with vertices (nodes)* $z_i, i = 1, 2, 3$*,* P_K *to be the set of polynomials of dgree* 1 *in two varibles* x_1, x_2 *on* K *and* $N_K = \{N_1, N_2, N_3\}$ *where* $N_i(p) = p(z_i), i = 1, 2, 3$.

Let M *be the* 3×3 *matrix whose* i*th row is* $(1, z_i^T)$ *which is nonsingular since* $det(M)$=$2Area(K)$*. It follows that the equations* $p(z_i) = 0$ *for the polynomial coefficients of* $p(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$ *have a unique zero solution so that* N_K *determines* P_K.

In this special case, we denote the nodal basis for P_K by $\lambda_j, j = 1, 2, 3$ with $\lambda_j(z_i) = \delta_{ij}$. Note that $\lambda_1 + \lambda_2 + \lambda_3 - 1 = 0$ (since this holds at each of the three nodes and a plane is uniquely determined by its values at three non-collinear points). Hence $\sum_{j=1}^{3} \lambda_j = 1$. The λ_j are sometimes called barycentric coordinates since (again by uniqueness of linear interpolation at 3 points), any x may be written as the weighted sum: $x = \sum_{j=1}^{3} \lambda_j(x) z_j$.

Given an element $(K, P_K, N_)K)$, let $N_K = \{N_1, ..., N_d\}$ and choose $\{pi : i = 1, ..., d\}$ to be the nodal basis for P_K then for all v in the domain of $N_i, \forall i$, we define the (local) interpolant

$$I_K v = \sum_{i=1}^{d} (N_i v) p_i \in P_K.$$

This interpolates v in the sense that

$$N_j(I_K v) = N_j(v) j = 1, ..., d.$$

9.2 Construction of a finite element space on a mesh

Definition 9.2.1. A mesh on $\Omega \subset \mathbb{R}^n$ is a subdivision T of Ω into closed triangles (n=2), tetrahedra (n=3) or simplices (general n) with the properties:

- $\bar{\Omega} = \cup_{K \in T} K$ and the elements $K \in T$ have pairwise disjoint interiors.
- If $K, K' \in T$ and $K \neq K'$ then $K \cap K'$ is either empty or a vertex or face of each element.

Note that these assumptions implicitly assume that Ω is polyhedral. Curved elements are also possible. We also call K an element.

Let

$$h_K = max\{|x - y| : x, y \in K\} \text{ and } h = \max h_K,\ K \in T.$$

We usually write T_h for T and consider a sequence of meshes with $h \to 0$. Also, let $\rho_K =$ the diameter of largest ball in $\mathbb{R}^n$ contained inside K. Note $\rho_K \leq h_K$.

By using I_K on each K we can define a global interpolation operator I_h for functions on Ω by

$$(I_h v)|_K = I_K v \forall\ K \in T_h.$$

(Note, $I_h v$ may not be well-defined on the interface between 2 neighboring elements.)

Definition 9.2.2. The finite element space is called H^m conforming if $I_h v \in H^m(\Omega)$ for $v \in C(\Omega)$.

We denote by $(\hat{K}, \hat{P}, \hat{N})$ a standard reference element with the diamter of $\hat{K}$ being order 1.

9.3 Approximation theory

We wish to consider the approximation properties of the finite element space by bounding the interpolation error.

Throughout, to make statements simpler to write down, if $A(h), B(h)$ are mesh dependent quantities, we write $A(h) \lessapprox B(h)$ if $A(h)/B(h)$ is bounded independently of h.

Definition 9.3.1. The mesh is called **regular** provided

$$h_K \lessapprox \rho_K, \ \forall K \in T_h \text{ as } h \to 0.$$

Theorem 9.3.2. *Assume the mesh sequence T_h is regular. Then if $m \geq 2$ and*

$$\mathbb{P}_{m-1}(n) \subseteq \hat{P}$$

for $i = 0, ..., m$, we have for each $K \in Th$,

$$||v - I_K v||_{H^i(K)} \lessapprox h_K^{m-i} |v|_{H^m(K)}, \quad \textit{for all} \ \ v \in H^m(K).$$

Moreover if the element is H^i conforming with $0 \leq i \leq m$, then

$$||v - I_h v||_{H^i(\Omega)} \lessapprox h^{m-i} |v|_{H^m(\Omega)}.$$

Chapter 10

Parabolic problems

10.1 Function spaces

Let X be a Banach space with norm$||\cdot||_X$.

Definition 10.1.1. The space $L^p(0,T;X)$ consists of all measurable functions $\eta : [0,T] \to X$ with

$$||\eta||_{L^p(0,T;X)} := (\int_0^T ||\eta(t)||_X^p dt)^{1/p} < \infty, \quad 1 \leq p < \infty$$

and

$$||\eta||_{L^\infty(0,T;X)} := \operatorname*{ess\,sup}_{0\leq t\leq T} ||\eta(t)||_X < \infty.$$

Definition 10.1.2. The space $C([0,T];X)$ consists of all continuous functions $\eta : [0,T] \to X$ with

$$||\eta||_{C([0,T];X)} := \max_{0\leq t\leq T} ||\eta(t)||_X < \infty.$$

Definition 10.1.3. Let $\eta \in L^1(0,T;X)$. We say that ξ is the weak derivative of η written

$$\eta' = \xi$$

provided

$$\int_0^T \phi'(t)\eta(t)dt = -\int_0^T \phi(t)\xi(t)dt$$

for all test functions $\phi \in C_0^\infty(0,T)$.

Definition 10.1.4. The space $H^1(0,T);X)$ consists of all functions $\eta \in L^2(0,T;X)$ with a weak derivative $\eta' \in L^2(0,T;X)$.

10.2 Parabolic equation

Let Ω be a bounded domain in $\mathbb{R}^d,\ d = 1, 2$. Let $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ and $(\cdot,\cdot)$ denote the $L^2(\Omega)$ inner product. Let f, p, q and u_0 be given functions in $C(\bar{\Omega})$ with $p(x) \geq p_m > 0, q(x) \geq 0$ for all $x \in \Omega$. Consider the following initial value problem

$$u_t = \nabla(p(x)\nabla u) - q(x)u + f(x), \quad x \in \Omega, \quad 0 < t \leq T \quad ,$$
$$u = 0,\ x \in \partial\Omega, \quad u(x,0) = u_0(x)\ x \in \Omega.$$

This may be formulated as a variational problem of the form:

Find $u \in H^1(0,T;H) \cap L^2(0,T;V)$ such that

$$(u_t, v) + a(u,v) = (f,v) \quad \forall v \in V \text{ and a.e. } \ t \in (0,T)$$

and

$$u(0) := u_0.$$

Here $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is the bounded, coercive symmetric bilinear form

$$a(w,v) := \int_\Omega p\nabla w \cdot \nabla v + qwv.$$

This follows multiplying by a test function $v \in H_0^1(\Omega)$, integrating by parts and using the boundary conditions on v.

It is convenient to recall the Poincare inequality

$$||v||_H \leq C_*||v||_V,\ \forall v \in V$$

where $||v||_H := ||v||_{L^2(\Omega)}$ and for $H_0^1(\Omega)$ we take the norm $||v||_V := ||\nabla v||_{L^2(\Omega)}$.

Lemma 10.2.1. *There is a positive K independent of u, u_0 and f such that*

$$||u(t)||_H \leq e^{-Kt}||u_0||_H + \frac{1}{K}(1 - e^{-Kt})||f||_H.$$

Proof. Taking $v = u$ in the variational formulation yields

$$(u_t, u) + a(u, u) = (f, u)$$

and applying the Cauchy Schwarz inequality on the RHS and the Poincare inequality on the LHS we find

$$\frac{1}{2}\frac{d}{dt}||u||_H^2 + \frac{p_m}{(C_*)^2}||u||_H^2 \leq ||f||_H||u||_H$$

which implies that ($K := \frac{p_m}{(C_*)^2}$) (since $\frac{1}{2}\frac{d}{dt}||u||_H^2 = ||u||_H\frac{d}{dt}||u||_H$)

$$\frac{d}{dt}||u||_H + K||u||_H \leq ||f||_H$$

and using an integrating factor

$$\frac{d}{dt}(e^{Kt}||u||_H) \leq e^{Kt}||f||_H.$$

Integrating yields the desired inequality.

□

10.2.1 Semi-discrete finite element approximation

Let V_h be a finite element subspace of V. Here we may consider piecewise linear functions on a partition of Ω ($d = 1$) or on a triangulation of a polygonal Ω ($d = 2$). Find $u(t) \in V_h, t \in [0, T]$ such that

$$((u_h)_t, v_h) + a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h \text{ and a.e. } \; t \in (0, T)$$

and

$$u_h(0) := u_0^h$$

where $u_0^h \in V_h$ is an approximation to u_0.

Lemma 10.2.2. *There is a positive K independent of u_h, u_0^h and f such that*

$$||u_h(t)||_H \leq e^{-Kt}||u_0^h||_H + \frac{1}{K}(1 - e^{-Kt})||f||_H.$$

Proof. This is the same as the proof for the continuous problem. □

Let $\{\phi_j\}_{j=1}^J$ be a basis for V_h. Let A, M, b be defined by

$$A_{ij} = a(\phi_j, \phi_i), M_{ij} = (\phi_i, \phi_j), F_i = (f, \phi_i).$$

Then writing $u_h(t) := \sum_{j=1}^J U(t)_j \phi_j$ we have that

$$M\frac{dU}{dt} + AU = F \quad , \quad U(0) = U_0$$

where $u_0^h = \sum_{j=1}^J U(t)_0 \phi_j$.

10.2.2 Fully discrete finite element approximation

A fully discrete numerical scheme for the initial value problem may be obtained by using difference quotients to approximate the time derivative.

The idea is to construct a sequence $\{u_h^n\}_{n=0}^N$ with $\Delta t := T/N$ and $u_h^n \in V_h$ approximating $u(n\Delta t)$.

The backward Euler scheme in time is:-

$$(\frac{u_h^{n+1} - u_h^n}{\Delta t}, v_h) + a(u_h^{n+1}, v_h) = (f, v_h) \quad \forall v_h \in V_h$$

with u_h^0 given.

The system of linear algebraic equations resulting from the backward Euler scheme is easily seen to be

$$LU^{n+1} = MU^n + b$$

where $L = M + \Delta t A$ and $b_i = \Delta t(f, \phi_i)$.

Lemma 10.2.3. Gronwall inequality

Let $z_k \geq 0, k = 0, 1, 2, ...n, ..N$ satisfy

$$z_{n+1} \leq \lambda z_n + \lambda G$$

where $\lambda > 0$ and $G \geq 0$. Then

$$z_n \leq \lambda^n z_0 + \frac{1 - \lambda^n}{1 - \lambda}\lambda G \ , \lambda \neq 1$$
$$z_n \leq z_0 + nG \ , \ \lambda = 1.$$

Proof. This follows by induction. We consider the case $\lambda \neq 1$. For a particular k, if

$$z_{k+1} \leq \lambda z_k + \lambda G$$

and

$$z_k \leq \lambda^k z_0 + \frac{1-\lambda^k}{1-\lambda}\lambda G$$

it follows that

$$z_{k+1} \leq \lambda(\lambda^k z_0 + \frac{1-\lambda^k}{1-\lambda}\lambda G) + \lambda G$$

and rearranging we find

$$z_{k+1} \leq \lambda^{k+1} z_0 + \frac{1-\lambda^k}{1-\lambda}\lambda G.$$

□

Lemma 10.2.4. *For the backward Euler scheme, the following stability bound holds*

$$||u_h^n||_H^2 \leq (1+K\Delta t)^{-n}||u_0||_H^2 + \frac{1}{K^2}(1-(1+K\Delta t)^{-n})||f||_H^2 \ \forall n \geq 0.$$

Proof. Taking $v_h = u_h^{n+1}$ in the backward Euler scheme yields

$$(\frac{u_h^{n+1}-u_h^n}{\Delta t}, u_h^{n+1}) + a(u_h^{n+1}, u_h^{n+1}) = (f, u_h^{n+1})$$

and rearranging we find

$$\frac{1}{2}(||u_h^{n+1}||_H^2 - ||u_h^n||_H^2 + ||u_h^{n+1}-u_h^n||_H^2) + \Delta t K||u_h^{n+1}||_H^2 \leq \frac{1}{2K}\Delta t||f||_H^2 + \frac{K}{2}\Delta t||u_h^{n+1}||_H^2$$

which yields

$$(1+K\Delta t)||u_h^{n+1}||_H^2 \leq ||u_h^n||_H^2 + \Delta t\frac{1}{K}||f||_H^2$$

from which using the Gronwall inequality we obtain the desired result. □

Lemma 10.2.5. *In the case $f = 0$ for the backward Euler scheme, the following stability bound holds*

$$||u_h^n||_H^2 \leq (1+2K\Delta t)^{-n}||u_0||_H^2 \ \forall n \geq 0.$$

Proof. Exercise □

The forward Euler scheme is:

$$(\frac{u_h^{n+1}-u_h^n}{\Delta t}, v_h) + a(u_h^n, v_h) = (f, v_h) \quad \forall v_h \in V_h$$

with u_h^0 given.

In order to show stability results for this scheme we require an *inverse inequality*:-

$$||v_h||_V \leq S(h)||v_h||_H \quad \forall v_h \in V_h.$$

That such an inequality holds follows from the fact that V_h is finite dimensional and all norms are equivalent in finite dimensions. However such an inequality does not hold for $v \in V$. For V_h being the space of piecewise linear functions it can be shown that $S(h) \leq \frac{c}{h}$.

Lemma 10.2.6. *In the case $f = 0$ for the forward Euler scheme, the following stability bound holds*

$$||u_h^n||_H^2 \leq (1 + K\Delta t)^{-n}||u_0||_H^2 \ \forall n \geq 0$$

provided the stability condition

$$\Delta t \leq \frac{p_m}{\gamma^2 S(h)^2}$$

holds

Proof. Taking $v_h = u_h^{n+1}$ in the forward Euler scheme yields

$$(\frac{u_h^{n+1}-u_h^n}{\Delta t}, u_h^{n+1}) + a(u_h^n, u_h^{n+1}) = 0$$

and rearranging we find

$$\frac{1}{2}(||u_h^{n+1}||_H^2 - ||u_h^n||_H^2 + ||u_h^{n+1}-u_h^n||_H^2) + \Delta t p_m ||u_h^{n+1}||_V^2 \leq \Delta t a(u_h^{n+1}, u_h^{n+1}-u_h^n)$$

and using

$$\begin{aligned}\Delta t a(u_h^{n+1}, u_h^{n+1}-u_h^n) &\leq \gamma \Delta t ||u_h^{n+1}||_V ||u_h^{n+1}-u_h^n||_V \\ &\leq \Delta t \frac{p_m}{2}||u_h^{n+1}||_V^2 + \frac{\Delta t \gamma^2}{2p_m} S(h)^2 ||u_h^{n+1}-u_h^n||_H^2\end{aligned}$$

which yields

$$(1 + K\Delta t)||u_h^{n+1}||_H^2 \leq ||u_h^n||_H^2$$

from which we obtain the desired inequality.

□

This is typical for *explicit* schemes for evolutionary PDEs in which in order to have stability the time step must satisfy a condition which restricts its size in terms of the spatial mesh size. In this case

$$\Delta t \leq Ch^2$$

for stability.

Chapter 11

Variational Inequalities

11.1 Projection theorem

Theorem 11.1.1. *Let K be a closed convex subset of a Hilbert space H. It follows that*

- *For all $x \in H$ there exists a unique $y \in K$ such that*

  $$||y - x|| = \inf_{\eta \in K} ||\eta - x||_H.$$

  We set

  $$y := \mathbb{P}_K x$$

  and call $\mathbb{P}_K : H \to K$ the projection operator from H onto K.

- $$y = \mathbb{P}_K x \iff y \in K \quad \text{and} \quad \langle y, \eta - y \rangle_H \geq \langle x, \eta - y \rangle_H \ \forall \eta \in K.$$

- *The operator $\mathbb{P}$ is non-expansive:-*

  $$||\mathbb{P}_K x_1 - \mathbb{P}_K x_2||_H \leq ||x_1 - x_2||_H.$$

11.2 Elliptic variational inequality

Let K be a closed convex subset of a Hilbert space V. Let $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ be a bounded coercive bilinear form satisfying

1) $a(\cdot,\cdot)$ is bounded, i.e.,

$$\exists \gamma > 0 \text{ s.t. } |a(v,w)| \leq \gamma \|v\|_V \|w\|_V \ \forall v, w \in V.$$

2) $a(\cdot,\cdot)$ is coercive i.e.,

$$\exists \alpha > 0 \text{ s.t. } a(v,v) \geq \alpha \|v\|_V^2 \ \forall v \in V.$$

and $l(\cdot) : V \to \mathbb{R}$ be a bounded linear functional, i.e.,

$$\exists c_l > 0 \text{ s.t. } |l(v)| \leq c_l \|v\|_V \ \forall v \in V.$$

Theorem 11.2.1. *There exists a unique $u \in K$ such that*

$$\textit{(VI)} \quad a(u, v-u) \geq l(v-u) \quad \forall \quad v \in K.$$

Furthermore

$$||u_1 - u_2||_V \leq \frac{1}{\alpha} ||l_1 - l_2||_{V^*}$$

$u_i, i = 1, 2$ solves (VI) for the linear forms l_1 and l_2, respectively.

11.3 Obstacle problem

Let $V := H_0^1(\Omega)$ where Ω is a bounded domain in $\mathbb{R}^d, d = 1, 2, 3$. Set

$$a(w,v) := \int_\Omega \nabla w \cdot \nabla v \quad , \quad l(v) := \int_\Omega fv$$

where $f \in L^2(\Omega)$ is given. Let $\psi \in H^1(\Omega) \cup C^0(\bar{\Omega})$ and $\psi|_{\partial\Omega} \leq 0$ and set

$$K := \{v \in H_0^1(\Omega) : v \geq \psi a.e. \in \Omega\}.$$

It follows that

1. K is non-empty.

   Set $\psi^+ := \frac{1}{2}(\psi + |\psi|) = \max(\psi,)$. Recall the following lemma

   Lemma 11.3.1. *Let $\theta : \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous, i.e.,*

   $$|\theta(t_1) - \theta(t_2)| \leq \lambda_\theta |t_1 - t_2| \quad \forall \quad t_1, t_2 \in \mathbb{R}.$$

   Suppose θ has a finite number of points of discontinuity. Then $\theta : H^1(\Omega) \to H^1(\Omega)$ is continuous and $\theta : H_0^1(\Omega) \to H_0^1(\Omega)$ is continuous in the case $\theta(0) = 0$.

Hence $\psi^+ \in H_0^1(\Omega)$ and since $\psi^+ \geq \psi$ it follows that K is non-empty.

2. K is convex because for $t \in [0,1]$, $t\eta + (1-t)v \geq \psi \forall \eta, v \in K$.

3. K is closed.

   This follows from the fact that convergence in $H_0^1(\Omega)$ implies convergence in $L^2(\Omega)$ and hence convergence almost everywhere for a sub-sequence. From which we find that $v_n \to v$ in V implies $v_{n_i} \to v$ *a.e.* in *in* Ω and if $v_n \in K$ then $v_{n_i} \geq \psi$ *a.e. in* Ω which implies by the convergence of v_n that $v \geq \psi$ *a.e. in* Ω and so $v \in K$.

Theorem 11.3.2. *There exists a unique solution to the obstacle problem: Find* $u \in K := \{v \in H_0^1(\Omega) : v \geq \psi$ *a.e. in* $\Omega\}$ *such that*

$$\int_\Omega \nabla u \cdot (\nabla v - \nabla u) \geq \int_\Omega f(v-u) \quad \forall v \in K.$$

11.3.1 Finite element approximation

Part III

Finite Differences

Chapter 12

Introduction to Finite Difference Methods

12.1 Finite Differences

The basic idea of finite difference methods is to seek approximations to solutions of the PDE on a lattice. To approximate the derivatives appearing in the PDE, differences between lattice values at neighbouring points are used. We introduce the idea by considering functions of a single variable x.

Let $x_j = j\Delta x$, $\Delta x \ll 1$ and consider a smooth function $u : I \to \mathbb{R}$ for some open $I \subset \mathbb{R}$. We set $u_j = u(x_j)$. By Taylor expansion we have:

$$u_{j\pm1} = u_j \pm \Delta x \frac{\partial u}{\partial x}(x_j) + \frac{\Delta x^2}{2}\frac{\partial^2 u}{\partial x^2}(x_j) \pm \frac{\Delta x^3}{6}\frac{\partial^3 u}{\partial x^3}(x_j) + \mathcal{O}(\Delta x^4) \quad (12.1)$$

provided that $u \in C^4(I, \mathbb{R})$, $x_j \in I$ and Δx is sufficiently small. [1] From this we see that

$$\frac{\partial^2 u}{\partial x^2}(x_j) = \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x^2} + \mathcal{O}(\Delta x^2) \quad (12.2)$$

$$:= \frac{\delta^2 u_j}{\Delta x^2} + \mathcal{O}(\Delta x^2) \quad (12.3)$$

provided that $u \in C^4(I, \mathbb{R})$, $x_j \in I$ and Δx is sufficiently small.

[1] We write all derivatives as partial here because, in subsequent applications of these ideas, u will typically be a function of several variables.

We can stop the Taylor expansion at different powers of Δx and obtain similar approximations for the first derivatives. For example

$$\frac{\partial u}{\partial x}(x_j) = \frac{u_{j+1}-u_j}{\Delta x} + \mathcal{O}(\Delta x), \tag{12.4}$$

$$:= \frac{\Delta_+ u_j}{\Delta x} + \mathcal{O}(\Delta x) \tag{12.5}$$

and

$$\frac{\partial u}{\partial x}(x_j) = \frac{u_j-u_{j-1}}{\Delta x} + \mathcal{O}(\Delta x), \tag{12.6}$$

$$:= \frac{\Delta_- u_j}{\Delta x} + \mathcal{O}(\Delta x) \tag{12.7}$$

provided that $u \in C^2(I,\mathbb{R})$, $x_j \in I$ and Δx is sufficiently small. With the assumption that u is three times continuously differentiable we can find an improved approximation to the first derivative:

$$\frac{\partial u}{\partial x}(x_j) = \frac{u_{j+1}-u_{j-1}}{2\Delta x} + \mathcal{O}(\Delta x^2), \tag{12.8}$$

$$:= \frac{\Delta_0 u_j}{2\Delta x} + \mathcal{O}(\Delta x^2) \tag{12.9}$$

provided that $u \in C^3(I,\mathbb{R})$, $x_j \in I$ and Δx is sufficiently small.

12.2 Time-stepping

We illustrate the idea of finite difference methods through time-stepping methods for ODEs. Consider the equation

$$\frac{du}{dt} = f(u)$$

and let $U^n \approx u(n\Delta t)$ denote an approximation. Such an approximation can be computed by the following methods:

- $\frac{U^{n+1}-U^n}{\Delta t} = f(U^n)$ – Forward Euler;
- $\frac{U^{n+1}-U^n}{\Delta t} = f(U^{n+1})$ – Backward Euler;
- $\frac{U^{n+1}-U^n}{\Delta t} = \theta f(U^{n+1}) + (1-\theta) f(U^n)$ – θ-method;
- $\frac{U^{n+1}-U^n}{\Delta t} = f(\theta U^{n+1} + (1-\theta)U^n)$ – one-leg θ-method;
- $\frac{U^{n+1}-U^{n-1}}{2\Delta t} = f(U^n)$ – Leap-frog method.

12.3 Norms

Consider a function $u : \Omega \to \mathbb{R}$ with $\Omega \subset \mathbb{R}^d$. When we discretize in space we will obtain lattice approximations U_k where k is a multi-index ranging over a lattice Ω_Δ. We use U to denote the vector obtained from this indexed set of U_k. We use Δ to denote the mesh-spacing. For simplicity we will always use the same mesh-spacing in all dimensions.

When considering maximum principles our topology will be the supremum norm in space and we use the notation

$$\|u\|_\infty = \sup_{x\in\Omega} |u(x))|, \tag{12.10}$$

with Ω differing from example to example.

For the discretization we use the notation

$$\|U\|_\infty = \max_{k\in\Omega_\Delta} |U_k|, \tag{12.11}$$

with Ω_Δ differing from example to example. No confusion should arise from the dual use of the notation $\|\cdot\|_\infty$ since it will always be clear from the context whether a function or a vector is being measured.

Chapter 13

Finite difference schemes for the diffusion equation

13.1 Introduction

In this chapter we study the Diffusion Problem (1.6) by means of finite difference approximations.

13.2 The Heat Equation

13.2.1 The PDE

Here we study the Diffusion Model Problem (1.6) in dimension $d = 1$, with $\Omega = (0, 1)$ and with $f = 0$. Let $g \in C([0, 1], \mathbb{R})$ and consider the problem:

$$\begin{array}{rcll} \frac{\partial u}{\partial t} & = & \frac{\partial^2 u}{\partial x^2} & (x,t) \in (0,1) \times (0,\infty), \\ u & = & 0 & (x,t) \in \{0,1\} \times (0,\infty), \\ u & = & g & (x,t) \in [0,1] \times \{0\}. \end{array} \tag{13.1}$$

Theorem 13.2.1. *Let $t > s > 0$. Then:*

$$\min_{0 \le y \le 1} u(y,s) \le u(x,t) \le \max_{0 \le y \le 1} u(y,s) \quad \forall x \in [0,1] \tag{13.2}$$

This well-posedness result can be used to establish uniqueness and continuous dependence of the solution on the initial data.

13.2.2 The Approximation

Again, letting $x_j = jh$ and $Jh = 1$, $J \in \mathbb{N}$, we introduce $U_j(t)$ as our approximation to $u(x_j, t)$ and consider the approximation

$$\begin{array}{rcll} \frac{dU_j}{dt} & = & \frac{1}{h^2}\delta^2 U_j & (j,t) \in \{1,\ldots,J-1\} \times [0,T], \\ U_j & = & 0, & (j,t) \in \{0,J\} \times [0,T], \\ U_j & = & g(x_j) & (j,t) \in \{0,1,....,J\} \times \{0\}. \end{array}$$

Adopting vector notation ($U = (U_1, \ldots, U_{J-1})^T$), we have:

$$\begin{array}{rcl} \dfrac{dU}{dt} + AU & = & 0, \\ U(0) & = & G \end{array}$$

where

$$\begin{array}{rcl} A & = & \dfrac{-1}{h^2} \begin{pmatrix} -2 & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & -2 \end{pmatrix}, \\ G & = & (g(x_1), \ldots, g(x_{J-1}))^T. \end{array}$$

Letting $t_n = n\Delta t$ and $N\Delta t = T$, $N \in \mathbb{N}$, we now apply the theta method (with $\theta \in [0,1]$) in time to get:

$$\begin{array}{rcll} \frac{U_j^{n+1} - U_j^n}{\Delta t} & = & \frac{\theta}{h^2}\delta^2 U_j^{n+1} + \frac{1-\theta}{h^2}\delta^2 U_j^n & (j,n) \in \{1,\ldots,J-1\} \times \{0,\ldots,N-1\} \\ U_j^n & = & 0 & (j,n) \in \{0,J\} \times \{1,\ldots,N\} \\ U_j^n & = & g(x_j) & (j,n) \in \{0,\ldots,J\} \times \{0\} \end{array} \tag{13.3}$$

Set

$$\lambda = \frac{\Delta t}{h^2}.$$

- *Explicit scheme* $\theta = 0$

$$U_j^{n+1} = \lambda U_{j-1}^n + (1-2\lambda)U_j^n + \lambda U_{j+1}^n$$

- *Fully implicit scheme* $\theta = 1$

$$-\lambda U_{j-1}^{n+1} + (1+2\lambda)U_j^{n+1} - \lambda U_{j+1}^{n+1} = U_j^n$$

- *Crank-Nicolson scheme* $\theta = 1/2$

Vectorially, we have, for $U^n = (U_1^n, \ldots, U_{J-1}^n)^T$

$$\begin{array}{rcl} (I + \Delta t\theta A)U^{n+1} & = & (I - \Delta t(1-\theta)A)U^n \\ U^0 & = & G \end{array} \tag{13.4}$$

Theorem 13.2.2. Discrete well-posedness *If* $\lambda(1-\theta) \leq \frac{1}{2}$ *then the solution at time level* $n+1$ *is unique and for* $j \in \{0, \ldots, J\}$,

$$\min_{k\in\{0,\ldots,J\}} U_k^n \leq U_j^{n+1} \leq \max_{k\in\{0,\ldots,J\}} U_k^n.$$

Proof. Because the schemes are linear uniqueness follows from the bounds and the fact that for the difference of two solutions the given data is zero.

We have, in the interior $(j,n) \in \{1, \ldots, J-1\} \times \{0, \ldots, N-1\}$,

$$(1+2\lambda\theta)U_j^{n+1} = \lambda\theta(U_{j-1}^{n+1} + U_{j+1}^{n+1}) + \lambda(1-\theta)(U_{j-1}^n + U_{j+1}^n) + [1 - 2\lambda(1-\theta)]U_j^n$$

Thus, define

$$U_{\max}^n = \max_{j\in\{0,\ldots,J\}} U_j^n.$$

The upper inequality simply states that

$$U_{\max}^{n+1} \leq U_{\max}^n$$

and it is this that we now prove.

Note that $1 - 2\lambda(1-\theta) \geq 0$, $(1-\theta) \geq 0$, $\theta \geq 0$ and $\lambda \geq 0$. For (j,n) in the interior,

$$\begin{array}{rcl} (1+2\lambda\theta)U_j^{n+1} & \leq & 2\lambda\theta U_{\max}^{n+1} + 2\lambda(1-\theta)U_{\max}^n + [1-2\lambda(1-\theta)]U_{\max}^n \\ & = & 2\lambda\theta U_{\max}^{n+1} + U_{\max}^n \end{array}$$

If $U^{n+1}_{\max}$ occurs for $j \in \{1, \ldots, J-1\}$ then

$$\begin{aligned}(1+2\lambda\theta)U^{n+1}_{\max} &\leq 2\lambda\theta U_{\max} + U^n_{\max} \\ \implies U^{n+1}_{\max} &\leq U^n_{\max}\end{aligned}$$

If $U^{n+1}_{\max}$ occurs for $j \in \{0, J\}$ then

$$U^{n+1}_{\max} = 0 \leq U^n_{\max}$$

Hence

$$U^{n+1}_{\max} \leq U^n_{\max} \quad n \in \{0, \ldots, N-1\}$$

and the upper inequality follows. The lower inequality is proved similarly, by considering

$$U^n_{\min} = \min_{j \in \{0,\ldots,J\}} U^n_j$$

□

13.2.3 Convergence

Let $u^n_j = u(x_j, t_n)$, noting that U^n_j is the computed approximation to u^n_j. Let

$$T^n_j = \frac{u^{n+1}_j - u^n_j}{\Delta t} - \frac{\theta}{h^2}\delta^2 u^{n+1}_j - \frac{(1-\theta)}{h^2}\delta^2 u^n_j \quad (j,n) \in \{1, \ldots, J-1\} \times \{0, \ldots, N-1\}$$

This is the truncation error. We define $T^n = (T^n_1, \ldots, T^n_{J-1}) \in \mathbb{R}^{J-1}$. By the techniques in Chapter 12, expanding about $\frac{t_{n+1}+t_n}{2}$, the following may be shown, using the discrete infinity norm.

Lemma 13.2.3. Consistency *If* $u(x,t) \in C^{4\times 2}([0,1] \times [0,T], \mathbb{R})$ *then*

$$\max_{n \in \{0,\ldots,(N-1)\}} \|T^n\|_\infty \leq C[\Delta t + h^2]$$

for some constant C *independent of* Δt *and* h. *Also, if* $\theta = \frac{1}{2}$ *and* $u(x,t) \in C^{4\times 3}([0,1] \times [0,T], \mathbb{R})$ *then*

$$\max_{n \in \{0,\ldots,(N-1)\}} \|T^n\|_\infty \leq C[\Delta t^2 + h^2]$$

for some constant C *independent of* Δt *and* h.

Assumption For simplicity we assume that

$$\max_{n\in\{0,\ldots,(N-1)\}} \|T^n\|_\infty \leq C[\Delta t^2 + (1-2\theta)\Delta t + h^2]$$

in the following. The Lemma 13.2.3 gives conditions under which this holds.

Remark

We prove convergence directly from this consistency result. Of course we are implicitly proving a stability result in the course of the proof.

Theorem 13.2.4. *Let $\lambda(1-\theta) \leq \frac{1}{2}$ and make the above assumption. Then*

$$\max_{n\in\{0,\ldots,N\}} \|u^n - U^n\|_\infty \leq CT\left[(1-2\theta)\Delta t + \Delta t^2 + h^2\right].$$

Proof. Let $e_j^n = u_j^n - U_j^n$. Then, by linearity,

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} = \frac{\theta}{h^2}\delta^2 e_j^{n+1} + \frac{1-\theta}{h^2}\delta^2 e_j^n + T_j^n \quad (j,n) \in \{1,\ldots,J-1\}\times\{0,\ldots,N-1\}$$

Thus, for $(j,n) \in \{1,\ldots,J-1\}\times\{0,\ldots,N-1\}$ we get:

$$(1+2\lambda\theta)e_j^{n+1} = \lambda\theta(e_{j-1}^{n+1} + e_{j+1}^{n+1}) + \lambda(1-\theta)(e_{j-1}^n + e_{j+1}^n) + [1-2\lambda(1-\theta)]e_j^n + \Delta t T_j^n$$

with

$$e_0^n = e_J^n = 0,$$
$$e_j^0 = 0,\ j = 1,\ldots,J-1.$$

We define

$$E^n = (e_1^n,\ldots,e_{J-1}^n)^T,\ T^n = (T_1^n,\ldots,T_{J-1}^n)^T.$$

Since $1-2\lambda(1-\theta) \geq 0$, $(1-\theta) \geq 0$, $\theta \geq 0$ and $\lambda \geq 0$ we have

$$(1+2\lambda\theta)|e_j^{n+1}| \leq 2\lambda\theta\|E^{n+1}\|_\infty + 2\lambda(1-\theta)\|E^n\|_\infty + [1-2\lambda(1-\theta)]\|E^n\|_\infty + \Delta t\|T^n\|_\infty$$

Hence, for $j \in \{1,\ldots,J-1\}$

$$(1+2\lambda\theta)|e_j^{n+1}| \leq 2\lambda\theta\|E^{n+1}\|_\infty + \|E^n\|_\infty + \Delta t\|T^n\|_\infty$$

which implies

$$\begin{aligned}(1+2\lambda\theta)\|E^{n+1}\|_\infty &\leq 2\lambda\theta\|E^{n+1}\|_\infty + \|E^n\|_\infty + \Delta t\|T^n\|_\infty \\ \Longrightarrow \|E^{n+1}\|_\infty &\leq \|E^n\|_\infty + \Delta t\|T^n\|_\infty.\end{aligned}$$

By induction, using $\|E^0\|_\infty = 0$ and $0 \leq n\Delta t \leq T$, setting $\tau := \max_{n\in\{0,\ldots,(N-1)\}} \|T^n\|_\infty$ we obtain

$$\|E^n\|_\infty \leq n\Delta t\tau \leq T\tau$$

as required. □

13.3 Fourier analysis

13.3.1 Example

Consider the following finite difference approximation of the diffusion equation

Explicit scheme $\theta = 0$

$$U_j^{n+1} = \lambda U_{j-1}^n + (1-2\lambda)U_j^n + \lambda U_{j+1}^n,\ j = 1, 2..., J-1$$

where

$$U_j^0 = \alpha_k \sin(k\pi x_j), j = 1, 2, .., J-1\ ,\ U_0^n = U_J^n = 0,\ n = 0, 1, 2, ...$$

and we take J is even.

We seek a solution in the form

$$U_j^n = \mu_k^n \alpha_k \sin(k\pi x_j), j = 1, 2, .., J-1, \quad n \geq 0.$$

Substitution into the scheme yields

$$\mu_k^{n+1} \sin(k\pi jh) = \mu_k^n(\lambda(\sin(k\pi(j-1)h) + \sin(k\pi(j+1)h)) + (1-2\lambda)\sin(k\pi jh))$$

and applying the appropriate trigonometric addition formula yields

$$\mu_k^{n+1} \sin(k\pi jh) = \mu_k^n(\lambda 2 \sin(k\pi jh)\cos(k\pi h) + (1-2\lambda)\sin(k\pi jh))$$

so that

$$\mu_k^{n+1} \sin(k\pi jh) = \mu_k^n(1 - 2\lambda(1-\cos(k\pi h))\sin(k\pi jh)$$

and

$$\mu_k^{n+1} = \mu_k^n(1 - 2\lambda(1-\cos(k\pi h)) = (1 - 4\lambda \sin^2(\frac{k\pi h}{2}))\mu_k^n.$$

Thus

$$\mu_k^n = (\mu_k)^n = \exp(-\omega_k \Delta t)^n = \exp(-\omega_k t_n)\alpha_k$$

where

$$\mu_k := \exp(-\omega_k \Delta t) = (1 - 4\lambda \sin^2(\frac{k\pi h}{2})).$$

It follows that the solution of the finite difference scheme may be written as

$$U_j^n = \alpha_k \exp(-\omega_k t_n) \sin(k\pi x_j), j = 1, 2, .., J-1, \quad n \geq 0.$$

This may be compared with the general solution of the diffusion equation

$$u(x,t) = \sum_{k=1}^{\infty} \alpha_k \exp(-k^2 t) \sin(k\pi x)$$

with the initial data

$$u(x,t) = \sum_{k=1}^{\infty} \alpha_k \sin(k\pi x).$$

We see that in this general solution the Fourier coefficients decay in time so that the Fourier series converges if the series converges for the initial data. It is clearly desirable for the numerical solution to have non-growing Fourier coefficients. In order for the discrete solution to converge as Δt and h converge to zero with λ fixed it is necessary for the discrete solution to be bounded at any fixed time.

Observe two situations:

- **Stability**

  $$0 < \lambda \leq \frac{1}{2}$$

  It follows that for all $\Delta t, h$ and k that

  $$|\mu_k| \leq 1$$

  since

  $$1 \geq 1 - 4\lambda \sin^2(\frac{k\pi h}{2}) \geq 1 - 2\sin^2(\frac{k\pi h}{2}) \geq -1.$$

  In this case we find the discrete coefficient

  $$\alpha_k \exp(-\omega_k t_n) = \alpha_k \mu_k^n$$

  does not grow as $n \to \infty$.

- **Instability**

  $$\lambda > \frac{1}{2}$$

It follows that for fixed $\lambda > \frac{1}{2}$ and h sufficiently small there exists $k = k^*$ such that

$$|\mu_{k^*}| > 1.$$

To show this, write $\lambda = \frac{1}{2} + \frac{\delta}{4}$ for some positive δ, and choosing $k^* := J - 1$ we find that, since $\sin^2(\frac{\pi(1-h)}{2}) > 1 - ch^2$ for h sufficiently small,

$$\mu_{k^*} = 1 - 4\lambda \sin^2(\frac{\pi(1-h)}{2}) < 1 - 4\lambda(1 - ch^2) = -1 - \delta + 4\lambda ch^2 < -1$$

for h sufficiently small.

Thus for initial data with $k = k^*$ and $\alpha_{k^*} \neq 0$ consider U_j^n for $j = J/2$ when J is even and k^* is odd so that

$$|\sin(k^*\pi jh)| = |\sin(k^*\pi/2)| = 1$$

we find that

$$|U_j^n| = |\alpha_{k^*}||\mu_{k^*}|^n \to \infty \text{ as } n \to \infty$$

for fixed λ.

This is an example of instability. Choosing to reduce h and Δt to zero whilst maintaining $\lambda > \frac{1}{2}$ will lead to divergence of the numerical solution in general.

13.3.2 Initial value problems with periodic boundary conditions

We place ourselves, for convenience, in the setting of periodic complex mesh functions $\mathcal{M}_h := \{V := \{V_j\}_{j\in\mathbb{Z}}$ such that $V_{j+J} = V_j\}$ where $Jh = 2\pi$ and $x_j := jh$. The discrete L_h^2 inner product is

$$(V, W)_h := h\sum_{j=1}^{J} V_j \bar{W}_j, \ ||V||_h := (V, V)_h^{\frac{1}{2}}.$$

We use $i := \sqrt{-1}$. Set $\Phi^k \in \mathcal{M}_h$ by

$$\Phi_j^k := \exp(ikjh).$$

It follows that

$$(\Phi^k, \Phi^m)_h = 2\pi\delta_{km}, \ k, m = 1, 2, ...J$$

Thus $\{\Phi^k\}_{k=1}^J$ form an orthogonal basis of $\mathcal{M}_h$. Any mesh function may be written in as

$$V := \sum_{k=1}^{J} \hat{V}_k \Phi^k \text{ or } V_j := \sum_{k=1}^{J} \hat{V}_k \Phi_j^k$$

where the $\hat{V}_k$ are unique.

Consider the constant coefficient finite difference formula

$$\sum_p \alpha_p U_{j+p}^{n+1} = \sum_p \beta_p U_{j+p}^n, \ \forall j \tag{13.5}$$

where U^{n+1} and U^n are periodic mesh functions. It follows that

$$\sum_p \alpha_p \sum_{k=1}^{J} \hat{U}_k^{n+1} \Phi_{j+p}^k = \sum_p \beta_p \sum_{k=1}^{J} \hat{U}_k^n \Phi_{j+p}^k$$

and rearranging

$$\sum_{k=1}^{J} \hat{U}_k^{n+1} \exp(ikjh) \sum_p \alpha_p \exp(ikph) = \sum_{k=1}^{J} \hat{U}_k^n \exp(ikjh) \sum_p \beta_p \exp(ikph)$$

which implies that

$$\hat{U}_k^{n+1} = \mu_k \hat{U}_k^n$$

where

$$\mu_k := \frac{\sum_p \beta_p \exp(ikph)}{\sum_p \alpha_p \exp(ikph)}$$

so that

$$\hat{U}_k^{n+1} = \mu_k^n \hat{U}_k^0.$$

The solution of the initial value problem may be written as

$$U^n = \sum_{k=1}^{J} \exp(-\omega_k t_n) \hat{U}_k^0 \Phi^k \tag{13.6}$$

$$U_j^n = \sum_{k=1}^{J} \hat{U}_k^0 \exp(-\omega_k t_n) \exp(ikx_j) \tag{13.7}$$

where

$$\exp(-\omega_k \Delta t) := \mu_k.$$

This may be contrasted with the solution

$$u(x,t) = \sum_{k=1}^{\infty} \hat{u}_k(0) \exp(-\omega_k t) \exp(ikx) \tag{13.8}$$

of the initial value problem

$$u_t = Lu, \quad u(x,0) = \sum_{k=1}^{\infty} \hat{u}_k(0) \exp(ikx) \tag{13.9}$$

$$Lu := P(\partial_x)u = \sum_{r=1}^{m} A_r \partial_x^r u, \quad P(z) := \sum_{r=1}^{m} A_m z^r \tag{13.10}$$

where

$$\omega_k = -P(ik). \tag{13.11}$$

Example

In the case of the heat equation

$$P(r) = r^2, \quad w_k = k^2$$

and

$$u(x,t) = \sum_{k=1}^{\infty} \hat{u}_k(0) \exp(-k^2 t) \exp(ikx).$$

On the other hand the finite difference scheme

$$(1 + 2\lambda\theta)U_j^{n+1} - \lambda\theta(U_{j-1}^{n+1} + U_{j+1}^{n+1}) = \lambda(1-\theta)(U_{j-1}^n + U_{j+1}^n) + [1 - 2\lambda(1-\theta)]U_j^n$$

has

$$\alpha_1 = \alpha_{-1} = -\lambda\theta, \ \alpha_0 = 1 + 2\lambda\theta, \quad \beta_1 = \beta_{-1} = \lambda(1-\theta), \ \beta_0 = 1 - 2\lambda(1-\theta)$$

so that

$$\mu_k = \frac{1 - 2\lambda(1-\theta) + \lambda(1-\theta)(\exp(ikh) + \exp(-ikh))}{1 + 2\lambda\theta - \lambda\theta(\exp(ikh) + \exp(-ikh))}$$

yielding

$$\mu_k = \frac{1 - 4\lambda(1-\theta)\sin^2(\frac{kh}{2})}{1 + 4\lambda\theta\sin^2(\frac{kh}{2})}.$$

Clearly $\mu_k < 1$. If $\theta \geq 1/2$ or if

$$\theta < \frac{1}{2}, \quad \lambda \leq \frac{1}{2(1-2\theta)}$$

then $|\mu_k| \leq 1$ and the scheme is said to be stable.

On the other hand if

$$\theta < \frac{1}{2}, \quad \lambda > \frac{1}{2(1-2\theta)}$$

then k can be chosen such that $|\mu_k| > 1$ and the scheme is unstable and does not converge.