Lecture Notes

On

FINITE ELEMENT METHODS FOR ELLIPTIC PROBLEMS

[1]

Amiya Kumar Pani

Industrial Mathematics Group
Department of Mathematics
Indian Institute of Technology, Bombay
Powai, Mumbai-4000 76 (India).

IIT Bombay, March 2012 .

[1]Workshop on 'Mathematical Foundation of Advanced Finite Element Methods (MFAFEM-2013) held in BITS,GOA during 26th December - 3rd January, 2014

‘ Everything should be made simple, but not simpler’.

- Albert Einstein.

Contents

HOW THESE NOTES CAME TO BE

I have never intended to write a *Lecture Note* on this topic, as there are many excellent books available in literature. But, when a group of enthusiastic faculty members in the Department of Mathematics at the Universidade Federale do Paraná (UFPR) requested me to give some lectures on the theory of finite elements, I immediately rushed to the Central Library of UFPR to check some books on this topic. To my dismay (may be it is a boon in disguise), I could not find any relevant books in the library here. More over, in my first lecture, I realised the urgent need of supplying some notes in order to keep their interest on. However, thank to Professor Vidar Thomée, I had his notes entitled *Lectures on Approximation of Parabolic Problems by Finite Elements* [13] and thank to the netscape for downloading [7] from the home page of Professor D. Estep from CALTEC. So in an attempt to explain the lectures given by Professor Vidar at IIT, Bombay , first four chapters are written. Thank to LATEX and Adriana, she typed first chapter promptly and then the flow was on. So by the end of the series, we had these notes. At this point let me point out that the write up is influenced by the style of Mercier [10] (although the book is not available here, but it is more due to my experience of using this book for a course at IIT, Bombay).

Sometimes back, I asked myself: Is there any novelty in the present approach? Well, (Bem!) the answer is certainly in negative. But for the following simple reason this one may differ from those books written by stalwarts of this field : Some questions are posed in the introductory chapter and throughout this series, attempts have been made to provide some answers. Moreover for finite element methods, the usefulness of Lax equivalence theorem has been highlighted throughout these notes and the role as well as the appropriate choice of intermediate or auxiliary projection has been brought out in the context of finite element approximations to parabolic problems.

It seems to be a standard practice that in the Preface the author should give a brief description of each chapter of his / her book. In order to pay respect to our tradition, I (although first person is hardly used in writing articles) am making an effort to present some rudiments of each chapter of this lecture note.

In the introductory chapter, some basic (!) questions are raised and in the subsequent chapters, some answers have been provided. To motivate the finite

element methods, a quick description on the finite difference method to the Poisson's equation is also given.

Chapter 2 is devoted to a quick review of the theory of distribution, generalised derivatives and some basic elements of Sobolev spaces. Concept of trace and Poincaré inequality are also discussed in this chapter.

In chapter 3, a brief outline of abstract variational theory is presented . A proof of Lax-Milgram Theorem is also included. Several examples on elliptic problems are discussed and the question of solvability of their weak formulations are examined.

Chapter 4 begins with a brief note on the finite dimensional approximation to the abstract variational problems. An introduction of finite element methods applied to two elliptic problems is also included. *A priori* error estimates are also derived. Based on Eriksson *et al.* [7], an adaptive procedure is outlined with *a posteriori* error estimates.

I am indebted to Professors Bob Anderssen, Purna C. Das, Ian Sloan, Graeme Fairweather, Vidar Thomée and Lars Wahlbin for introducing me the beautiful facets of Numerical Analysis. The present set of notes is a part of the lectures given in the Department of Mathematics at Federal University of Parana, Curitiba. This would have not been possible without the support of Professor Jin Yun Yuan. He helped me in typing and correcting the manuscript. He is a good human being and a good host and I shall recommend the readers to visit him. I greately acknowledge the financial support provided by the UFPR. Thank to the GORORORBA group for entertaining me during lunch time and I also acknowledge the moral support provided by the evening café club members (Carlos, Corréa Eidem, Fernando (he is known for his improved GORORORBA), Jose, Zeca). Perhapse, it would be incomplete if I do not acknowledge the support and encouragement given by my wife Tapaswini. She allowed me to come here in a crucial juncture of her life. Thanks to the email facility through which I get emails from my son Aurosmit and come to know more about my new born baby ' Anupam' (I am waiting eagerly to see him, he was born when I was away to Curitiba). I would like to express my heart felt thank to the family of Professor Jin Yun and Professor Beatriz for showing me some nice places in and around Curitiba. Finally, I am greately indebted to my friend Professor Kannan Moudgalya and his family for their help and moral support provided to my family during my abscence.

Amiya Kumar Pani

Chapter 1

Introduction

In this introductory chapter, we briefly describe various issues related to finite element methods for solving differential equations.

1.1 Background.

We shall now start with a mathematical model which describes the vibration of a drum. Let $\Omega \subset \mathbb{R}^2$ be a thin membrane fixed to the brim of a hallow wooden structure like:

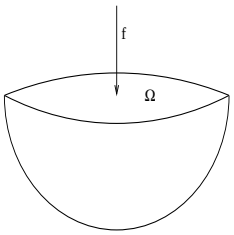
Given an external force f (as show in the figure) we are interested to find the displacement u at any point (x, y) in the domain Ω . The above model gives rise to the following partial differential equation:

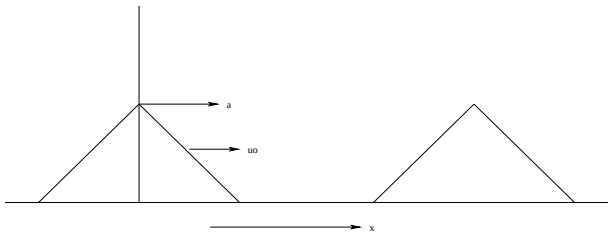
Find the displacement $u(x, y)$ satisfying

$$-\Delta u(x, y) = f(x, y), (x, y) \in \Omega, \tag{1.1}$$

$$u(x, y) = 0, (x, y) \in \partial\Omega,$$

where $\Delta = \dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2}$ and $\partial\Omega$ is the boundary of the domain Ω. Since the boundary $\partial\Omega$ is fixed, there is no displacement and hence, $u(x, y) = 0$ on the boundary $\partial\Omega$.





By a solution (classical) u of (1.1), we mean a twice continuously differentiable function u which satisfies the partial differential equation (1.1) at each point (x, y) in Ω and also the boundary condition. Note that f has to be continuous, if we are looking for the classical solutions of (1.1). For the present problem, the external force f may not be imparted continuously. In fact f can be a point force, i.e., a force applied at a point in the domain. Thus a physically more relevant problem is to allow more general f, that is, f may have some discountinuities. Therefore, there is a need to generalize the concept of solutions. In early 1930, S.L. Sobolev came across with a similar situation while he was dealing with the following first order hyperbolic equation:

$$\begin{aligned} u_t + a u_x &= 0,\ t > 0, \quad -\infty < x < \infty \\ u(x, 0) &= u_0(x), \quad -\infty < x < \infty. \end{aligned} \tag{1.2}$$

Here, a is a real, positive constant and u_0 is the initial profile. It is well know the exact solution $u(x, t) = u_0(x - at)$. In this case, the solution preserves the shape of the initial profile. However, if u_0 is not differentiable (say it has a kink) the solution u is still
meaningful physically, but not in the sense of classical solution. These observations were instrumental for the development of the modern partial differential equations.

Again comming back to our vibration of **drum** problem, we note that in mechanics or in physics, the same model is described through a minimization of total energy, say $J(v)$, i.e., minimization of $J(v)$ subjects to the set V of all possible admissible displacements, where

$$J(v) = \underbrace{\frac{1}{2}\int_\Omega (|\frac{\partial v}{\partial x}|^2 + |\frac{\partial v}{\partial y}|^2)\, dxdy}_{Kinetic\ Energy\ mass\ being\ unit} - \underbrace{\int_\Omega f v\, dxdy}_{Potential\ Energy} \tag{1.3}$$

and V is a set of all possible displacements v such that the above integrals are meaningful and $v = 0$ on $\partial\Omega$. More precisely, we cast the above problem as Find $u \in V$ such that $u = 0$ on $\partial\Omega$ and u minimize J, that is,

$$J(u) = \underset{v \in V}{Min}\, J(v). \tag{1.4}$$

The advantage of the second formulation is that the displacement u may be once continuously differentiable and the external force f may be of general form, i.e., f may be square integrable and may allow discontinuties. Further, it is observed that every solution u of (1.1) satisfies (1.4) (it is really not quite aparrent, but it is indeed possible, this point we shall closely examine later on in the course of my lectures). However, the converse need not be true, since in (1.4) the solution u is only once continuosly differentiable.

We shall see subsequently that (1.4) is a weak formulation of (1.1), and it allows physically more relevant external forces, say even point force.

Choice of the admissible space V. At this stage, it is worthwile to analyse the space V, which will motivate the introduction of Sobolev space in Chpater 2. With $f \in L^2(\Omega)$ (Space of all square integrable functions), V may be considered as:

$$\{v \in C^1(\Omega) \cap C(\overline{\Omega}) : \int_\Omega |v|^2 dxdy < \infty, \int_\Omega (|\frac{\partial v}{\partial x}|^2 + |\frac{\partial v}{\partial y}|^2)\, dxdy < \infty \ \textit{and}\ v = 0 \ \textit{on}\ \partial\Omega\},$$

where $C^1(\Omega)$, the space of one time continuously differentiable functions in Ω is such that $C^1(\Omega) = \{v : v, v_x, v_y \in C(\Omega)\}$ and $C(\overline{\Omega})$ is the space of continuous functions defined on $\overline{\Omega}$ with $\overline{\Omega} = \Omega \cup \partial\Omega$. Hence, $u_{|\partial\Omega}$ is properly defined for $u \in C(\overline{\Omega})$. Unfortunately, V is not complete with measurement (norm) given by

$$\|u\|_1 = \sqrt{\int_\Omega (|u|^2 + |\nabla u|^2)\, dxdy},$$

where $\nabla u = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y})$ and $|\nabla u|^2 = |\frac{\partial u}{\partial x}|^2 + |\frac{\partial u}{\partial y}|^2$. Roughly speaking, completeness means that all possible Cauchy sequences should find their limits inside that space. In fact, if V is not complete, we add the limits to make it complete. One is curious to know '*why do we require completeness ?*' In practice, we need to solve the above problem by using some approximation schemes, i.e., we should like to approximate u by a sequence of approximate solutions $\{u_n\}$. Many times $\{u_n\}$ forms a Cauchy sequence (that is a part of convergence analysis). Unless the space is complete, the limit may not be inside that space. Therefore, a more **desirable** space is the completion of V. Subsequently, we shall see that the completion of V is $H_0^1(\Omega)$. This is a Hilbert Sobolev Space and is stated as: $\{v \in L^2(\Omega) : \frac{\partial v}{\partial x}, \frac{\partial v}{\partial y} \in L^2(\Omega)$ and $u(x, y) = 0, (x, y) \in \partial\Omega\}$. Obviously, the square integrable function may not have partial derivatives in the usual sense. In order to attach a meaning, we shall generalize the concept of differentiation and that we shall discuss *prior* to the introduction of Sobolev spaces. One should note that the meaning of the H^1-function v on Ω which satisfies $v = 0$ has to be understood in a general sense.

If we accept (1.4) as a more general formulation, and the equation (1.1) is its Euler form, then it is natural to ask:

'*Does every PDE have a weak form which is of the form (1.4) ?*'

The answer is simply in negative. Say, for a flow problem with a **transport** or **convective** term :

$$-\Delta u + \vec{b} \cdot \nabla u = f,$$

it does not have an energy formulation like (1.4). So, next question would be:

'*Under what conditions on PDE, such a minimization form exists ?*'

More over,

'*Is it possible to have a more general weak formulation which in a particular situation coincides with minimization of the energy ?*'

This is what we shall explore in the course of these lectures.

If formally we multiply (1.1) by $v \in V$ (space of admissible displacements) and apply Gauss divergence theorem, the contribution due to boundary terms becomes zero as $v = 0$ on $\partial\Omega$. Then we obtain

$$\int_\Omega (\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y})\, dxdy = \int_\Omega f v\, dxdy. \tag{1.5}$$

For flow problem with a tranport term $\vec{b}.\nabla u$ (1.5) we have an extra term $\int_\Omega \vec{b}.\nabla u\, v\, dxdy$ added to the left hand side of (1.5). This is a more general weak formulation and we shall also examine the relation between (1.4) and (1.5). Given such a weak formulation, '*Is it possible to establish its wellposedness* [1] *?*' Subsequently in Chapter 3, we shall settle this issue by using **Lax-Milgram Lemma**.

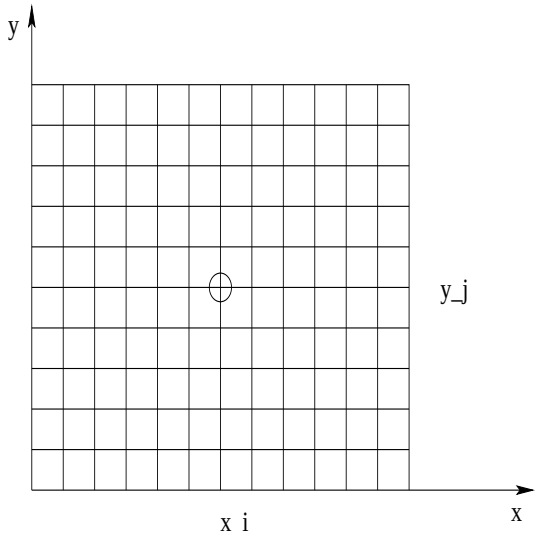
Very often problem like (1.1) doesn't admit exact or analytic solutions. For the problem (1.1), if the boundary is irregular, i.e.,Ω need not be a square or a circle, it is difficult to obtain an analytic solution. Even when analytic solution is known, it may contain complicated terms or may be an infinite series. In both the cases, one resorts to numerical approximations. One of the objectives of the numerical procedures for solving differential equations is to cut down the degrees of freedom (the solutions lie in some infinite dimensional spaces like the Hilbert Sobolev spaces described above) to a finite one so that the discrete problem can be solved by using computers. Broadly speaking, there are three numerical methods for solving PDE's: **Finite Difference Methods, Finite Element Procedures and Boundary Integral Techniques**. In fact, we have also **Spectral Methods**, but we may prefer to give some comments on the class of spectral methods rather describing it as a separate class of methods, while dealing with finite element techniques. Below, we discuss only finite difference methods and present a comparision with finite element methods.

1.2 Finite Difference Methods (FDM).

Let us assume that Ω is a unit square,i.e., $\Omega = (0,1) \times (0,1)$ in (1.1)

In order to derive a finite difference scheme, we first discretize the domain and then discretize the equation. For fixed positive integers M and N, let h_1,

[1] The problem is said to be wellposed (in the sense of Hadamard) if it has a solution, the solution is unique and it depends continuously on the data



the spacing in x-direction be $\frac{1}{M}$ with mesh points $x_i = ih_1, i = 0, 1, \ldots, M$ and let $h_2 := \frac{1}{N}$ with mesh points or nodal points in the direction of y as $y_j = jh_2, j = 0, 1, \ldots, N$.

For simplicity of exposition, consider uniform partition in both directions, i.e., $M = N$ and $h = h_1 = h_2$. The set of all nodal points $\{(x_i, y_j), 0 \leq i, j \leq M\}$ forms a mesh in $\overline{\Omega}$. The finite difference scheme is now obtained by replacing the second derivatives in (1.1) at each nodal pints (x_i, y_j) by central difference quotients. Let u at (x_i, y_j) be called u_{ij}. For fixed j using Taylor series expansions of $u(x_{i+1}, y_j)$ and $u(x_{i-1}, y_j)$ around (x_i, y_j), it may be easily found out that

$$\frac{\partial^2 u}{\partial x^2}|_{(x_i,y_j)} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h^2} + O(h^2). \tag{1.6}$$

Similarly for fixed i, replace

$$\frac{\partial^2 u}{\partial y^2}|_{(x_i,y_j)} = \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h^2} + O(h^2). \tag{1.7}$$

Now, we have at each interior points $(x_i, y_j), 1 \leq i, j \leq M - 1$,

$$-\frac{1}{h^2}[u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j}] = f_{ij} + O(h^2), \tag{1.8}$$

where $f_{ij} = f(x_i, y_j)$. However, it is not possible to solve the above system (because of the presence of $O(h^2)$-term). In order to drop this term, define U_{ij}, an approximation of u_{ij} as solution of the following algebraic equations:

$$U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1} - 4U_{i,j} = -h^2 f_{ij}, 1 \leq i, j \leq M-1 \quad (1.9)$$

with $U_{0,j} = 0 = U_{M,j}, 0 \leq j \leq M$ and $U_{i,0} = U_{i,M} = 0, 0 \leq i \leq M$.

Writing a vector U in lexicographic ordering, i.e.,

$$U = [U_{1,1}, U_{2,1}, \ldots, U_{M-1,1}, U_{1,2}, \ldots, U_{M-1,2}, \ldots, U_{1,M-1}, \ldots, U_{M-1,M-1}]^T,$$

it is easy to put (1.9) in a block diagonal form

$$AU = F \quad (1.10)$$

where $A = diag[B, \ldots, B]$ and each block $B = triad[1, -4, 1]$ is of size $(M-1) \times (M-1)$. The size of the matrix A is $(M-1)^2 \times (M-1)^2$. Note that the matrix A is very sparse (only three or five non-zero elements in each row of length $(M-1)^2$). This is an example of a large sparse system which is diagonally dominant (not strictly) [2] and symmetric.

One may be curious to know (*prior* to the actual computation) :

'*Is this system (1.10) solvable?*'

Since A is diagonally dominant and irreducible [3], then A is invertible. How to compute U more efficiently is a problem related to Numerical Linear Algebra. However, it is to be noted that for large and sparse systems with large bandwidth, the iterative methods are more successful.

One more thing which bothers a numerical analyst (even the user community) is the convergence of the discrete solution to the exact solution as the mesh size becoming smaller and smaller. Sometimes, it is important to know '*how fast it converges*'. This question is connected with the order of convergence. Higher rate of convergence implies faster computational process. Here, a measurement (norm) is used for quantification. At this stage, let us recall one of the important theorem called **Lax- Ritchmyer Equivalence Theorem** on the convergence of discrete solution. It roughly states that '*For a wellposed linear problem, a stable numerical approximation is convergent if and only if it is consistent*'.

The consistency is somewhat related to the (local) truncation error. It is defined as the amount by which the exact or true solution $u(x_i, y_j)$ does not satisfy the difference scheme (1.9) by $u(x_i, y_j)$. From (1.8), the truncation error say τ_{ij} at (x_i, y_j) is $O(h^2)$. Note that this is the maximum rate of convergence we can expect once we choose this difference scheme. The concept of stability is connected with the propagation of round off or chopping off errors during the course of computation. We call a numerical method stable , of course with respect to some measurement, if small changes (like these accumulation of errors) in the data do not give rise to a drastic change in the solution. Roughly speaking, computations in different machines should not give drastic changes

[2] A matrix $A = [a_{ij}]$ is diagonally dominant if $|a_{ii}| \geq \sum_j |a_{ij}|, \forall i$. If strict inequality holds , then A is strictly diagonally dominant.

[3] The matrix A is irreducible if it does not have a subsystem which can be solved independently.

in the numerical solutions, i.e., stability will ensure that the method does not depend on a particular machine we use. Mostly for stability, we bound with respect to some measurements the changes in solutions (difference between unperturbed and pertubed solutions) by the perturbation in data. However, for linear discrete problems a uniform *a priori* bound (uniform with respect to the discretization parameter) on the computed solution yields a stability of the system.

For (1.10), the method is stable with respect some norm say $\|\cdot\|_\infty$ if $\|U\|_\infty \leq C\|F\|_\infty$, where C is independent of h and $\|U\|_\infty = \max\limits_{0\leq i,j\leq M} |U_{ij}|$. This is an easy consequence of invertibility of A and boundedness of $\|A^{-1}\|_\infty$ independent of h. It is possible to discuss stability with respect to other norms like Euclidean norm.

If

$$\overline{u} = [u_{1,1}, \ldots, u_{2,1}, \ldots, u_{M-1,1}, \ldots, u_{M-1,M-1}]^T,$$

then the error $\overline{u} - U$ satisfies

$$A(\overline{u} - U) = A\overline{u} - AU = A\overline{u} - F = \tau,$$

where τ is a vector representing the local truncation errors. Hence, using stability

$$\|\overline{u} - U\|_\infty \leq C\|\tau\| \leq Ch^2,$$

where C depends on maximum norm of the 4th derivative of the exact solution u. It is to be remarked here that unless the problem is periodic it is impossible to achieve $u \in C^4(\overline{\Omega})$, if Ω is a square (in this case at most $u \in C^2(\Omega) \cap C(\overline{\Omega})$). Therefore, the above error estimate or order of convergence looses its meaning unless the problem is periodic. On the other hand, if we choose boundary $\partial\Omega$ of Ω to be very nice (i.e., the boundary can be represented locally by the graph of a nice function), then u may be in $C^4(\Omega)$, but we loose order of convergence as mesh points may not fall on the boundary of Ω and we may use some interpolations to define the boundary values on the grids.

One possible plus point "*why it is so popular amongst the user community*" is that it is easy to implement and in the interior of the domain, it yields a reasonable good approximation. Some of the disadvantages, we would like to recall are:

- It is difficult to apply FDM to the problems with irregular boundaries
- It requires higher smoothness on the solutions for the same order of convergence (this would be clarified in the context of finite element method

). Nicer boundary may lead to smoother solutions, but there is a deterioration in the order of convergence near the boundary.

- In order to mimick the basic properties of the physical system in the entire domain, it is desirable to refine the mesh only on the part of the domain, where there is a stiff gradient (i.e., the magnitude of ∇u is large) or a boundary layer. Unfortunately, for FDM, it is difficult to do refinements locally. Therefore, a finer refinement on the entire domain will dramatically increase the computational storage and cost.

- The computed solution is obtained only on the grids. Therefore, for computation of solution on the points other than the grid points, we need to use some interpolation techniques.

However, the finite element methods take care of the above points, i.e., it can be very well applied to problems with irregular domains, requires less smoothness to obtain the same order of convergence (say for $O(h^2)$, we need $u \in H^2 \cap H_0^1$ and this smoothless for u is easy to derive for a square domain); it is possible to refine the mesh locally and finally, the solution is computed at each and every point in the domain.

1.3 Finite Element Methods (FEM).

A basic step in FEM is the reformulation (1.5) of the original problem (1.1). Note that there is a reduction in the differentiation and this is a more desirable property for the numerical computation (roughly speaking, numerical differentiation brings ill conditioning to the system).

Then by discretizing the domain, construct a finite dimensional space V_h where h is the discretization parameter with property that the basis functions have small supports in Ω . Pose the reformulated problem in the finite dimensional setting as : Find $u_h \in V_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h dxdy = \int_\Omega f v_h dxdy \tag{1.11}$$

The question is to be answered

'*Is the above system (1.11) solvable ?*'

Since (1.11) leads to a system of linear solutions, one appeals to the techiniques in numerical linear algebra to solve this system.

Next question, we may like to pose as:

'*Does u_h converge to u in some measurements as $h \longrightarrow 0$?*'

If so, *Is it possible to find its order of convergence ?*

We may wonder, whether we have used (Lax-Ritchmyer) equivalence theorem here. Consistency is related to the truncation of any $v \in V$ by $v_h \in V_h$. If V is a Hilbert space, then v_h is the best approximation of v in V_h and it is possible quantify the error $v - v_h$ in the best approximation. The beauty of FEM is that stability is straightforward and it mimicks for the linear problems, the *a priori*

bounds for the true solution. Note that the form of the weak formulation of the problem does not change for the discrete problem, therefore, the wellposedness of the original (linear) problem invariably yields the wellposedness of the discrete problem and hence, the the proof of stability becomes easier.

These *a priori* error estimates $u - u_h$ in some measurements tell the user that the method works and we have confidence in computed numbers. However, many more methods were developed by user community without bothering about the convergence theory. But they have certain checks like if the computed solution stabilizes after some decimal points as h decreases, then computed results make sense. But this thumb rule may go wrong.

One more useful problem, we may address is related to the following:

"*Given a tolerance (say* ε *) and a measurement (say a norm* $\|\cdot\|$*), CAN we compute a reliable and efficient approximate solution* u_h *such that* $\|u-u_h\| \leq \varepsilon$ *?*"

By efficient approximate solution, we mean computed solution with minimal computational effort. This is a more relevant and practical question and I shall try my best to touch upon these points in the course of my lectures.

Chapter 2

Theory of Distribution and Sobolev Spaces

In this chapter, we quickly review L^2-space, discuss distributional derivatives and present some rudiments of Sobolev spaces.

2.1 Review of L^2- Space.

To keep the presentation at an elementary level, we try to avoid introducing the L^2-space through measure theoretic approach. However, if we closely look at the present approach, the notion of integration is assumed tactically.

Let Ω be a bounded domain in $\mathbb{R}^{\mathrm{d}}(\mathrm{d} = 1, 2, 3)$ with smooth boundary $\partial\Omega$. All the functions, we shall consider in the remaining part of this lecture note, are assumed to be real valued. It is to be noted that most of the results developed here can be carried over to the complex case with appropriate modifications.

Let $L^2(\Omega)$ be the space of all square integrable functions defined in Ω, i.e.,

$$L^2(\Omega) := \{v : \int_\Omega |v|^2 \, dx < \infty\}.$$

Define a mapping $(\cdot, \cdot) : L^2(\Omega) \times L^2(\Omega) \mapsto \mathbb{R}$ by

$$(v, w) := \int_\Omega v(x)w(x) \, dx.$$

It satisfies all the properties of an innerproduct except for: $(v, v) = 0$ does not imply that v is identically equal to zero. For an example, consider $\Omega = (0, 1)$ and set for $n > 1$, $v_n(x) = n$ if $x = \frac{1}{n}$ and is equal to zero elsewhere. Now $(v_n, v_n) = \int_0^1 |v_n|^2 \, dx = 0$, but $v_n \neq 0$. Here, the main problem is that there are infinitely many functions whose square integrals are equal to zero, but the functions are not identically equaly to zero. It was **Lebesgue** who generalized the concept of identically equal to zero in early 20th century. If we wish to

make $L^2(\Omega)$ an innerproduct space, then define an equivalence relation on it as follows:

$$v \equiv w \text{ if and only if } \int_\Omega |v|^2\,dx = \int_\Omega |w|^2\,dx.$$

With this equivalence relation, the set $L^2(\Omega)$ is decomposed into disjoint classes of equivalent functions. In the language of measure theory, we call this relation almost equal everywhere (a.e.). Now $(v, v) = 0$ implies $v = 0\,a.e.$ This is indeed a generalization of the notion of " *identically equals to zero*" to " *zero almost everywhere*". Note that when $v \in L^2(\Omega)$, it is a representer of that equivalence class. The induced L^2-norm is now given by

$$\|v\| = (v, v)^{\frac{1}{2}} = \left(\int_\Omega |v|^2\,dx\right)^{\frac{1}{2}}.$$

With this norm, $L^2(\Omega)$ is a complete (the proof of this is a nontrivial task and is usually proved using measure theoretic arguments) innerproduct space, i.e., it is a Hilbert space.

Some Properties.

(i) The space $C(\overline{\Omega})$ (space of all bounded and uniformly continuous functions in $\Omega$is continuously imbedded [1] in $L^2(\Omega)$. In other words, it is possible to identify a uniformly continuous function in each equivalence class. Note that $v \in L^2(\Omega$ without being in $C(\overline{\Omega})$. For example the function $v(x) = x^{-\alpha}, x \in (0, 1)$ is in $L^2(0, 1)$ provided $0 < \alpha < 2$, but not in $C[0, 1]$.

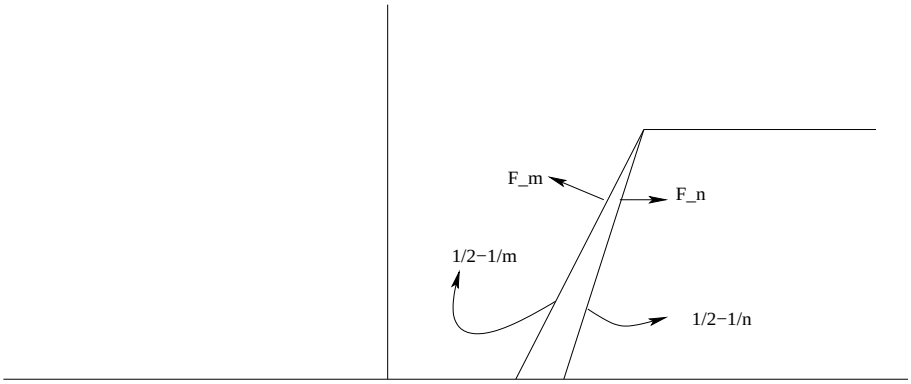
(ii) The space $C(\Omega)$ may not be a subspace of $L^2(\Omega)$. For example, let $\Omega = (0, 1)$ and set $v(x) = \frac{1}{x}$. Then, $v \in C(\Omega)$, but not in $L^2(\Omega)$. Of course, v be in $L^2(\Omega)$ without being in $C(\Omega)$. For example the function $v(x) = 0\,0 \le x < \frac{1}{2}$ and for $x \in [\frac{1}{2}, 1], v(x) = 1$ is discontinuous at $x = \frac{1}{2}$, but it is not in $L^2(0, 1)$

(iii) Let $L^1(\Omega) = \{v : \int_\Omega |v|\,dx < \infty\}$. For bounded domain Ω, the space $L^2(\Omega)$ is continuously imbedded in $L^1(\Omega)$. However for unbounded domains, the above imbedding may not hold. For example, when $\Omega = (1, \infty)$, consider $v(x) = \frac{1}{x}$. Then, v does not belong to $L^1(\Omega)$, while it is in $L^2(\Omega)$. For bounded domains, there are functions which are in $L^1(\Omega)$, but not in $L^2(\Omega)$. One such example is $v(x) = \frac{1}{x^\alpha},\ 1 > \alpha \ge \frac{1}{2}$ in $(0, 1)$.

The space $V = \{v \in C(\Omega) : \int_\Omega |v|^2\,dx < \infty\}$ is not complete with respect to L^2-norm. We can now recall one example in our analysis course. Consider a sequence $\{f_n\}_{n\ge 2}$ of continuous functions in $C(0, 1)$ as
that is each $f_n \in C(0, 1)$ and

$$f_n(x) = \begin{cases} 1, & \frac{1}{2} \le x < 1 \\ 0, & 0 < x \le \frac{1}{2} - \frac{1}{n}. \end{cases}$$

[1] The identity map $i : C(\overline{\Omega}) \mapsto L^2(\Omega)$ is continuous that is $\|v\| \le C \max_{x\in\overline{\Omega}} |v(x)|$.



Being a continuous function , the graph of f_n is shown as in the above figure. It is apparently clear from the graphs of f_n and f_m that

$$\|f_n - f_m\| = \left(\int_\Omega |f_n(x) - f_m(x)|^2 \, dx \right)^{\frac{1}{2}}$$

can be made less than and equal to a preassigned $\epsilon > 0$ by choosing a positive integer N with $m, n > N$. Therefore, $\{f_n\}$ forms a Cauchy sequence. However, f_n converges to a discontinuous function f, where $f(x) = 1$ for $\frac{1}{2} \leq x < 1$ and it is zero elsewhere. Thus, the space V is not complete. If we complete it with respect to L^2-norm, we obtain $L^2(\Omega)$-space.

Sometimes, we shall also use the space of all locally integrable functions in Ω denote by$L^1_l oc(\Omega)$ which is defined as

$$L^1_{loc}(\Omega) := \{v \in L^1(K), \text{ for every compact set } K \text{ with } \overline{K} \subset \Omega\}.$$

Now, the space $L^1(\Omega)$ is continuously imbedded in $L^1_{loc}(\Omega)$. When Ω is unbounded say $\Omega = (0, \infty)$, constant functions, x^α with α integers, e^x are all in $L^1_{loc}(0, \infty)$ with out being in $L^1(0, \infty)$. Similarly, when Ω is bounded that is say $\Omega = (0, 1)$, the functions $\frac{1}{x^\alpha}$, $\alpha \geq 1$, $\log \frac{1}{x}$ are all locally integrable but not in $L^1(0, 1)$. Note that $L^1_{loc}(\Omega)$ contains all of $C(\Omega)$ without growth restriction.

Similarly, when $1 \leq p \leq \infty$, the space of all p-th integrable functions will be denoted by $L^p(\Omega)$. For $1 \leq p < \infty$, the norm is defined by

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

and for $p = \infty$

$$\|v\|_{L^\infty(\Omega)} := \text{essup}_{x \in \Omega} |v(x)|.$$

2.2 A Quick Tour to the Theory of Distributions.

We would like to generalize the concept of differentiation to the functions belonging to a really large space. One way to catch hold of such a space is to construct a smallest nontrivial space and then take its topological dual. As a preparation to choose such a space, we recall some notations and definitions.

A multi-index $\alpha = (\alpha_1, \cdots, \alpha_d)$ is a d-tuple with nonnegative integer elements. By an order of α, we mean $|\alpha| = \sum_{i=1}^{d} \alpha_i$. The notion of multi-index is very useful for writing the higher order partial derivatives in compact form. Define α^{th} order partial derivative of v as

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1}, \cdots, \partial x_d^{\alpha_d}}, \ \alpha = (\alpha_1, \cdots, \alpha_d).$$

Example 2.1. When $d = 3$ and $\alpha = (1, 2, 1)$ with $|\alpha| = 4$, the

$$D^\alpha v = \frac{\partial^4 v}{\partial x_1 \partial x_2^2 \partial x_3}.$$

In case $|\alpha| = 2$, we have the following 6 possibilities :$(2, 0, 0)$, $(1, 1, 0)$, $(0, 2, 0)$, $(0, 1, 1)$, $(0, 0, 2)$, $(1, 0, 1)$ and D^α will represent all the partial derivatives of order 2. When $\alpha \leq 3$, D^α would mean all partial derivatives up and including 3.

Let $C^m(\Omega) := \{v : D^\alpha v \in C(\Omega), |\alpha| \leq m\}$. Similarly, we define the space $C^m(\overline{\Omega})$ as m-times continuously differentiable functions with uniformly continuous derivatives up to order m. By $C^\infty(\Omega)$, we mean a space of infinitely differentiable functions in Ω, i.e., $C^\infty(\Omega) = \cap C^m(\Omega)$. The support of a function v called $supp\, v$ is defined by

$$supp\, v := \overline{\{x \in \Omega : v(x) \neq 0\}}.$$

If this set is compact (i.e., if it is bounded in this case) and $supp\, v \subset\subset \Omega$, then v is said to have 'compact support' with respect to Ω. Denote by $C_0^\infty(\Omega)$, a space of all infinitely differentiable functions with compact support in Ω. In fact, this is a nonempty set and it can be shown as has been done below that it has as many elements as the points in this set Ω.

Example 2.2. For $\Omega = \mathbb{R}$, consider the function

$$\phi(x) = \begin{cases} \exp(\frac{1}{|x|^2 - 1}), & |x| < 1, \\ 0, & |x| \geq 1. \end{cases}$$

In order to show that $\phi \in C_0^\infty(\Omega)$, it is enough to check the continuity and differentiability properties only at the points $x = \pm 1$. Apply L'Hopsital rule to conclude the result.

Similarly, when $\Omega = \mathbb{R}^d$, consider

$$\phi(x) = \begin{cases} \exp(\frac{1}{\|x\|^2-1}), & \|x\| < 1, \\ 0, & \|x\| \geq 1, \end{cases}$$

where $\|x\|^2 = \sum_{i=1}^{d} |x_i|^2$. Then, since the function is radial, we can in a similar manner prove that $\phi \in C_0^\infty(\Omega)$. For all other points we can simply translate and show the existence of such a function.

Exersise 2.1 Show that $\phi \in C_0^\infty(\mathbb{R}^d)$

When Ω is a bounded domain in $\mathbb{R}^d$, consider for any point $x_0 \in \Omega$ and small $\epsilon > 0$ the following function

$$\phi_\epsilon(x) = \begin{cases} \exp(\frac{\epsilon}{\|x-x_0\|^2-\epsilon}), & \|x - x_0\| < \epsilon, \\ 0, & \|x - x_0\| \geq \epsilon. \end{cases}$$

This function has the desired property and can be constructed for each and every point in the domain. Since the domain Ω is open, it is possible to find ϵ for point $x_0 \in \Omega$ so that the the ϵ-ball around x_0 is inside the domain. Therefore, the space $C_0^\infty(\Omega)$ is a nontrivial vector space and this is the smallest space we are looking for. In order to consider the topological dual of this space, we need to equip $C_0^\infty(\Omega)$ with a topology with respect to which we can discuss convergence of a sequence. Since we shall be only using the concept of convergence of a sequence to discuss the continuity of the linear functionals defined on $C_0^\infty(\Omega)$, we refrain from defining the topology. We say :

'There is a structure (topology) on $C_0^\infty(\Omega)$ with respect to which the convergence of a sequence $\{\phi_n\}$ to ϕ has the following meaning, i.e., $\phi_n \mapsto \phi$ in $C_0^\infty(\Omega)$ if the following conditions are satisfied

(i) there is a common compact set K in Ω with $\overline{K} \subset \Omega$ such that

$$supp\, \phi_n,\ supp\, \phi \subset K.$$

(ii) $D^\alpha \phi_n \mapsto D^\alpha \phi$ uniformly in K, for all multi-indices α.

When $C_0^\infty(\Omega)$ is equipped with such a topology, we call it $\mathcal{D}(\Omega)$, the space of test functions. **Definition.** A continuous linear functional on $\mathcal{D}(\Omega)$ is called a **distribution**, i.e., $T : \mathcal{D}(\Omega) \mapsto \mathbb{R}$ is called a distribution if T is linear and $T(\phi_n) \mapsto T(\phi)$ as $\phi_n \mapsto \phi$ in $\mathcal{D}(\Omega)$. The space of all distributions will be denoted by $\mathcal{D}'(\Omega)$. This is the topological dual of $\mathcal{D}(\Omega)$.

We shall use $< \cdot, \cdot >$ for duality pairing between $\mathcal{D}'(\Omega)$ and $\mathcal{D}(\Omega)$.

Example 2.3. *Every integrable function defines a distribution.*

Let $f \in L^1(\Omega)$. Then for $\phi \in \mathcal{D}(\Omega)$, set T_f as

$$T_f(\phi) = \int_\Omega f\phi\, dx.$$

We now claim that $T_f \in \mathcal{D}'(\Omega)$. Linearity of T_f is a consequence of additivity as well as homogeneity property of the integral. It, therefore, remains to show that

T_f is continuous, i.e., as $\phi_n \mapsto \phi$ in $\mathcal{D}(\Omega)$, we need to prove $T_f(\phi_n) \mapsto T_f(\phi)$. Note that

$$|T_f(\phi_n) - T_f(\phi)| = \left| \int_\Omega f(\phi_n - \phi)\, dx \right|.$$

Since ϕ_n converges to ϕ in $\mathcal{D}(\Omega)$, there is a common compact set K in Ω such that $supp\, \phi_n$, $supp\, \phi \subset K$ and $\phi_n \mapsto \phi$ uniformly in K. Therefore,

$$|T_f(\phi_n) - T_f(\phi)| \leq \int_K |f(\phi_n - \phi)|\, dx \leq \max_{x \in K} |\phi_n(x) - \phi(x)| \int_\Omega |f|\, dx.$$

Since $\int_\Omega |f|\, dx < \infty$, $T_f(\phi_n) \mapsto T_f(\phi)$, as $n \mapsto \infty$ and the result follows.

If f and g are in $L^1(\Omega)$ with $f = g$ a.e., then $T_f = T_g$. Thus, we identify $T_f = f$ and $T_f(\phi) =< f, \phi >$. Therefore, $L^1(\Omega) \subset \mathcal{D}'(\Omega)$.

Excercise 2.1. Show that every square integrable function defines a distribution.
Excercise 2.2. Prove that every locally integrable function defines a distribution.

Example 2.4. The Dirac delta δ function (it is a misnomer and it is a functional, but in literature it is still called a function as P. A. M. Dirac called this a function when he discovered it) concentrated at the origin ($0 \in \Omega$) is defined by

$$\delta(\phi) = \phi(0),\ \forall \phi \in \mathcal{D}(\Omega).$$

This, in fact, defines a distribution. Linearity is trivial. As $\phi_n \mapsto \phi$ in $\mathcal{D}(\Omega)$, i.e., $\phi_n \mapsto \phi$ uniformly on a common compact support, then,

$$\delta(\phi_n) = \phi_n(0) \mapsto \phi(0) = \delta(\phi),\ \forall \phi \in \mathcal{D}(\Omega).$$

Therefore, $\delta \in \mathcal{D}'(\Omega)$ and we write $\delta(\phi) =< \delta, \phi >= \phi(0)$.
Note that the Dirac delta function can not be generated by locally integrable function (the proof can be obtained by using arguments by contradiction [2]). Therefore, depending on whether the distribution is generated by locally integrable function or not, we have two types of distributions: *Regular* and *Singular* distributions.

(i) *Regular.* The distributions which are generated by locally integrable functions.

[2]Suppose that the Dirac delta is genereted by a locally integreble function say f. Then we define the distribution $T_f(\phi) = \delta(\phi)$ that is $T_f(\phi) = \phi(0)$ for $\phi \in \mathcal{D}(\mathbb{R}^{\mathrm{d}})$. For $\epsilon > 0$ it is possible to find $\phi_\epsilon \in \mathcal{D}(\mathbb{R}^{\mathrm{d}})$ with $supp\phi_\epsilon \subset B_\epsilon(0)$, $0 \leq \phi_\epsilon \leq 1$ and $\phi_\epsilon \equiv 1$ in $B_{\epsilon/2}(0)$. Note that $\delta(\phi_\epsilon) = 1$. But on the other hand

$$\delta(\phi_\epsilon) = T_f(\phi_\epsilon) = \int_{\mathbb{R}^{\mathrm{d}}} f\phi_\epsilon\, dx = \int_{B_\epsilon(0)} f\phi_\epsilon\, dx \leq \int_{B_\epsilon(0)} |f| dx.$$

Since f is localy integrable , $\int_{B_\epsilon(0)} |f| dx \longrightarrow 0$ as $\epsilon \longrightarrow 0$, that is $\delta(\phi_\epsilon) \longrightarrow 0$. This leads to a contradiction and hence the Direc delta function can not be genereted by locally integrable functions

(ii) *Singular.* The distributions is called singular if it is not regular. For example, the Dirac delta function is a singular distribution.

Note that if $T_1, T_2 \in \mathcal{D}'(\Omega)$ are two regular distributions satisfying

$$< T_1, \phi >=< T_2, \phi > \ \forall \phi \in \mathcal{D}(\Omega),$$

then $T_1 = T_2$. This we shall not prove , but like to use it in future[3]

We can also discuss the convergence of the sequence of distributions $\{T_n\}$ to T in $\mathcal{D}'(\Omega)$. Say, $\{T_n\} \mapsto T$ in $\mathcal{D}'(\Omega)$ if

$$< T_n, \phi > \mapsto < T, \phi >, \ \forall \phi \in \mathcal{D}(\Omega).$$

As an example, we consider $\epsilon > 0$

$$\rho_\epsilon(x) = \begin{cases} A\epsilon^{-d} \exp\left(\frac{-\epsilon^2}{\epsilon^2 - \|x\|^2}\right) \phi(x), & \|x\| < \epsilon, \\ 0, & \|x\| \geq \epsilon, \end{cases}$$

where $A^{-1} = \int_{\|x\| \leq 1} \exp\left(\frac{-1}{1-\|x\|^2}\right) dx$. Note that $\int_{\mathbb{R}^d} \rho_\epsilon \, dx = 1$. It is easy to check that $\rho_\epsilon \in D'(\mathbb{R}^d)$. We now claim that $\rho_\epsilon \longrightarrow \delta$ in $D'(\mathbb{R}^d)$ as $\epsilon \longrightarrow 0$. For $\phi \in D(\mathbb{R}^d)$, we note that

$$\begin{aligned} \langle \rho_\epsilon, \phi \rangle &= A\epsilon^{-d} \int_{\mathbb{R}^d} \exp\left(\frac{-\epsilon^2}{\epsilon^2 - \|x\|^2}\right) \phi(x) \, dx. \\ &= A \int_{\mathbb{R}^d} \exp\left(\frac{-1}{1 - \|y\|^2}\right) \phi(\epsilon y) \, dy \\ &= \phi(0) + A \int_{\mathbb{R}^d} \exp\left(\frac{-1}{1 - \|y\|^2}\right) (\phi(\epsilon y) - \phi(0)) \, dy. \end{aligned}$$

As $\epsilon \longrightarrow 0$, using the Lebesgue dominated convergence theorem, we can take the limit under the integral sign and therefor, the integral term become zero. Thus we obtain

$$\lim_\epsilon \langle \rho_\epsilon, \phi \rangle = \phi(0) = \langle \delta, \phi \rangle.$$

Hence the result follows.

Distributional Derivatives. In order to generalize the concept of differentiation, we note that for $f \in C^1([a,b])$ and $\phi \in \mathcal{D}([a,b])$, integration by parts yields

$$\int_a^b f'(x)\phi(x) \, dx = f\phi|_{x=a}^{x=b} - \int_a^b f(x)\phi'(x) \, dx.$$

Since $supp\, \phi \subset (a,b)$, $\phi(a) = \phi(b) = 0$. Thus,

$$\int_a^b f'(x)\phi(x) \, dx = -\int_a^b f(x)\phi'(x) \, dx = - < f, \phi' >, \forall \phi \in \mathcal{D}(\Omega).$$

[3]This is a cosequence of the following Lebesque lemma: Le T_f and T_g are two regular distribution genereted by locally integrable functions f and g, respectively. Then $T_f = T_g$ if and only if $f = g$ a.e.

We observe that the meaning of derivative of f can be attached through the term on the right hand side. However, the term on the right hand side is still be meaningful even if f is not differentiable. Note that f can be a distribution. This motivates us to define a derivative of a distribution.

If $T \in \mathcal{D}'(\Omega)$ $(\Omega = (a, b))$, then define DT as

$$(DT)(\phi) = - < T, D\phi >, \quad \forall\ \phi \in \mathcal{D}(\Omega).$$

Here, $D\phi$ is simply the derivative of ϕ. Observe that (DT) is a linear map. For $\phi_1,\ \phi_2 \in \mathcal{D}(\Omega)$ and $a \in \mathbb{R}$,

$$\begin{aligned}(DT)(\phi_1 + a\phi_2) &= - < T, D(\phi_1 + a\phi_2) > \\ &= - < T, D\phi_1 > -a < T, D\phi_2 > \\ &= (DT)(\phi_1) + a(DT)(\phi_2).\end{aligned}$$

For continuity, we first infer that as $\phi_n \to \phi$ in $\mathcal{D}(\Omega)$, $D\phi_n \to D\phi$ on $\mathcal{D}(\Omega)$. Now,

$$(DT)(\phi_n) = - < T, D\phi_n > \to - < T, D\phi >= (DT)(\phi).$$

Therefore, $DT \in \mathcal{D}'(\Omega)$.

More precisely, if T is a distribution, i.e., $T \in \mathcal{D}'(\Omega)$, then αth order distributional derivative, say $D^\alpha T \in \mathcal{D}'(\Omega)$, is defined by

$$< D^\alpha T, \phi >= (-1)^\alpha < T, D^\alpha \phi >, \quad \phi \in \mathcal{D}(\Omega).$$

Example 2.5. Let $\Omega = [-1, 1]$ and let $f = |x|$ in Ω. Obviously, $f \in \mathcal{D}'(-1, 1)$. To find its distributional derivative Df, write first for $\phi \in \mathcal{D}(\Omega)$

$$\begin{aligned}< Df, \phi > &= - < f, D\phi > \\ &= -\int_{-1}^{1} f(x)\phi'(x)dx \\ &= -\int_{-1}^{0} (-x)\phi'(x)dx - \int_{0}^{1} x\phi'(x)dx.\end{aligned}$$

On integrating by parts and using $\phi(1) = \phi(-1) = 0$, it follows that

$$\begin{aligned}< Df, \phi > &= x\phi]_{-1}^{0} - \int_{-1}^{0} \phi(x)dx - x\phi]_{0}^{1} + \int_{0}^{1} \phi(x)dx \\ &= \int_{-1}^{1} (-1)\phi(x)dx + \int_{0}^{1} 1\phi(x)dx \\ &= \int_{-1}^{1} H(x)\phi(x)dx \\ &= < H, \phi >, \quad \forall\ \phi \in \mathcal{D}(\Omega),\end{aligned}$$

where

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

As $< Df, \phi >=< H, \phi >$, $\forall \phi \in \mathcal{D}(\Omega)$, we obtain $Df = H$. The first distributional derivative of f is a Heaviside step function.

Again to find DH, note that

$$\begin{aligned} < DH, \phi > &= - < H, D\phi >= -\int_{-1}^{1} H(x)\phi'(x)dx \\ &= \int_{-1}^{0} \phi'(x)dx - \int_{0}^{1} \phi'(x)dx \\ &= \phi(0) - \phi(-1) - \phi(1) + \phi(0) \\ &= 2\phi(0) = 2 < \delta, \phi > . \end{aligned}$$

Therefore,

$$DH = 2\delta \quad \text{or} \quad D^2 f = 2\delta.$$

In the above example, we observe that if a locally integrable function has a classical derivative a.e., which is also locally integrable, then then the classical derivative may not be equal to the distributional derivative. However, if a function is absolute continous, then its classical derivative coincides with its distributional derivative.

Exercises:

1.) Find Df, when

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

$f(x) = \log x$, $0 < x < 1$, $f(x) = (\log \frac{1}{x})^k$, $k < \frac{1}{2}$.

2.) Find Df, when $f(x) = \sin(\frac{1}{x})$, $0 < x < 1$ and $f(x) = \frac{1}{x}$, $0 < x < 1$.

3.) Find $D^2 f$ when $f(x) = x|x|$, $-1 < x < 1$.

For $\Omega \subset \mathbb{R}^d$, let $f \in C^1(\bar{\Omega})$ and $\phi \in \mathcal{D}(\Omega)$. Then using integration by parts, we have

$$\int_\Omega \frac{\partial f}{\partial x_i}\phi dx = -\int_\Omega f \frac{\partial \phi}{\partial x_i}dx = - < f, D_i\phi > .$$

As in one variable case, set the first generalized partial derivative of a distribution $T \in \mathcal{D}'(\Omega)$ as

$$< D_i T, \phi >= - < T, D_i\phi >, \quad \forall \phi \in \mathcal{D}(\Omega).$$

It is easy to check that $D_i T \in \mathcal{D}'(\Omega)$. Similarly, if $T \in \mathcal{D}'(\Omega)$ and α is a multi-index, then the α^{th} order distributional derivative $D^\alpha T \in \mathcal{D}'(\Omega)$ is defined by

$$< D^\alpha T, \phi >= (-1)^{|\alpha|} < T, D^\alpha \phi >, \quad \forall \phi \in \mathcal{D}(\Omega).$$

Example 2.6. Let $\Omega = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 < 1\}$. Set radial function $f(r) = (\log \frac{1}{r})^k$, $k < \frac{1}{2}$, where $r = \sqrt{x_1^2 + x_2^2} < 1$. This function is not continuous at the origin, but we shall show that it is in $L^2(\Omega)$. Now using coordinate transformation (transforming to polar coordinates), we find that

$$\int_\Omega |f|^2 \, dx_1 dx_2 = \pi \int_0^1 |f(r)|^2 r \, dr.$$

This integral is finite, because $r(\log(\frac{1}{r}))^{2k} \mapsto 0$ as $r \mapsto 0$. Hence, it defines a distribution. Its partial distributional derivatives may be found out as $D_i f = \frac{\partial f}{\partial x_i} = -k(\log(\frac{1}{r}))^{k-1} \frac{x_i}{r^2}$ $i = 1, 2$.

The derivative map $D^\alpha : \mathcal{D}' \to \mathcal{D}'$ is continuous. To prove this, let $\{T_n\}$ a sequence in $\mathcal{D}'$ converge to T in $\mathcal{D}'$. Then, for $\phi \in \mathcal{D}$

$$\begin{aligned} < D^\alpha T_n, \phi > &= (-1)^{|\alpha|} < T_n, D^\alpha \phi > \\ &\to (-1)^{|\alpha|} < T, D^\alpha \phi > \\ &= < D^\alpha T, \phi > . \end{aligned}$$

Hence, the result follows.

2.3 Elements of Sobolev Spaces

We define the (Hilbert) Sobolev space $H^1(\Omega)$ as the set of all functions in $L^2(\Omega)$ such that all its first partial distributional derivatives are in $L^2(\Omega)$, i.e.,

$$H^1(\Omega) = \{v \in L^2(\Omega) : \frac{\partial v}{\partial x_i} \in L^2(\Omega), \ 1 \le i \le d\}.$$

Note that v being an $L^2(\Omega)$-function, all its partial distribuitional derivatives exist. But for $H^1(\Omega)$ what more we require is that $\frac{\partial v}{\partial x_i} \in L^2(\Omega)$, $1 \le i \le d$. This is, indeed, an extra condition and it is not true that every L^2-function has its first distributional partial derivatives in L^2. In the second part of Example 2.5, we have shown that $DH = 2\delta \notin L^2(-1, 1)$, where as $H \in L^2(-1, 1)$.

Let us define a map $(\cdot, \cdot)_1 : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ by

$$(u, v)_1 = (u, v) + \sum_{i=1}^d (\frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i}), \quad \forall u, v \in H^1(\Omega). \tag{2.1}$$

It is a matter of an easy exercise to check that $(\cdot, \cdot)_1$ forms an inner-product on $H^1(\Omega)$, and $(H^1(\Omega), (\cdot, \cdot)_1)$ is an inner-product space. The induced norm $\| \cdot \|_1$ on $H^1(\Omega)$ is set as

$$\|u\|_1 = \sqrt{(u, u)_1} = \sqrt{\|u\|^2 + \sum_{i=1}^d \|\frac{\partial u}{\partial x_i}\|^2}. \tag{2.2}$$

In the next theorem, we shall prove that $H^1(\Omega)$ is complete.

Theorem 2.1 *The space $H^1(\Omega)$ with $\|\cdot\|_1$ is a Hilbert space.*

Proof. Consider an arbitrary Cauchy sequence $\{u_n\}$ in $H^1(\Omega)$. We claim that it converges to an element in $H^1(\Omega)$. Given a Cauchy sequence $\{u_n\}$ in $H^1(\Omega)$, the $\{u_n\}$ and $\{\frac{\partial u_n}{\partial x_i}\}$, $i = 1, 2, \ldots, d$ are Cauchy in $L^2(\Omega)$. Since $L^2(\Omega)$ is complete, we can find u and u^i, $(i = 1, 2, \ldots, d)$ such that

$$u_n \to u \ \text{ in } \ L^2(\Omega),$$

$$\frac{\partial u_n}{\partial x_i} \to u^i, \ \ 1 \le i \le d, \ \text{ in } \ L^2(\Omega).$$

To complete the proof, it is enough to show that $u^i = \frac{\partial u}{\partial x_i}$. For $\phi \in \mathcal{D}(\Omega)$, consider

$$\begin{aligned} < \frac{\partial u_n}{\partial x_i}, \phi > &= - < u_n, \frac{\partial \phi}{\partial x_i} > = - \int_\Omega u_n \frac{\partial \phi}{\partial x_i} dx \\ &\to - \int_\Omega u \frac{\partial \phi}{\partial x_i} dx = - < u, \frac{\partial \phi}{\partial x_i} > \\ &= < \frac{\partial u}{\partial x_i}, \phi > . \end{aligned}$$

However,

$$\begin{aligned} < \frac{\partial u_n}{\partial x_i}, \phi > &= \int_\Omega \frac{\partial u_n}{\partial x_i} \phi dx \\ &\to \int_\Omega u^i \phi dx = < u^i, \phi > . \end{aligned}$$

Thus,

$$< u^i, \phi > = < \frac{\partial u}{\partial x_i}, \phi >, \quad \forall \phi \in \mathcal{D}(\Omega),$$

and we obtain

$$u^i = \frac{\partial u}{\partial x_i}.$$

Therefore,

$$u_n \to u \ \text{ in } \ L^2(\Omega) \quad \text{and} \quad \frac{\partial u_n}{\partial x_i} \to \frac{\partial u}{\partial x_i} \ \text{ in } \ L^2(\Omega),$$

i.e.,

$$u_n \to u \ \text{ in } \ H^1(\Omega).$$

Since $\{u_n\}$ is arbitrary, we have shown the completeness of $H^1(\Omega)$.

For positive integer m, we also define the higher order Sobolev spaces $H^m(\Omega)$ as

$$H^m(\Omega) := \{v \in L^2(\Omega) : D^\alpha v \in L^2(\Omega), |\alpha| \le m\}.$$

This is again a Hilbert space with respect to the inner product

$$(u,v)_m := \sum_{|\alpha|\leq m} (D^\alpha u, D^\alpha v),$$

and the induced norm

$$\|u\|_m := (\sum_{|\alpha|\leq m} \|D^\alpha u\|^2)^{1/2}.$$

In fact, it is a matter of an easy induction on m to show that $H^m(\Omega)$ is a complete innerproduct space.

Why should we restrict ourselves to L^2-setting alone, we can now define Sobolev spaces $W^{m,p}(\Omega)$ of order (m,p), $1 \leq p \leq \infty$ by

$$W^{m,p}(\Omega) := \{v \in L^p(\Omega) : D^\alpha v \in L^p(\Omega), |\alpha| \leq m\}.$$

This is a Banach space with norm

$$\|u\|_{m,p} := \left(\sum_{|\alpha|\leq m} \int_\Omega |D^\alpha u|^p \, dx \right)^{1/p}, \; 1 \leq p < \infty,$$

and for $p = \infty$

$$\|u\|_{m,\infty} := \max_{|\alpha|\leq m} \|D^\alpha u\|_{L^\infty(\Omega)}.$$

When $p = 2$, we shall call, for simplicity, $W^{m,2}(\Omega)$ as $H^m(\Omega)$ and its norm $\|\cdot\|_{m,2}$ as $\|\cdot\|_m$.

Exercise 2.2 *Show that $W^{1,p}(\Omega), (1 \leq p \leq \infty)$ is a Banach space and the using mathematical induction show that $W^{m,p}(\Omega)$ is also a Banach space.*

Some Properties of H^1- Space. Below, we shall discuss both negative and positive properties of H^1- Hilbert Sobolev space.
Negative Properties.

(i) For $d > 1$, the functions in $H^1(\Omega)$ may not be continuous.

In Example 2.6 , we have noted that the function $f(r)$ is not continuous at the origin, but it has distributional derivatives of all order. Now,

$$\frac{\partial f}{\partial x_i} = -k(\log(\frac{1}{r}))^{k-1} \frac{x_i}{r^2}, \; i = 1, 2,$$

and hence,

$$\begin{aligned} \int_\Omega |\frac{\partial f}{\partial x_1}|^2 \, dxdy &= 4 \int_0^{\pi/2} \int_0^1 k^2 \left(\log(\frac{1}{r}) \right)^{2(k-1)} \frac{1}{r} \cos^2\theta \, dr d\theta \\ &= \pi k^2 \int_0^1 r^{-1} \left(\log(\frac{1}{r}) \right)^{2(k-1)} dr. \end{aligned}$$

Here, we have used change of variables : $x_1 = r\cos\theta$, $x_2 = \sin\theta$, $0 \le \theta \le \pi$. It is easy to check that the integral is finite provided $k < \frac{1}{2}$ (only care is needed at $r = 0$). Therefore, $\frac{\partial f}{\partial x_1} \in L^2(\Omega)$. Similarly, it can be shown that $\frac{\partial f}{\partial x_2} \in L^2(\Omega)$ and hence, $f \in H^1(\Omega)$.

When $d = 1$, i.e., $\Omega \subset \mathbb{R}$, then every function in $H^1(\Omega)$ is absolutely continuous. This can be checked easily as

$$u(x) - u(x_0) = \int_{x_0}^{x} Du(\tau)\, d\tau, \;\; x_0, x_1 \in \Omega,$$

where Du is the distributional derivative of u.

(ii) The space $\mathcal{D}(\Omega)$ is not dense in $H^1(\Omega)$. To see this, we claim that the orthogonal complement of $\mathcal{D}(\Omega)$ in $H^1(\Omega)$ is not a trivial space. Let $u \in (\mathcal{D}(\Omega))^{\perp}$ in $H^1(\Omega)$ and let $\phi \in \mathcal{D}(\Omega)$. Then

$$\begin{aligned} 0 = (u,\phi)_1 &= (u,\phi) + \sum_{i=1}^{d} (\frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i}) \\ &= < u, \phi > - \sum_{i=1}^{d} < \frac{\partial^2 u}{\partial x_i^2}, \phi > \\ &= < -\Delta u + u, \phi >, \; \forall \phi \in \mathcal{D}(\Omega). \end{aligned}$$

Therefore,

$$-\Delta u + u = 0, \text{ in } \mathcal{D}'(\Omega).$$

We shall show that there are enough functions satisfying the above equation. Let Ω be a unit ball in $\mathbb{R}^d$. Consider $u = e^{r\cdot x}$ with $r \in \mathbb{R}^d$. Note that

$$\Delta u = |r|^2 e^{r\cdot x} = \begin{cases} |r|^2 u, \\ u, \qquad |r| = 1. \end{cases}$$

When $d = 1$, the solution u with $r = \pm 1$ belongs to $(\mathcal{D}(\Omega))^{\perp}$. For $d > 1$, there are infinitely many r's lying on the boundary of the unit ball in $\mathbb{R}^d$ for which $u \in (\mathcal{D}(\Omega))^{\perp}$. Therefore, the space $\mathcal{D}(\Omega)$ is not dense in $H^1(\Omega)$.

One then curious to know :

'*What is the closure of $\mathcal{D}(\Omega)$ in $H^1(\Omega)$-space ?*'

Then

'*Is it possible to characterize such a space?*'

Let $H_0^1(\Omega)$ be the closure of $\mathcal{D}(\Omega)$ in $H^1(\Omega)$-space. This is a closed subspace of $H^1(\Omega)$. The characterization of such a space will require the concept of values of the functions on the boundary or the restriction of functions to the boundary $\partial\Omega$. Note that the boundary $\partial\Omega$ being a $(d-1)$-dimensional manifold is a set of measure zero, and hence, for any

$v \in H^1(\Omega)$, its value at the boundary or its restriction to the boundary $\partial\Omega$ does not make sense.

When Ω is a bounded domain with Lipschitz continuous boundary $\partial\Omega$ (i.e., the boundary can be parametrized by a Lipschitz function), then the space $C^1(\overline{\Omega})$ (in fact $C^\infty(\overline{\Omega})$) is dense in $H^1(\Omega)$. The proof, we shall not pursue it here, but refer to Adams [1]. For $v \in C^1(\overline{\Omega})$, its restriction to the boundary or value at the boundary is meaningful as $v \in C(\overline{\Omega})$. Since $C^1(\overline{\Omega})$ is dense in $H^1(\Omega)$, for every $v \in H^1(\Omega)$ there is a sequence $\{v_n\}$ in $C^1(\overline{\Omega})$ such that $v_n \mapsto v$ in $H^1(\Omega)$. Thus, for each v_n, its restriction to the boundary $\partial\Omega$ that is $v_n|_{\partial\Omega}$ makes sense and further, $v_n|_{\partial\Omega} \in L^2(\partial\Omega)$. This sequence, infact, converges in $L^2(\partial\Omega)$ to an element say $\gamma_0 v$. This is what we call the trace of v. More precisely, we define the trace $\gamma_0 : H^1(\Omega) \mapsto L^2(\partial\Omega)$ as a continuous linear map satisfying

$$\|\gamma_0 v\|_{L^2(\partial\Omega)} \leq C\|v\|_1,$$

and for $v \in C^1(\overline{\Omega})$, $\gamma_0 v = v|_{\partial\Omega}$. This is, indeed, a generalization of the concept of the values of H^1-functions on the boundary or the restriction to the boundary. The range of γ_0 called $H^{1/2}(\partial\Omega)$ is a proper and dense subspace of $L^2(\partial\Omega)$. The norm on this space is defined as

$$\|g\|_{1/2,\partial\Omega} = \inf\{\|v\|_1 : v \in H^1(\Omega) \text{ and } \gamma_0 v = g\}.$$

Now the space $H_0^1(\Omega)$ is characterized by

$$H_0^1(\Omega) := \{v \in H^1(\Omega) : \gamma_0 v = 0\}.$$

For a proof, again we refer to Adams [1]. Sometimes, by abuse of language, we call $\gamma_0 v = 0$ as simply $v = 0$ on $\partial\Omega$.

Positive Properties of H_0^1 and H^1- Spaces. We shall see subsequently that the denseness of $\mathcal{D}(\Omega)$ and $C^\infty(\overline{\Omega})$, respectively in $H_0^1(\Omega)$ and $H^1(\Omega)$ with respect to H^1-norm will play a vital role. Essentially, many important results are first derived for smooth functions and then using the denseness property are carried over to H_0^1 or H^1-spaces.

In the following, we shall prove some inequalities. If $v \in V := \{w \in C^1[a,b] : w(a) = w(b) = 0\}$, then for $x \in [a,b]$

$$v(x) = \int_a^x \frac{dv}{dx}(s)\, ds.$$

A use of Hölder's inequality yields

$$|v(x)| \leq (x-a)^{1/2}\Big(\int_a^b |\frac{dv}{dx}|^2\, ds\Big)^{1/2}.$$

On squaring and again integrating over x from a to b, it follows that

$$\begin{aligned}\int_a^b |v(x)|^2\,dx &\leq (\int_a^b (x-a)\,dx)(\int_a^b |\frac{dv}{dx}|^2\,ds)\\ &\leq \frac{(b-a)^2}{2}\int_a^b |\frac{dv}{dx}|^2\,dx.\end{aligned}$$

Thus,

$$\|v\|_{L^2(a,b)} \leq C(a,b)\|v'\|_{L^2(a,b)},$$

where the constant $C(a,b) = \frac{(b-a)}{\sqrt{2}}$. This is one dimensional version of Poincaré inequality. We note that in the above result only $v(a) = 0$ is used. The same conclusion is also true when $v(b) = 0$. Further, it is easy to show that

$$\max_{x\in[a,b]} |v(x)| \leq (b-a)^{1/2}\|v'\|_{L^2(a,b)}.$$

As a consequence of the Poincaré inequality, $\|v'\|_{L^2(a,b)}$ defines a norm on V which is equivalent to $H^1(a,b)$-norm, i.e., there are positive constants say C_1 and C_2 such that

$$C_1\|v\|_1 \leq \|v'\|_{L^2(a,b)} \leq C_2\|v\|_1,$$

Where $C_2 = 1$ and $C_1 = (\frac{(b-a)^2}{2} + 1)^{-1/2}$.

We shall now generalize this to the functions in $H_0^1(\Omega)$, when Ω is a bounded domain in IR^d.

Theorem 2.3 *(Poincaré Inequality). Let Ω be a bounded domain in IR^d. Then there exists constant $C(\Omega)$ such that*

$$\int_\Omega |v|^2\,dx \leq C(\Omega)\int_\Omega |\nabla v|^2\,dx,\ \ \forall v \in H_0^1(\Omega).$$

Proof. Since $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$, it is enough to prove the above inequality for the functions in $\mathcal{D}(\Omega)$. Now because of denseness, for every $v \in H_0^1(\Omega)$, there is a sequence $\{v_n\}$ in $\mathcal{D}(\Omega)$ such that $v_n \mapsto v$ in H^1-norm. Suppose for every v_n in $\mathcal{D}(\Omega)$, the above result holds good,i.e.,

$$\int_\Omega |v_n|^2\,dx \leq C(\Omega)\int_\Omega |\nabla v_n|^2\,dx,\ \ v_n \in \mathcal{D}(\Omega), \tag{2.3}$$

then taking limit as $n \mapsto \infty$, we have $\|v_n\| \mapsto \|v\|$ as well as $\|\nabla v_n\| \mapsto \|\nabla v\|$, and hence, the result follows for $v \in H_0^1(\Omega)$.

It remains to prove (2.3). Since Ω is bounded, enclose it by a box in $\mathbb{R}^{\mathrm{d}}$, i.e., say $\Omega \subset \Pi_{i=1}^d [a_i, b_i]$. Then extend each v_n even alongwith its first partial derivatives to zero outside of Ω. This is possible as $v_n \in \mathcal{D}(\Omega)$. Now, we have (like in one dimensional case) for $x = (x_1, \cdots, x_i, \cdots, x_d)$

$$v_n(x) = \int_{a_i}^{x_i} \frac{\partial v_n}{\partial x_i}(x_1, \cdots, x_{i-1}, s, x_{i+1}, \cdots, x_d)\,ds.$$

Then using Hölder's inequality

$$|v_n(x)| \leq (x_i - a_i)^{1/2} \left(\int_{a_i}^{b_i} |\frac{\partial v_n}{\partial x_i}|^2 \, ds \right)^{1/2}.$$

Squaring and then integrating both sides over the box, it is found that

$$\int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} |v_n(x)| \, dx \leq \frac{(b_i - a_i)^2}{2} \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} |\frac{\partial v_n}{\partial x_i}|^2 \, dx.$$

Therefore the estimate (2.3) follows with the constant $C = \frac{1}{2}(b_i - a_i)^2$. This completes the rest of the proof.

Remarks. The above theorem is not true if $v \in H^1(\Omega)$ (say, when $\Omega = (0,1)$ and $v = 1$, $v \in H^1(\Omega)$ and it does not satisfy Poincaré ineguality) or if Ω is not bounded. But when Ω is within a strip , the result still holds (the proof of theorem only uses that the domain is bounded in one direction that is in the direction of x_i). The best constant $C(\Omega)$ can be $\frac{1}{\lambda_{min}}$, where λ_{min} is the minimum positive eigenvalue of the following Dirichlet problem

$$-\Delta u = \lambda u, \ x \in \Omega$$

with Dirichlet boundary condition

$$u = 0, \text{ on } \partial\Omega.$$

Using the Poincaré inequality , it is easy to arrive at the following result.

Corollary 2.4 *For $v \in H_0^1(\Omega)$, $\|\nabla v\|$ is, infact, a norm on $H_0^1(\Omega)$-space which is equivalent to H^1-norm.*

Using the concept of trace and the denseness property, we can generalize the Green's formula to the functions in Sobolev spaces. For details , see Kesavan [9]

Lemma 2.5 *Let u and v be in $H^1(\Omega)$. Then for $1 \leq i \leq d$, the following integration by parts formula holds*

$$\int_\Omega u \frac{\partial v}{\partial x_i} \, dx = -\int_\Omega \frac{\partial u}{\partial x_i} v \, dx + \int_{\partial\Omega} uv\nu_i \, ds, \tag{2.4}$$

where ν_i is the ith component of the outward normal ν to the boundary $\partial\Omega$.

Further, if $u \in H^2(\Omega)$, then the following Green's formula holds

$$\sum_{i=1}^{d} \int_\Omega \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx = -\sum_{i=1}^{d} \int_\Omega \frac{\partial^2 u}{\partial x_i^2} v \, dx + \sum_{i=1}^{d} \int_{\partial\Omega} \nu_i \gamma_0(\frac{\partial u}{\partial x_i}) \gamma_0(v) \, ds,$$

or in usual vector notation

$$(\nabla u, \nabla v) = (-\Delta u, v) + \int_{\partial\Omega} \frac{\partial u}{\partial \nu} v \, ds. \tag{2.5}$$

We now state another useful property of H^1-space. For a proof, we may refer to Adams [1].

Theorem 2.6 *(Imbedding Theorem). Let Ω be a bounded domain with Lipschitz boundary $\partial\Omega$. Then $H^1(\Omega)$ is continuously imbedded in $L^p(\Omega)$ for $1 \le p < q$ if $d = 2$ or $p \le q$, for $d \ge 2$, where $\frac{1}{q} = \frac{1}{2} - \frac{1}{d}$. Further, the above imbedding is compact provided $p < q$.*

Dual Spaces of H^1 and H_0^1. We denote by $H^{-1}(\Omega)$, the dual space of $H_0^1(\Omega)$, i.e., it consists of all continuous linear functionals on $H_0^1(\Omega)$. The norm on $H^{-1}(\Omega)$ is given by the standard dual norm and now is denoted by

$$\|f\|_{-1} = \sup_{0 \neq v \in H_0^1(\Omega)} \frac{| < f, v > |}{\|v\|_1}. \tag{2.6}$$

The following Lemma characterizes the elements in the dual space $H^{-1}(\Omega)$.

Lemma 2.7 *A distribution $f \in H^{-1}(\Omega)$ if and only if there are functions $f_\alpha \in L^2(\Omega)$ such that*

$$f = \sum_{|\alpha| \le 1} D^\alpha f_\alpha. \tag{2.7}$$

For an example, the Dirac delta function δ belongs to $H^{-1}(-1, 1)$, because there exists Heaviside step function

$$H(x) = \begin{cases} 1, & 0 < x < 1, \\ 0, & -1 < x \le 0 \end{cases}$$

in $L^2(-1, 1)$ such that $\delta = DH$.

Note that

$$H_0^1(\Omega) \subset L^2(\Omega) \subset H^{-1}(\Omega).$$

In fact the imbedding is continuous in each case. This forms a Gelfand triplet. For $f \in L^2(\Omega)$ and $v \in H_0^1(\Omega)$, we have the following identification

$$< f, v >= (f, v), \tag{2.8}$$

where $< \cdot, \cdot >$ is a duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Similarly, we can define the dual space $(H^1(\Omega))'$ of $H^1(\Omega)$. Its norm is again given through the dual norm and is defined by

$$\|f\|_{(H^1(\Omega))'} = \sup_{0 \neq v \in H^1(\Omega)} \frac{| < f, v > |}{\|v\|_1}. \tag{2.9}$$

In the present case, we may not have a characterization like the one in the previous Lemma.

Remark. For positive integer m, we may define $H_0^m(\Omega)$ as the closure of $\mathcal{D}(\Omega)$ with respect to the norm $\|\cdot\|_m$. For smooth boundary Ω, we can characterize the space H_0^2 by

$$H_0^2(\Omega) := \{v \in H^2(\Omega) : \gamma_0 v = 0,\ \gamma_0(\frac{\partial v}{\partial \nu}) = 0\}. \tag{2.10}$$

Similarly, the dual space $H^{-m}(\Omega)$ is defined as the set of all distributions f such that there are functions $f_\alpha \in L^2(\Omega)$ with $f = \sum_{|\alpha|\leq m} D^\alpha f_\alpha$. The norm is given by

$$\|f\|_{-m} = \sup_{0\neq v\in H_0^m(\Omega)} \frac{|<f, v>|}{\|v\|_m}. \tag{2.11}$$

Note. There are some exellent books available on the theory of Distribution and Sobolev Spaces like Adams [1], Dautray and Lions [5]. In this chapter, we have discussed only the basic rudiment theory which is barerly minimum for the development of the subsequent chapters.

Chapter 3

Abstract Elliptic Theory

We present in this chapter a brief account of the fundamental tools which are used in the study of linear elliptic partial differential equations. To motivate the materials to come, we again start with the mathematical model of vibration of drum: Find u such that

$$-\Delta u = f \quad \text{in} \quad \Omega \tag{3.1}$$

$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{3.2}$$

where $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a bounded domain with smooth boundary $\partial\Omega$ and f is a given function.

Multiply (3.1) by a nice function v which vanishes on $\partial\Omega$ (say $v \in \mathcal{D}(\Omega)$) and then integrate over Ω to obtain

$$\int_\Omega (-\Delta u) v dx = \int_\Omega f v dx.$$

Formally, a use of Gauss divergence theorem for the integral on the left hand side yields

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

Since $v = 0$ on $\partial\Omega$, the first term on the right hand side vanishes and hence, we obtain

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega f v dx.$$

Therefore, the problem (3.1) is reformulated as: Find a function $u \in V$ such that

$$a(u, v) = L(v), \quad \forall v \in V, \tag{3.3}$$

where $a(u, v) = \int_\Omega \nabla u \cdot \nabla v dx$, and $L(v) = \int_\Omega f v dx$. Here, the space V (the space of all admissable displacements for the problem (1.1)) has to be chosen approprietely. For the present case, one thing is quite apparent that every element of V should vanish on $\partial\Omega$.

How does one choose V ? The choice of V should be such that the integrals in (3.3) are finite. For $f \in L^2(\Omega)$, space V may be chosen in such a way that $\nabla u \in L^2(\Omega)$ and $u \in L^2(\Omega)$. The largest such space satisfying the above conditions and $u = 0$ on $\partial\Omega$ is, in fact, the space $H_0^1(\Omega)$. Thus, we choose $V = H_0^1(\Omega)$. Since $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$, the problem (3.3) makes sense for all $v \in V$.

Now, observe that every (classical) solution of (3.1) – (3.2) satisfies (3.3). However, the converse need not be true. Then, one is curious to know whether, in some weaker sense, the solution of (3.3) does satisfy (3.1) – (3.2) . To see this, note that

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega f v dx, \quad \forall v \in \mathcal{D}(\Omega) \subset H_0^1(\Omega).$$

Using the definition of distributional derivatives, it follows that

$$\begin{aligned} \int_\Omega \nabla u \cdot \nabla v \, dx &= \sum_{i=1}^{d} \int_\Omega \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_i} \, dx \\ &= \sum_{i=1}^{d} < \frac{\partial u}{\partial x_i}, \frac{\partial v}{\partial x_i} > = - \sum_{i=1}^{d} < \frac{\partial^2 u}{\partial x_i^2}, v > \\ &= < -\Delta u, v > \ \forall v \in \mathcal{D}(\Omega), \end{aligned}$$

and hence,

$$< -\Delta u, v > = < f, v > \quad \forall v \in \mathcal{D}(\Omega).$$

Obviously,

$$-\Delta u = f \quad \text{in} \ \ \mathcal{D}'(\Omega). \tag{3.4}$$

Thus, the solution of (3.3) satisfies (3.1) in the sense of distribution, i.e., in the sense of (3.4) .

Conversely, if u satisfies (3.4), then following the above steps backward, we obtain

$$a(u, v) = L(v) \ \forall v \in \mathcal{D}(\Omega).$$

As $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$, the above equation holds for $v \in H_0^1(\Omega)$ and hence, u is a solution of (3.3).

Since most of the linear elliptic equations can be recast in the abstract form (3.3), we examine, below, some sufficient conditions on $a(\cdot, \cdot)$ and L so that the problem (3.3) is wellposed in the sense of Hadamard.

3.1 Abstract Variational Formulation

Let V and H be real Hilbert spaces with $V \hookrightarrow H = H' \hookrightarrow V'$ where V' is the dual space of V. Given a continuous linear functional L on V and a bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$, the abstract variational problem is to seek a function $u \in V$ such that

$$a(u, v) = L(v) \ \ \forall v \in V. \tag{3.5}$$

In the following Theorem, we shall prove the wellposedness of (3.5). This theorem is due to P.D. Lax and A.N. Milgram (1954) and hence, it is called Lax-Milgram Theorem.

Theorem 3.1 *Assume that the bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is bounded and coercive, i.e., there exist two constants M and $\alpha_0 > 0$ such that*

$$\text{(i)} \qquad |a(u,v)| \le M\|u\|_V\|v\|_V \quad \forall u, v \in V$$

and

$$\text{(ii)} \qquad a(u,v) \ge \alpha_0\|v\|_V^2 \quad \forall v \in V.$$

Then, for a given $L \in V'$ the problem (3.5) has one and only one solution u in V. Moreover,

$$\|u\|_V \le \frac{1}{\alpha_0}\|L\|_{V'}, \tag{3.6}$$

i.e., the solution u depends continuously on the given data L.

Remark: In addition if we assume $a(\cdot,\cdot)$ is symmetric, then as a consequence of Riesz-Representation Theorem[1], it is easy to show the wellposedness of (3.5). Since $a(\cdot,\cdot)$ is symmetric, define an innerproduct on V as $((u,v)) = a(u,v)$. The introduced norm is $|||u||| = ((u,u))^{\frac{1}{2}} = (a(u,u))^{\frac{1}{2}}$. The space $(V, |||\cdot|||)$ is a Hilbert space. Due to boundedness and coercivity of $a(\cdot,\cdot)$, we have

$$\alpha_0\|v\|_V^2 \le ((u,v)) = |||v|||^2 \le M\|v\|_V^2.$$

Hence, the two norms are equivalent. Thus, both these spaces have the same dual V'. Then the problem (3.5) can be reformulated in this new setting as: Given $L \in V'$, find $u \in (V, |||\cdot|||)$ such that

$$((u,v)) = L(v) \quad \forall v \in V. \tag{3.7}$$

Now, an application of Riesz-Representation Theorem ensures the existence of a unique element $u \in (V, |||\cdot|||)$ such that (3.7) is satisfied. Further,

$$\sqrt{\alpha_0}\|u\|_V \le |||u||| = \|L\|_{V'}$$

and hence, the result follows.

When $a(\cdot,\cdot)$ is not symmetric, we prove the result as follows:

Proof of Theorem 3.1. We shall first prove the estimate (3.6). Using coercivity, the property (ii) of the bilinear form, we have

$$\alpha_0\|u\|_V^2 \le a(u,u) = |L(u)| \le \|L\|_{V'}\|u\|_V.$$

Therefore, the result (3.6) follows. For continuous dependence, let $L_n \to L$ in V' and let u_n be the solution of (3.5) corresponding to L_n. Then, using linearity property,

$$a(u_n - u, v) = (L_n - L)(v).$$

[1]For every continuous linear functional l on a Hilbert space V, there exists one and only one element $u \in V$ such that $l(v) = (u,v)_V$, $\forall v \in V$ and $\|u\|_V = \|l\|_{V'}$.

Now, an application of the estimate (3.6) yields

$$\|u_n - u\|_V \leq \frac{1}{\alpha_0}\|L_n - L\|_{V'}.$$

Hence, $L_n \to L$ in V' implies $u_n \to u$ in V, and this completes the proof of the continuous dependence of u on L. Uniqueness is a straight forward consequence of the continuous dependence.

For existence, we shall put (3.5) in an abstract operator theoretic form and then apply Banach contraction mapping theorem [2]

For $w \in V$, define a functional l_w by $l_w = a(w, v)\ \forall v \in V$. Note that l_w is linear functional on V. Using boundedness of the bilinear form, we obtain

$$|l_w(v)| = |a(w, v)| \leq M\|w\|_V\|v\|_V,$$

and hence,

$$\|l_w\|_{V'} = \sup_{v\in V}\frac{|l_w(v)|}{\|v\|_V} = \sup_{v\in V}\frac{|a(w, v)|}{\|v\|_V} \leq M\|w\|_V < \infty.$$

Thus, $l_w \in V'$. Since $w \mapsto l_w$ is a linear map, setting $A : V \to V'$ by $Aw = l_w$, we have

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

Indeed, $w \mapsto Aw$ is a linear map and A is continuous with $\|Aw\|_V \leq M\|w\|_V$. Let $\tau : V' \mapsto V$ denote the Riesz mapping such that

$$\forall f \in V',\ \forall v \in V,\ f(v) = (\tau f, v)_V,$$

where $(\cdot, \cdot)_V$ is the innerproduct in V. The variational problem (3.5) is now equivalent to the operator equation

$$\tau Au = \tau L \quad \text{in } V. \tag{3.8}$$

Thus, for the wellposedness of (3.5), it is sufficient to show the wellposedness of (3.8). For $\rho > 0$, define a map $\mathcal{J} : V \to V$ by

$$\mathcal{J}v = v - \rho\,(\tau Av - \tau L) \quad \forall v \in V. \tag{3.9}$$

Suppose $\mathcal{J}$ in (3.9) is a contraction, then by contraction mapping theorem there is a unique element $u \in V$ such that $\mathcal{J}u = u$, i.e., $u - \rho\,(\tau Au - L) = u$ and u satisfies (3.8). To complete the proof, we now show that $\mathcal{J}$ is a contraction. For $v_1, v_2 \in V$, we note that

$$\begin{aligned}
\|\mathcal{J}v_1 - v_2\|_V^2 &= \|v_1 - v_2 - \rho\,(\tau Av_1 - \tau Av_2)\,\|_V^2 \\
&\leq \|v_1 - v_2\|_V^2 - 2\rho a(v_1 - v_2, v_1 - v_2) + \rho^2 a(v_1 - v_2, \tau A(v_1 - v_2)) \\
&\leq \|v_1 - v_2\|_V^2 - 2\rho\alpha_0\|v_1 - v_2\|_V^2 + \rho^2 M^2\|v_1 - v_2\|_V^2.
\end{aligned}$$

[2] A map $\mathcal{J} : V \to V$ (V can be taken even as a Banach space or even a metric space) is said to be a contraction if there is a constant $0 \leq \beta < 1$ such that $\|\mathcal{J}u - \mathcal{J}v\|_V \leq \beta\|u - v\|_V$ for $u, v \in V$. Every contraction mapping has a unique fixed point, i.e., $\exists!\ u_0 \in V$ such that $\mathcal{J}u_0 = u_0$.

Here, we have used the coercive property (ii) and boundedness property of A. Altogether,

$$\|\mathcal{J}v_1 - \mathcal{J}v_2\|_V^2 \leq (1 - 2\rho\alpha_0 + \rho^2 M^2)\|v_1 - v_2\|_V^2,$$

and hence,

$$\|\mathcal{J}w_1 - \mathcal{J}w_2\|_V \leq \sqrt{(1 - 2\rho\alpha_0 + \rho^2 M^2)}\|w_1 - w_2\|_V.$$

With a choice of $\rho \in (0, \frac{\alpha_0}{2M})$, we find that $\mathcal{J}$ is a contraction and this completes the proof.

Remark. There are also other proofs that are available in the literature. One depends on continuation argument by looking at the deformation from the symmetric case and other depends on the existence of a solution to an abstract operator equation. However, the present existential proof gives an algorithm to compute the solution as a fixed point of $\mathcal{J}$. Based on Faedo-Galerkin method, another proof is presented in Temam [11] (see, pages 24–27).

Corollary 3.2 *When the bilinear form $a(\cdot,\cdot)$ in symmetric, i.e., $a(u,v) = a(v,u)$, $u, v \in V$, and coercive, then the solution u of (3.5) is also the only element of V that minimizes the following energy functional J on V, i.e.,*

$$J(u) = \min_{v \in V} J(v), \tag{3.10}$$

where $J(v) = \frac{1}{2}a(v,v) - L(v)$. Moreover, the converse is also true.

Proof. Let u be a solution of (3.5). Write $v = u + w$, for some vector $w \in V$. Note that

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

Using definition of J and symmetric property of $a(\cdot,\cdot)$, we have

$$J(v) = J(u) + (a(u,w) - L(w)) + \frac{1}{2}a(w,w) \geq J(u).$$

Here, we have used the fact that u is a solution of (3.5) and coercivity (ii) of the bilinear form, Since v is arbitrary (it can be done by varying w), u satisfies (3.10).

To prove the converse (direct proof), let u be a solution of (3.10). Define a scalar function for $t \in \mathbb{R}$ as

$$\Phi(t) = J(u + tv).$$

Indeed, $\Phi(t)$ has a minimum at $t = 0$. If $\Phi(t)$ is differentiable, then $\Phi(t)'|_{t=0} = 0$. Note that

$$\begin{aligned} \Phi(t) &= \frac{1}{2}a(u+tv, u+tv) - L(u+tv) \\ &= \frac{1}{2}a(u,u) + \frac{t}{2}(a(u,v) + a(v,u)) + \frac{t^2}{2}a(v,v) - L(u) - tL(v). \end{aligned}$$

This is a quadratic functional in t and hence, differentiable. Using symmetric property of $a(\cdot,\cdot)$, the derivative of Φ is

$$\Phi'(t) = a(u,v) - L(v) + ta(v,v),\ \forall t.$$

Now $\Phi'(0) = 0$ implies $a(u,v) = L(v)\ \forall v \in V$, and hence, the result follows.

3.2 Some Examples.

In this section, we discuss some examples of linear elliptic partial differential equations and derive wellposedness of the weak solutions using Lax-Milgram Theorem. In all the following examples, we assume that Ω is a bounded domain with Lipschitz continuous boundary $\partial\Omega$.

Example 3.3 *(Homogeneous Dirichlet Boundary Value Problem Revisited)*

For the problem (3.1)-(3.2), the weak formulation is given by (3.3) in $V = H_0^1(\Omega)$ for wellposedness, the bilinear form $a(\cdot,\cdot)$ satisfies both the properties:

$$a(v,v) = \int_\Omega |\nabla v|^2\, dx = \|v\|_V^2,$$

and

$$|a(u,v)| = |\int_\Omega \nabla u \cdot \nabla v\, dx| \leq \|u\|_V \|v\|_V.$$

The functional L is linear and using the identification (2.9) in Chapter 2, we have

$$L(v) = (f,v) =< f, v >\ \ \forall v \in V.$$

Hence, using Poincaré inequality it follows that

$$\|L\|_{V'} = \sup_{0 \neq v \in V} \frac{|L(v)|}{\|v\|_V} \leq C\|f\|.$$

An application of the Lax-Milgram Theorem yields the wellposedness of (3.3). Moreover,

$$\|u\|_1 \leq C\|f\|.$$

Since, $f \in L^2(\Omega)$, it is straight forward to check that $\|\Delta u\| \leq \|f\|$ and the following regularity result can be shown to hold

$$\|u\|_2 \leq C\|f\|.$$

For a proof, see , Kesavan [9].

Since, the bilinear form is symmetric, by Corollary 3.2 the problem (3.3) is equivalent to the following minimization problem:

$$J(u) = \min_{v \in H_0^1(\Omega)} J(v),$$

where

$$J(v) = \frac{1}{2}\int_\Omega |\nabla v|^2\, dx - \int_\Omega fv\, dx.$$

Example 3.4 *(Nonhomogeneous Dirichlet Problem)*
Given $f \in H^{-1}(\Omega)$ and $g \in H^{\frac{1}{2}}(\partial\Omega)$, find a function u satisfying

$$-\Delta u = f, \quad x \in \Omega \tag{3.11}$$

with non-homogeneous Dirchlet boundary condition

$$u = g, \qquad x \in \partial\Omega.$$

To put (3.11) in the abstract variational form, the choice of V is not difficult to guess. Now since $H^{\frac{1}{2}}(\partial\Omega) =$ Range of γ_0, where $\gamma_0 : H^1(\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$, then for $g \in H^{\frac{1}{2}}(\partial\Omega)$, there is an element say $u_0 \in H^1(\Omega)$ such that $\gamma_0 u_0 = g$. Therefore, we recast the problem as: Find $u \in H^1(\Omega)$ such that

$$u - u_0 \in H_0^1(\Omega), \tag{3.12}$$

$$a(u - u_0, v) =< f, v > -a(u_0, v) \ \forall v \in H_0^1(\Omega), \tag{3.13}$$

where $a(v, w) = \int_\Omega \nabla v \cdot \nabla w dx$. We obtain (3.13) by using Gauss divergence Theorem.

To check the conditions in the Lax-Milgram Theorem, note that $V = H_0^1(\Omega)$, and the bilinear form $a(\cdot, \cdot)$ satisfies

$$|a(u, v)| \leq \int_\Omega |\nabla u||\nabla v| \, dx \leq \|\nabla u\|\|\nabla v\| = \|u\|_V \|v\|_V.$$

In $H_0^1(\Omega)$, $\|\nabla u\|$ is in fact a norm (see Corollary 2.1) and is equivalent to $H^1(\Omega)-$norm. For coercivity,

$$a(u, v) = \int_\Omega |\nabla u|^2 \, dx = \|u\|_V^2.$$

Since $a(\cdot, \cdot)$ is continuous, the mapping $L : v \to < f, v > -a(u_0, v)$ is a continuous linear map on $H_0^1(\Omega)$ and hence, it belongs to $H^{-1}(\Omega)$. An application of Lax-Milgram Theorem yields the wellposedness of (3.12)-(3.13). Further,

$$\|u\|_V - \|u_0\|_V \leq \|u - u_0\|_V = \|\nabla(u - u_0)\| \leq \|L\|_{H^{-1}(\Omega)} \leq \|f\|_{-1} + \|\nabla u_0\|.$$

Therefore, using the equivalence of the norms on $H_0^1(\Omega)$ and $H^1(\Omega)$, we obtain

$$\|u\|_1 \leq c(\|f\|_{-1} + \|u_0\|_1),$$

for all $u_0 \in H^1(\Omega)$ such that $\gamma_0 u_0 = g$. However, taking infimum over $u_0 \in H^1(\Omega)$ for which $\gamma_0 u_0 = g$, it follows that

$$\|u\|_1 \leq C(\|f\|_{-1} + \inf_{u_0 \in H^1(\Omega)} \{\|u_0\|_1 : \gamma_0 u_0 = g\}) = C(\|f\|_{-1} + \|g\|_{\frac{1}{2},\partial\Omega}),$$

(using definition of $H^{\frac{1}{2}}(\partial\Omega)$). It remains to check that *in what sense u satisfies (3.10)-(3.11)?* Choose $v \in \mathcal{D}(\Omega)$ in (3.13). This is possible as $\mathcal{D}(\Omega) \subset H_0^1(\Omega)$. Then

$$a(u,v) = - < \Delta u, v > = < f, v > \quad \forall v \in \mathcal{D}(\Omega).$$

Since $f \in H^{-1}(\Omega)$, $\Delta u \in H^{-1}(\Omega)$ and hence, u satisfies

$$u - u_0 \in H_0^1(\Omega), \tag{3.14}$$

$$-\Delta u = f, \quad \text{in} \quad H^{-1}(\Omega). \tag{3.15}$$

Converse part is straight forward as $\mathcal{D}(\Omega)$ is dense in $H_0^1(\Omega)$. However $u - u_0 \in H_0^1(\Omega)$ if only if $\gamma_0 u_0 = g$ and hence, the non-homogeneous boundary condition is satisfied.

Remark. It is even possible to show the following higher regularity result provided $f \in L^2(\Omega)$ and $g \in H^{\frac{3}{2}}(\partial\Omega)$:

$$\|u\|_2 \leq c(\|f\| + \|g\|_{\frac{3}{2},\partial\Omega}),$$

where

$$\|g\|_{\frac{3}{2},\partial\Omega} = \inf_{v \in H^2(\Omega)} \{\|v\|_2 : \gamma_0 v = g\}.$$

Since $a(\cdot,\cdot)$ is symmetric, $u - u_0$ will minimize $J(v)$ over $H_0^1(\Omega)$, where

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

Example 3.5 *(Neumann Problem): Find u such that*

$$-\Delta u + cu = f \quad \text{in} \quad \Omega, \tag{3.16}$$

$$\frac{\partial u}{\partial \nu} = g \quad \text{on} \quad \partial\Omega. \tag{3.17}$$

Here assume that $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$ and $c > 0$.

To find out the space V, the bilinear form and the linear functional, we multiply (3.16)-(3.17) by a smooth function v and apply Green's Formula:

$$\int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \frac{\partial u}{\partial \nu} v \, ds + \int_\Omega cuv \, dx = \int_\Omega fv \, dx.$$

Using Neumann boundary condition

$$\int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega cuv \, dx = \int_\Omega fv \, dx + \int_{\partial\Omega} gv \, ds. \tag{3.18}$$

Therefore, set

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + cuv) \, dx,$$

and

$$L(v) = \int_{\partial\Omega} gv\, ds + \int_{\Omega} fv\, dx.$$

The largest space for which these integrals are finite is $H^1(\Omega)$-space. So we choose $V = H^1(\Omega)$. To verify the sufficient conditions in the Lax-Milgram Theorem, note that the bilinear form $a(\cdot,\cdot)$ satisfies:

(i) the boundedness property as

$$\begin{aligned} |a(u,v)| &= |\int_{\Omega} (\nabla u \cdot \nabla v + cuv)\, dx| \\ &\leq \|\nabla u\|\|\nabla v\| + \max_{x\in\overline{\Omega}} c(x)\|u\|\|v\| \\ &\leq \max(1, \max_{x\in\overline{\Omega}} c(x))\|u\|_1\|v\|_1. \end{aligned}$$

(ii) the coercive condition, since

$$a(v,v) = \int_{\Omega} (\|\nabla v\|^2 + c|v|^2)\, dx \geq \min(1,c)\|v\|_1^2 = \alpha\|v\|_1^2.$$

Further, L is a linear functional and it satisfies

$$\begin{aligned} |L(v)| &\leq \int_{\partial\Omega} |g||v|\, ds + \int_{\Omega} |f||v|\, dx \\ &\leq \|g\|_{L^2(\partial\Omega)}\|\gamma_0 v\|_{L^2(\partial\Omega)} + \|f\|\|v\|. \end{aligned}$$

As $\|v\| \leq \|v\|_1$ and from the trace result $\|\gamma_0 v\|_{L^2(\partial\Omega)} \leq C\|v\|_1$, we obtain

$$\|L\|_{(H^1)'} = \sup_{0\neq v\in H^1(\Omega)} \frac{|L(v)|}{\|v\|_1} \leq C(\|f\| + \|g\|_{L^2(\partial\Omega)}).$$

Therefore, L is a continuous linear functional on $H^1(\Omega)$. Apply the Lax-Milgram Theorem to obtain the wellposedness of the weak solution of (3.18). Moreover, the following regularity also holds true

$$\|u\|_1 \leq C(\|f\| + \|g\|_{L^2(\partial\Omega)}).$$

Choose $v \in \mathcal{D}(\Omega)$ as $\mathcal{D}(\Omega) \subset H^1(\Omega)$ and using Green's formula, we find that

$$a(u,v) =< -\Delta u + cu, v >,$$

and hence,

$$< -\Delta u + cu, v >=< f, v >,\ \forall \mathcal{D}(\Omega).$$

Therefore, the weak solution u satisfies the equation

$$-\Delta u + cu = f,\ \text{ in } \mathcal{D}'(\Omega).$$

Since $f \in L^2(\Omega)$, $-\Delta u + cu \in L^2(\Omega)$. Further, it can be shown that $u \in H^2(\Omega)$.

For the boundary condition, we note that for $v \in H^1(\Omega)$

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

Since $u \in H^2(\Omega)$, we obtain

$$\int_\Omega (\Delta u + cu)v dx = \int_\Omega f v dx, \quad \forall v \in H^1(\Omega).$$

Using Green's Theorem, it follows that

$$\int_\Omega (\nabla u \nabla v + cuv) \, dx = \int_{\partial\Omega} \frac{\partial u}{\partial \nu} v \, ds + \int_\Omega f v \, dx.$$

Comparing this with the weak formulation (3.18), we find that

$$\int_{\partial\Omega} (g - \frac{\partial u}{\partial \nu} v) \, ds = 0, \quad \forall v \in H^1(\Omega).$$

Hence,

$$\frac{\partial u}{\partial \nu} = g \ \text{ on } \ \partial\Omega.$$

Remark: If $g = 0$, i.e., $\frac{\partial u}{\partial \nu} = 0$ on $\partial\Omega$, **CAN** we impose this condition on $H^1(\Omega)$-space as we have done in case of the homogeneous Dirichlet problem? The answer is simply in negative, because the space

$$\{v \in H^1(\Omega) : \frac{\partial u}{\partial \nu} = 0 \ \text{ on } \ \partial\Omega\}$$

is not closed. Since for Neumann problem, the boundary condition comes naturally when Green's formula is applied, we call such boundary conditions as natural boundary conditions and we do not impose homogeneous natural conditions on the basic space. In contrast, for second order problem, the Dirichlet boundary condition is imposed on the solution space and this condition is called essential boundary condition.

In general for elliptic operator of order $2m$, $(m \geq 1, \text{ integer})$ [3], the boundary conditions containing derivatives of order strictly less than m are called essential, where as the boundary conditions containing higher derivatives upto $2m-1$ and greater than or equal to m are called natural. As a rule, the homogeneous essential boundary conditions are imposed on the solution space.

Example 3.6 *(Mixed Boundary Value Problems). Consider the following mixed boundary value problem in general form: Find u such that*

$$Au = f, \quad x \in \Omega, \tag{3.19}$$

[3]For higher order (order greater than one) partial differential equations, one essential condition for equations to be elliptic is that it should be of even order.

$$u = 0 \qquad \text{on} \quad \partial\Omega_0, \tag{3.20}$$

$$\frac{\partial u}{\partial \nu_A} = g \qquad \text{on} \quad \partial\Omega_1, \tag{3.21}$$

where

$$Au = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}(a_{ij}\frac{\partial u}{\partial x_i}) + a_0 u, \quad \frac{\partial u}{\partial \nu_A} = \sum_{ij=1}^{d} a_{ij}\frac{\partial u}{\partial x_i}\nu_j,$$

the boundary $\partial\Omega = \partial\Omega_0 \cup \partial\Omega_1$ *with* $\partial\Omega_0 \cap \partial\Omega_1 = \emptyset$, *and* ν_j *is the* j^{th} *component of the external normal to* $\partial_1\Omega$.

We shall assume the following conditions on the coefficientes a_{ij}, a_0 and the forcing function f:

(i) The coefficientes a_{ij} and a_0 are continuous and bounded by a common positive constant M with $a_0 > 0$.

(ii) The operator A is uniformly elliptic, i.e., the associated quadratic form is uniformly positive definite. In other words, there is a positive constant α_0 such that

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \geq \alpha_0 \sum_{i=1}^{d} |\xi_i|^2, \quad 0 \neq \xi = (\xi_1, \ldots, \xi_d)^T \in \mathbb{R}^d. \tag{3.22}$$

(iii) Assume, for simplicity, that $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega_1)$.

Since on one part of the boundary $\partial\Omega_0$, we have zero essential boundary condition, we impose this on our solution space and thus, we consider

$$V = \{v \in H^1(\Omega) : v = 0 \ \text{ on } \ \partial\Omega_0\}.$$

This is a closed subspace of $H^1(\Omega)$ and is itself a Hilbert space. The Poincaré inequality holds true for V, provided volume or surface area of $\partial\Omega_0$ is bigger than zero (In the proof of the Poincaré inequality, we need only vanishing of elements on a part of boundary, see the proof). For weak formulation, multiply (3.19) by $v \in V$ and integrate over Ω. A use of integration by parts formula yields

$$\begin{aligned} \int_\Omega Auv\,dx &= -\sum_{i,j=1}^{d} \int_{\partial\Omega} a_{ij}\frac{\partial u}{\partial x_i}\nu_j v\,ds + \sum_{i,j=1}^{d} \int_\Omega a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,dx \\ &+ \int_\Omega a_0 uv\,dx = -\int_{\partial\Omega_1} gv\,ds + a(u,v), \end{aligned}$$

where

$$a(u,v) := \sum_{i,j=1}^{d} \int_\Omega a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,dx + \int_\Omega a_0 uv\,dx.$$

Therefore, the weak formulation becomes

$$a(u,v) = L(v), \quad v \in V, \tag{3.23}$$

where $L(v) = \int_\Omega fv\,dx + \int_{\partial\Omega_1} gv\,ds$.

For wellposedness of (3.23), let us verify the sufficient conditions of the Lax-Milgram Theorem.

(i) For boundedness, note that

$$|a(u,v)| = |\sum_{i,j=1}^{d} \int_\Omega a_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j}\,dx + \int_\Omega a_0 uv\,dx|$$

$$\leq M(\sum_{i,j=1}^{d} \|\frac{\partial u}{\partial x_i}\|\|\frac{\partial v}{\partial x_j}\| + \|u\|\|v\|).$$

Use $\sum_{i=1}^d a_i b_i \leq \sqrt{\sum_{i=1}^d a_i^2}\sqrt{\sum_{i=1}^d b_i^2}$ to obtain

$$\sum_{i=1}^{d} \|\frac{\partial u}{\partial x_i}\| \sum_{j=1}^{d} \|\frac{\partial v}{\partial x_j}\| \leq d(\sum_{i=1}^{d} \|\frac{\partial u}{\partial x_i}\|^2)^{\frac{1}{2}}(\sum_{j=1}^{d} \|\frac{\partial v}{\partial x_j}\|^2)^{\frac{1}{2}} \leq d\|\nabla u\|\|\nabla v\|.$$

Since Poincaré inequality holds, i.e., $\|v\| \leq c(\Omega)\|\nabla v\|$, $v \in V$, the seminorm $\|\nabla v\|$ is infact a norm on V and this is equivalent to $\|\cdot\|_1$-norm. Thus,

$$|a(u,v)| \leq C(M,d)\|u\|_V\|v\|_V.$$

(ii) For coercivity, observe that

$$a(v,v) = \int_\Omega \sum_{i,j=1}^{d} a_{ij}\frac{\partial v}{\partial x_i}\frac{\partial v}{\partial x_j}\,dx + \int_\Omega a_0|v|^2\,dx.$$

For the last term, use $a_0 > 0$ so that $\int_\Omega a_0|v|^2\,dx \geq 0$. For the first term, a use of uniform ellipticity with $\frac{\partial v}{\partial x_i} = \xi_i$ and $\frac{\partial v}{\partial x_j} = \xi_j$ yields

$$\int_\Omega \sum_{i,j=1}^{d} a_{ij}\frac{\partial v}{\partial x_i}\frac{\partial v}{\partial x_j}\,dx \geq \alpha_0 \sum_{i=1}^{d}\int_\Omega |\frac{\partial v}{\partial x_i}|^2\,dx = \alpha_0\|\nabla v\|^2 = \alpha_0\|u\|_V^2.$$

Alltogether, we obtain

$$a(v,v) \geq \alpha_0\|v\|_V^2.$$

Finally, the linear functional L satisfies

$$\begin{aligned} |L(v)| &\leq \int_\Omega |fv|\,dx + \int_{\partial\Omega_1} |g||\gamma_0 v|\,ds \\ &\leq \|f\|\|v\| + \|g\|_{L^2(\partial\Omega_1)}\|\gamma_0 v\|_{L^2(\partial\Omega_1)}. \end{aligned}$$

From the trace result

$$\|\gamma_0 v\|_{L^2(\partial\Omega_1)} \leq C\|v\|_1 \leq C\|v\|_V,$$

and now using Poincaré inequality, it follows that

$$\|L\|_{V'} = \sup_{0\neq v\in V} \frac{|L(v)|}{\|v\|_V} \leq C(\|f\| + \|g\|_{L^2(\partial\Omega)}).$$

An appeal to the Lax-Milgram Theorem yields the wellposedness of the weak solution $u \in V$ of (3.23). Moreover,

$$\|u\|_1 \leq C(M, \alpha_0, d, \Omega)(\|f\| + \|g\|_{L^2(\partial\Omega)}).$$

In order to explore in what sense the weak solution u of (3.23) satisfies (3.19)–(3.21), take $v \in \mathcal{D}(\Omega)$ in (3.23). Since $\mathcal{D}(\Omega) \subset V$, using the definition of distributional derivatives, (3.23) becomes

$$< Au, v >=< f, v >, \ \forall v \in \mathcal{D}(\Omega).$$

Thus,

$$Au = f, \ \text{ in } \mathcal{D}'(\Omega).$$

For the boundary conditions, it is found that for $v \in V$,

$$\begin{aligned} A(u,v) &= \int_\Omega Auv\,dx + \int_{\partial\Omega} \frac{\partial u}{\partial \nu_A} v\,ds \\ &= \int_\Omega fv\,dx + \int_{\partial\Omega_1} \frac{\partial u}{\partial \nu_A} v\,ds, \ \ (v = 0 \text{ on } \partial\Omega_0), \end{aligned}$$

and

$$L(v) = \int_\Omega fv\,dx + \int_{\partial\Omega_1} gv\,ds.$$

From (3.23), we obtain

$$\int_{\partial\Omega_1} \frac{\partial u}{\partial \nu_A} v\,ds = \int_{\partial\Omega_1} gv\,ds, \ \forall v \in V,$$

and hence, the boundary conditions are satisfied.

Note that even if f and g are smooth, the weak solution may not be in $H^2(\Omega)$, because of the common points or the common curves between $\partial\Omega_0$ and $\partial\Omega_1$.

Since the bilinear form is symmetric (as $a_{ij} = a_{ji}$, if not redefine $\overline{a}_{ij} = \frac{(a_{ij}+a_{ji})}{2}$, and this is possible), the equation (3.23) is equivalent to the minimization of the energy functional J, where

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

So far, all the problems, we have discussed have symmetric bilinear forms and hence, each of those leads to an equivalent energy formulation. The example considered, below, does not have an equivalent energy minimization counter part.

Example 3.7 *Consider the following flow fluid problem with a transport term*

$$-\Delta u + \vec{b} \cdot \nabla u + c_0 u = f \quad \text{in} \quad \Omega, \tag{3.24}$$

$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{3.25}$$

where $\vec{b} = (b_1, \ldots, b_d)^T$ *is a constant vector and* $c_0 \geq 0$.

With an obvious choice of space V as $H_0^1(\Omega)$, we multiply the equation (3.24) by $v \in H_0^1(\Omega)$ and the integrate over Ω to have

$$\int_\Omega (-\Delta u) v \, dx + \int_\Omega (\vec{b} \cdot \nabla u) v \, dx + \int_\Omega c_0 u v \, dx = \int_\Omega f v \, dx, \quad v \in V.$$

For the first term apply Green's formula and now rewrite the weak formulation as

$$a(u, v) = L(v), \quad v \in V, \tag{3.26}$$

where

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega (\vec{b} \cdot \nabla u) v \, dx + \int_\Omega c_0 u v \, dx$$

and $L(v) = \int_\Omega f v \, dx$. Note that $a(\cdot, \cdot)$ is not symmetric as

$$a(u, v) - a(v, u) = \int_\Omega [(\vec{b} \cdot \nabla u) v - (\vec{b} \cdot \nabla v) u] \, dx \neq 0.$$

It is straight forward to check $a(\cdot, \cdot)$ is bounded (use Cauchy-Schwarz inequality, boundedness of $\vec{b}$ and c_0).

For coercicity,

$$a(v, v) = \int_\Omega |\nabla v|^2 dx + \int_\Omega (\vec{b} \cdot \nabla v) v dx + \int_\Omega c_0 |v|^2 dx.$$

Note that

$$\begin{aligned} \int_\Omega (\vec{b} \cdot \nabla v) v dx &= \sum_{i=1}^d \int_\Omega b_i \frac{\partial v}{\partial x_i} v dx \\ &= \frac{1}{2} \sum_{i=1}^d b_i \int_\Omega \frac{\partial (v^2)}{\partial x_i} dx \quad (\vec{b} \text{ is constant}) \\ &= \frac{1}{2} \sum_{i=1}^d b_i \int_{\partial\Omega} v^2 \nu_i ds = 0 \text{ as } v \in V. \end{aligned}$$

Thus,

$$\begin{aligned} a(v, v) &= \int_\Omega |\nabla v|^2 dx + \int_\Omega c_0 |v|^2 dx \\ &\geq \|\nabla v\|^2 = \|v\|_V^2. \end{aligned}$$

Here, we have used $\int_\Omega c_0|v|^2dx \geq 0$ as $c_0 \geq 0$. Note that when the vector $\vec{b}$ is not constant, under some restrictions on $\vec{b}$ it is possible to check coercivity of $a(\cdot,\cdot)$. Using the Lax-Milgram Theorem, there exists a unique solution u to (3.26) and

$$\|u\|_1 \leq c\|f\|.$$

Note that the minimization of the functional (in this case as $\vec{b}$ is constant) J, where

$$J(v) = \frac{1}{2}\int_\Omega (|\nabla v|^2 + c_0|v|^2)dx - \int_\Omega fvdx$$

will yield a wrong solution. In general when $\vec{b}$ is not constant or we use other boundary conditions, it is not even possible to write down the corresponding energy like functional J.
Exercises. Find the admissible space, and derive the weak formulation. Using the Lax-Milgram Theorem prove the wellposedness of weak solutions.

1.)
$$-\frac{d}{dx}(a(x)\frac{du}{dx}) = f, \quad 0 < x < 1, \quad u(0) = 0, \quad \frac{du}{dx}(1) = 0.$$
Here $a(x) \geq a_0 > 0$, $f \in L^2(0,1)$.

2.)
$$-\frac{d}{dx}(a(x)\frac{du}{dx}) + b(x)\frac{du}{dx} + c(x)u = f, \quad \frac{du}{dx}(0) = \frac{du}{dx}(1) = 0.$$
Please find approperiate conditions on b, relating a and c. Here again $a(x) \geq a_0 > 0$, $c \geq c_0 > 0$, and a, b, c are bounded by a common constant M.

3.) The following Robinson Problem:
$$\begin{cases} -\Delta u + \lambda u = f & in\ \Omega \\ \alpha u + \frac{\partial u}{\partial \nu} = 0 & on\ \partial\Omega \end{cases}$$
where $\alpha \geq 0$ and $\lambda > 0$ constants (sometimes this boundary condition is called Fourier Boundary Condition).

In all the above cases, discuss in what sense the weak solutions satisfy the corresponding differential equations.

Example 3.8 *(Biharmonic Problems). Consider the following fourth order partial differential equations*

$$-\Delta^2 u = f \quad \text{in} \quad \Omega, \tag{3.27}$$

$$u = 0 \quad \text{on} \quad \partial\Omega, \tag{3.28}$$

$$\frac{\partial u}{\partial \nu} = 0 \quad \text{on} \quad \partial\Omega. \tag{3.29}$$

To choose an appropriate function space, multiply the partial differential equation by a smooth function $v \in \mathcal{D}(\Omega)$ or by $C^2(\Omega)$ such that $v = 0$ and $\frac{\partial v}{\partial \nu} = 0$ on $\partial\Omega$ (both these boundary conditions are essential boundary conditions for the 4th order problem) and integrate over Ω. Apply Gauss divergence theorem to the first term, i.e.,

$$\begin{aligned}\int_\Omega \Delta(\Delta u)v\,dx &= -\int_\Omega \nabla(\Delta u)\cdot\nabla v\,dx + \int_{\partial\Omega}\nabla(\Delta u)\cdot\nu v\,ds \\ &- \int_{\partial\Omega}\Delta u(\nabla u\cdot\nu v)\,ds + \int_\Omega \Delta u\Delta v\,dx.\end{aligned}$$

Since both the boundary terms vanish as $v = 0$ and $\frac{\partial v}{\partial \nu} = 0$ on $\partial\Omega$, we have

$$\int_\Omega \Delta^2 uv dx = (\Delta u, \Delta v) = a(u, v).$$

Similarly,

$$L(v) = \int_\Omega fv dx.$$

The largest space for which these integrals are meaningful is $u \in H^2(\Omega)$. Since the boundary conditions are essential type, we impose it on the space. Thus, the admissible space $V = H_0^2(\Omega)$, where

$$H_0^2(\Omega) = \{v \in H^2(\Omega) : v = 0, \quad \frac{\partial v}{\partial \nu} = 0 \ \text{ on } \ \partial\Omega\}.$$

Since $\mathcal{D}(\Omega)$ is dense in $H_0^2(\Omega)$, we have the following weak formulation: Find $u \in H_0^2(\Omega)$ such that

$$a(u, v) = L(v), \quad \forall v \in H_0^2(\Omega). \tag{3.30}$$

Using Poincaré inequality for ∇v as $\nabla v \cdot \nu = 0$ on $\partial\Omega$, we have

$$\|\nabla v\|^2 \le c(\Omega)\|\Delta v\|^2.$$

Similarly, as $v = 0$ on $\partial\Omega$

$$\|v\|^2 \le c(\Omega)\|\nabla v\|^2.$$

Therefore, $\|\Delta v\|$ is a norm on $H_0^2(\Omega)$ which is equivalent to $H^2(\Omega)-$norm. Note that

$$\|v\|^2_{H^2(\Omega)} = \|v\|^2 + \|\nabla v\|^2 + \sum_{|\alpha|=2}\|D^\alpha v\|^2.$$

Since for the last term, contains not only $\|\Delta v\|^2$ but also the cross derivatives, we need to apply interchange of derivatives like: Consider $v \in \mathcal{D}(\Omega)$, then we shall use denseness to extend it to $H_0^2(\Omega)$.

$$\int_\Omega (\frac{\partial^2 v}{\partial x_i \partial x_j})^2\,dx =< D_{ij}v, D_{ij}v >= - < \frac{\partial v}{\partial x_i}, \frac{\partial^3 v}{\partial x_i \partial^2 x_j} >=< \frac{\partial^2 v}{\partial x_i^2}, \frac{\partial^2 v}{\partial x_j^2} > .$$

Use denseness of $\mathcal{D}(\Omega)$ in $H_0^2(\Omega)$, so that

$$\int_\Omega |\frac{\partial^2 v}{\partial x_i \partial x_j}|^2 dx = \int_\Omega \frac{\partial^2 v}{\partial x_i^2}\frac{\partial^2 v}{\partial x_j^2} dx.$$

Taking summation over $i, j = 1, \ldots, d$, we obtain

$$\sum_{|\alpha|=2} D^\alpha v\|^2 = \sum_{i=1}^d \sum_{j=1}^d \int_\Omega |\frac{\partial^2 v}{\partial x_i \partial x_j}|^2 dx = \int_\Omega |\Delta v|^2 dx.$$

A use of Poincaré inequality yields

$$\|v\|^2_{H^2(\Omega)} \leq C\|\Delta v\|^2,$$

where C is a generic positive constant. Thus

$$\|\Delta v\|^2 \leq \|v\|_2^2 \leq C\|\Delta v\|^2,$$

and these two norms are equivalent. In order to verify the conditions in the Lax-Milgram Theorem, we observe that

$$|a(u,v)| \leq \int_\Omega |\Delta u||\Delta v| dx \leq \|\Delta u\|\|\Delta v\| = \|u\|_V \|v\|_V,$$

and

$$a(v,v) = \int_\Omega |\Delta v|^2 dx = \|v\|_V^2.$$

Further,

$$|L(v)| = |\int_\Omega fv dx| \leq \|f\|\|v\| \leq C\|f\|\|\Delta v\| = C\|f\|\|v\|_V,$$

and hence,

$$\|L\|_{V'} = \|L\|_{-2} \leq C\|f\|.$$

Application of the Lax-Milgram Theorem yields the wellposedness of the weak solution $u \in H_0^2(\Omega)$. Moreover,

$$\|u\|_2 \leq C\|\Delta u\| \leq C\|f\|.$$

In fact, it is an easy exercise to check that the weak solution u satisfies the PDE in the sense of distribution.

Example 3.9 *(Steady State Stokes Equations). Consider the steady state Stokes system of equations: Given the vector $\vec{f}$, find a velocity vector $\vec{u}$ and pressure p of an incompressible fluid such that*

$$-\Delta \vec{u} + \nabla p = \vec{f}, \text{ in } \Omega, \tag{3.31}$$

$$\nabla \cdot \vec{u} = 0, \text{ in } \Omega, \tag{3.32}$$

$$\vec{u} = 0, \text{ on } \partial\Omega, \tag{3.33}$$

where Ω is a bounded domain with smooth boundary $\partial\Omega$ and $\vec{u} = (u_1, \cdots, u_d)$.

Here, the incompressibility condition is given by the equation (3.33). However, it is this condition which makes the problem difficult to solve.

For variational formulation, we shall use the following function spaces. Let

$$\mathcal{V} := \{\vec{v} \in (\mathcal{D}(\Omega))^d : \nabla \cdot \vec{v} = 0, \text{ in } \Omega\},$$

and further, let V be the closure of $\mathcal{V}$ with respect to $\mathbf{H_0^1}$-norm, where

$$\mathbf{H}_0^1 := \{\vec{v} \in (H_0^1(\Omega))^d : \vec{v} = 0 \text{ in } \Omega\}.$$

In fact,

$$V := \{\vec{v} \in \mathbf{H}_0^1(\Omega) : \nabla \cdot \vec{v} = 0 \text{ in } \Omega\}.$$

For a proof, see Temam [11]. The norm on $\mathbf{H}_0^1(\Omega)$ will be denoted by

$$\|\vec{v}\|_{\mathbf{H}_0^1(\Omega)} = (\sum_{i=1}^{d} \|\nabla v_i\|^2)^{\frac{1}{2}}.$$

for weak formulation, form an innerproduct of (3.31) with $\vec{v} \in \mathcal{V}$ and then integrate by parts the first term on the left hand side to obtain

$$\begin{aligned} -\int_\Omega \Delta\vec{u} \cdot \vec{v}\, dx &= \sum_{i,j=1}^{d} \int_\Omega \frac{\partial^2 u_j}{\partial x_i^2} v_j\, dx \\ &= -\sum_{i,j=1}^{d} \int_{\partial\Omega} \frac{\partial u_j}{\partial x_i} \nu_i v_j\, ds + \sum_{i,j=1}^{d} \int_\Omega \frac{\partial u_j}{\partial x_i} \frac{\partial v_j}{\partial x_i}\, dx. \end{aligned}$$

Since $\vec{v} = 0, \text{ on } \partial\Omega$, it follows that

$$-\int_\Omega \Delta\vec{u} \cdot \vec{v}\, dx = (\nabla\vec{u}, \nabla\vec{v}).$$

for the second term on the left hand side, use again integration by parts to find that

$$\begin{aligned} \int_\Omega \nabla p \cdot \vec{v}\, dx &= \sum_{i=1}^{d} \int_\Omega \frac{\partial p}{\partial x_i} v_i\, dx \\ &= \sum_{i=1}^{d} \int_{\partial\Omega} p\nu_i v_i\, ds - \int_\Omega p\nabla \cdot \vec{v}\, dx = 0. \end{aligned}$$

Here, we have used $\vec{v} \in \mathcal{V}$. Now the admissible space is V. Since $\mathcal{V}$ is dense in V, the weak formulation of (3.31)–(3.33) reads as : ‘ Find $\vec{u} \in V$ such that

$$a(\vec{u}, \vec{v}) = L(\vec{v}), \ \forall \vec{v} \in V, \tag{3.34}$$

where $a(\vec{u}, \vec{v}) = (\nabla\vec{u}, \nabla\vec{v})$ and $L(\vec{v}) = (\vec{f}, \vec{v})$. It is now straight forward to check that the bilinear form is coercive as well as bounded in V and moreover, the

linear form L is continuous. note that the Poincaré inequality is also valid for the functions in V. Therefore, apply the Lax-Milgram Theorem to conclude the wellposedness of the weak solution $\vec{u}$ of (3.34).

Since the bilinear form is symmetric, (3.34) is equivalent to the following minimization problem:

$$J(\vec{u}) = \min_{\vec{v} \in V} J(\vec{v})$$

where

$$J(\vec{v}) = \frac{1}{2}\|\nabla \vec{v}\|^2 - (\vec{f}, \vec{v}).$$

Now suppose that $\vec{u}$ satisfies (3.34), then we shall show that $\vec{u}$ satisfies (3.31)-(3.33) in some sense. Now $\vec{u} \in V$ implies that $\nabla \cdot \vec{u} = 0$ in the sense of distribution. Then for $v \in V$, using integration by parts in (3.34), we obtain

$$< -\Delta \vec{u} - f, v >= 0, \quad \forall v \in V. \tag{3.35}$$

At this stage, let us recall some results from Temam [11] (see, pp. 14–15).

Lemma 3.10 *Let Ω be an open set in $\mathbb{R}^d$ and $\vec{F} = (f_1, \ldots, f_d)$ with each $f_i \in \mathcal{D}'(\Omega)$, $i = 1, \ldots, d$. A necessary and sufficient condition that $\vec{F} = \nabla p$, for some $p \in \mathcal{D}'(\Omega)$ is that*

$$< \vec{F}, v >= 0, \quad \forall v \in V.$$

Now using the above Lemma with $\vec{F} = \Delta \vec{u} + \vec{F}$, we have

$$-\Delta \vec{u} - \vec{F} = -\nabla p \tag{3.36}$$

for some $p \in \mathcal{D}'(\Omega)$.

Note. For more details on regularity results, see Agmon [2], [5] and also Kesavan [9].

Chapter 4

Elliptic Equations

In the first part of this chapter, we discuss the finite dimensional approximation of the abstract variational formulations described in the third chapter and derive error estimates using Cea's Lemma. The later part will be devoted to the discussion of some examples, derivation of the finite element equations and adaptive methods for elliptic equations.

4.1 Finite Dimensional Approximation to the Abstract Variational Problems.

Let us first recall the abstract variational problem discussed in the previous chapter. Given a real Hilbert space V and a linear functional L on it, find a function $u \in V$ such that

$$a(u, v) = L(v) \ \forall v \in V, \tag{4.1}$$

where the bilinear form $a(\cdot, \cdot)$ and the linear functional L satisfy the conditions in the Lax- Milgram Theorem in chapter 3.

Let V_h be a finite dimensional subspace of V, which is parametrized by a parameter h. Let us pose the the problem (4.1) in this finite dimensional setting as: seek a solution $u_h \in V_h$ such that

$$a(u_h, v_h) = L(v_h) \ \forall v_h \in V_h. \tag{4.2}$$

In literature, this formulation is known as Galerkin method. We may curious to know whether the problem (4.2) is wellposed. Since V_h is a finite dimensional with dimension say N_h, we may write down a basis of V_h as $\{\phi_i\}_{i=1}^{N_h}$. Let

$$u_h(x) = \sum_{i=1}^{N_h} \alpha_i \phi_i,$$

where α_i's are unknowns. Similarly, represent $v_h(x) = \sum_{i=j}^{N_h} \beta_j \phi_j$, for some β_j's. On substitution in (4.2), we obtain

$$\sum_{i,j=1}^{N_h} \alpha_i \beta_j a(\phi_i, \phi_j) = \sum_{j=1}^{N_h} \beta_j L(\phi_j).$$

In matrix form

$$\beta^T A \alpha = \beta^T b, \tag{4.3}$$

where the global stiffness matrix $A = [a_{ij}]$ with $a_{ij} = a(\phi_i, \phi_j)$ and the global load vector $b = (b_j)$ with $b_j = L(\phi_j)$. Since the system (4.2) is true for all $v \in V_h$, therefore, the equation (4.3) holds for all $\beta \in \mathbb{R}^{N_h}$. Thus,

$$A\alpha = b. \tag{4.4}$$

In other words, it is enough to consider the problem (4.2) as

$$a(u_h, \phi_j) = L(\phi_j), \;\; j = 1, \ldots, N_h,$$

that is replacing v_h by the basis function ϕ_j. Since the bilinear form $a(\cdot, \cdot)$ is coercive, the matrix A is positive definite. To verify this, we note that for any vector $0 \neq \beta \in \mathbb{R}^{N_h}$

$$\beta^T A \beta = \sum_{i,j=1}^{N_h} \beta_i \beta_j a(\phi_i, \phi_j) = a(\sum_{i=1}^{N_h} \beta_i \phi_i, \sum_{j=1}^{N_h} \beta_j \phi_j) \geq \alpha_0 \| \sum_{i=1}^{N_h} \beta_i \phi_i \|^2,$$

and hence, the linear system (4.4) has a unique solution $\alpha \in \mathrm{R}^{N_h}$. In other words, the problem (4.2) has a unique solution $u_h \in V_h$. This can also be proved directly by applying Lax-Milgram Theorem to (4.2) as $V_h \subset V$. When the bilinear form is symmetric, then the matrix A becomes symmetric. Therefore, the problem (4.4) can be written equivalently as a minimization problem like

$$J(\alpha) = \min_{\beta \in \mathbb{R}^{N_h}} J(\beta),$$

where $J(\beta) = \frac{1}{2} \beta^T A \beta - \beta^T b$, or simply, we write

$$J(u_h) = \min_{v_h \in V_h} J(v_h),$$

where $J(v_h) = \frac{1}{2} a(v_h, v_h) - L(v_h)$. Note that we identify the space V_h and $\mathbb{R}^{N_h}$ through the basis $\{\phi_i\}$ of V_h and the natural basis $\{e_i\}$ of $\mathbb{R}^{N_h}$, i.e., u_h is a solution of (4.2) if and only if α is a solution of $A\alpha = b$, where $u_h = \sum_{i=1}^{N_h} \alpha_i \phi_i$.

One of the attractive feature of finite element methods is that the finite dimensional spaces are constructed with basis functions having small support. For example, if the bilinear form is given by an integral like

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

then $a(\phi_i, \phi_j) = 0$, if $supp\,\phi_i \cap supp\,\phi_j = \emptyset$. Therefore, small support of each ϕ_i entails the sparseness of the corresponding stiffness matrix A. For computational purpose, this is a desirable feature as fast iterative methods are available for solving large sparse linear systems.

Below, we prove the Cea's Lemma which deals with the estimates of the error $e = u - u_h$, committed in the process of approximation.

Lemma 4.1 *(Cea's Lemma). Let u and u_h, respectively, be the solutions of (4.1) and (4.2). Let all the conditions for the Lax-Milgram Theorem hold. Then there is a constant C independent of the discretization parameter h such that*

$$\|u - u_h\|_V \leq C \inf_{\chi \in V_h} \|u - \chi\|_V.$$

Proof. Since $V_h \subset V$, we have for $\chi \in V_h$

$$a(u, \chi) = L(\chi),$$

and

$$a(u_h, \chi) = L(\chi).$$

Hence, taking the difference and using linearity of the form $a(\cdot, \cdot)$, we obtain

$$a(u - u_h, \chi) = 0, \;\; \chi \in V_h. \tag{4.5}$$

Using coercivity property of the bilinear form, it follows that

$$\begin{aligned} \alpha_0 \|u - u_h\|_V^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - \chi) + a(u - u_h, \chi - u_h) \\ &= a(u - u_h, u - \chi). \end{aligned}$$

Here, the second term in the last but one step is zero by using (4.5). Since the bilinear form is bounded, we infer that

$$\|u - u_h\|_V^2 \leq \frac{M}{\alpha_0} \|u - u_h\|_V \|u - \chi\|_V.$$

In case $u = u_h$, the result follows trivially. Now for $u \neq u_h$, we can cancel one term $\|u - u_h\|_V$ from both sides and then taking infimum over $\chi \in V_h$ (as χ is arbitrary), we obtain the desired estimate with $C = \frac{M}{\alpha_0}$. This completes the rest of the proof.

Remark. In literature, the property (4.5) is called as **Galerkin orthogonality**. The space V_h being a finite dimensional subspace of the real Hilbert space V is closed in V , and hence, the infimum is attained. In fact, the infimum is the best approximation say $P_1 u$ of u onto V_h. Given V_h, the maximum truncation error which can be expected is $\|u - P_1 u\|_V$. This is what we call the consistency condition. The stability is a straight forward consequence of the boundedness and coercivity of the bilinear form. Say , choose $v_h = u_h$ in (4.2) to obtain

$$\alpha_0 \|u_h\|_V^2 \leq a(u_h, u_h) = |L(u_h)| \leq \|L\|_{V'} \|u_h\|_V,$$

and hence,

$$\|u_h\|_V \leq \frac{\|L\|_{V'}}{\alpha_0}.$$

This is bounded independent of the parameter h and therefore, the Galerkin method is stable with respect to V- norm. Now we can apply the Lax-Ritchmyer equivalence theorem to infer directly the convergence result. Note that when the bilinear form $a(\cdot,\cdot)$ is symmetric, the Galerkin approximation u_h is the best approximation with respect to the energy norm $(a(v,v))^{1/2}$, $v \in V$.

4.2 Examples.

In this section, we examine through a couple of examples, the essential features of the finite element spaces which play a vital role in the error estimates especially in the optimal rate of convergence. It may be difficult for us to spend more time on the construction of the finite element spaces and the derivation of their approximation properties. But we indicate the basic steps through the two examples given below and for more details, we refer to some excellent books by Ciarlet [4], Johnson [8], and Brenner and Scott [3].

Example 4.2 *(Two point boundary value problems). Consider the following simple linear two point boundary value problem*

$$-(a(x)u')' + a_0(x)u = f(x),\ x \in I := (0,1), \tag{4.6}$$

with boundary conditions

$$u(0) = 0, \quad u'(1) = 0,$$

where $a \geq \alpha_0 > 0$, $a_0 \geq 0$, $\max_{x\in[0,1]}(|a|,|a_0|) \leq M$ *and* $f \in L^2(I)$.

One of the basic step in the finite element method is the weak or variational formulation of the original problem (4.6). Following the procedures in the last chapter, it is straight forward to guess the admissible space V as

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

Now form an innerproduct between (4.6) and $v \in V$, and use integration by parts to the first term on the left hand side. The boundary term at $x = 0$ vanishes as $v(0) = 0$, while the other boundary term becomes zero , because of the homogeneous Neumann condition $u'(1) = 0$. Then the weak form of (4.6) is to seek a function $u \in V$ such that

$$a(u,v) = (f,v) \ \ \forall v \in V, \tag{4.7}$$

where

$$a(u,v) = \int_0^1 (au'v' + a_0uv)\,dx.$$

Construction of finite element spaces V_h. For any positive integer N, let $\{0 = x_0 < x_1 < \cdots < x_N = 1\}$ be a partition of the interval I into sub-intervals called elements $I_i = (x_{i-1}, x_i)$, $i = 1, \cdots, N$ with length $h_i = x_i - x_{i-1}$. Let $h = \max_{1 \le i \le N} h_i$ be the parameter associated with the finite element mesh. The simplest finite element space defined on the above partition is the following C^0-piecewise linear spline space, i.e.,

$$V_h := \{v_h \in C^0([0,1]) : v_h|_{I_i} \in P_1(I_i), i = 1, 2, \cdots, N,\ v_h(0) = 0\},$$

where $P_1(I_i)$ is a linear polynomial. On each subintervals or elements I_i, v_h coincides with a linear polynomial like $a_i + b_i x$, where a_i and b_i are unknowns. Therefore, we have $2N$ unknowns or degrees of freedom. But the continuity requirement on the internal nodal points or the mesh points $(x_1, \cdots, x_{N-1})$ would impose $N-1$ additional conditions and one more due to imposition of left hand boundary condition. Thus, the total degrees of freedom or the dimension of V_h is N. Now, the specification of values at the nodal points $(x_1, \cdots, x_N)$ uniquely determines the elements of V_h. Setting the Lagrange interpolating polynomials $\{\phi_i\}_{i=1}^N$ as elements in V_h which satisfy

$$\phi_i(x_j) = \delta_{ij},\ 1 \le i, j \le N,$$

where the Krönecker delta function $\delta_{ij} = 1$, when $i = j$ and is equal to zero for $i \neq j$. It is an easy excercise to check that $\{\phi_i\}_{i=1}^N$ forms a basis for V_h. More over, for $i = 1, \cdots, N-1$

$$\phi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_i}, & x_{i-1} \le x \le x_i, \\ \frac{x_{i+1} - x}{h_{i+1}}, & x_i \le x \le x_{i+1}, \end{cases}$$

and for $i = N$

$$\phi_N(x) = \frac{1}{h_N}(x - x_{N-1}).$$

Each of these hat functions ϕ_i, $i = 1, \cdots, N-1$ has support in $I_i \cup I_{i+1}$ and *supp* $\phi_N = I_N$.

As mentioned earlier, each element $v_h \in V_h$ is uniquely determined by its values at the nodal points x_j, $j = 1, \ldots, N$, that is,

$$v_h(x) = \sum_{i=1}^{N} v_h(x_i)\phi_i(x).$$

Given a smooth function v with $v(0) = 0$, it can be approximated by its nodal interpolant $I_h v$ in V_h as

$$I_h v(x_i) = v(x_i),\ \ i = 1, \ldots, N.$$

Thus, $I_h v(x) = \sum_{i=1}^N v(x_i)\phi_i(x)$.

To estimate the error in the interpolation, we quickly prove the following Theorem.

Theorem 4.3 *Let v be an element in $H^2(I)$ which vanishes at $x = 0$. Let $I_h v$ be its nodal interpolant which is defined as above. Then the following estimates*

$$\|(v - I_h v)'\| \leq C_1 h \|v''\|,$$

and

$$\|v - I_h v\| \leq C_2 h^2 \|v''\|$$

hold true.

Proof. We shall first prove the result for individual elements and then sum up to obtain the result for the entire interval. On each element I_j, the interpolant $I_h v$ is a linear function and is such that $I_h v(x_{j-1}) = v(x_{j-1})$ and $I_h v(x_j) = v(x_j)$. Then we have a representation of $I_h v$ as

$$I_h v(x) = v(x_{j-1})\phi_{1,j} + v(x_j)\phi_{2,j} \ \ \forall x \in I_j, \tag{4.8}$$

where $\phi_{1,j} = \frac{x_j - x}{h_j} = 1 - \phi_{2,j}$ and $\phi_{2,j} = \frac{x - x_{j-1}}{h_j}$ are local basis functions on I_j such that $\phi_{i,j}$, $i = 1, 2$ takes value 1 at the left and right hand nodes, respectively, and zero elsewhere. Using Taylors formula for $i = j, j-1$, we have

$$v(x_i) = v(x) + v'(x)(x_i - x) + \int_x^{x_i} (x_i - s) v''(s)\, ds. \tag{4.9}$$

Substitute (4.9) in (4.8) and use the relations $\phi_{1,j} + \phi_{2,j} = 1$ and $(x_{j-1} - x)\phi_{1,j} + (x_j - x)\phi_{2,j} = 0$ to obtain the following representation of interpolation error

$$v(x) - I_h v(x) = -[\phi_{1,j} \int_{x_{j-1}}^{x} (s - x_{j-1}) v''(s)\, ds + \phi_{2,j} \int_x^{x_j} (x_j - s) v''(s)\, ds]. \tag{4.10}$$

First squaring both sides and using $(a + b)^2 \leq 2(a^2 + b^2)$, we integrate over I_j. Repeated applications of Cauchy-Schwarz inequality with analytic evaluation of L^2-norms of the local basis functions, we obtain the desired L^2 estimate for the interpolation error on each subinterval I_j. Since $\|v - I_h v\|^2 = \sum_{j=1}^{N} \|v - I_h v\|^2_{L^2(I_j)}$, the result follows for the L^2- norm

For the other estimate, we differentiate the error (4.10). Note that for the integrals , differentiation with respect to x will give values of v'' almost every where and then repeat the above procedure to complete the rest of the proof.

Remark. When $v \in W^{2,\infty}(I)$, then use (4.10) with some standard modifications to obtain the following maximum norm error estimate

$$\|v - I_h v\|_{L^\infty(I)} + h\|(v - I_h v)'\|_{L^\infty(I)} \leq C_3 h^2 \|v''\|_{L^\infty(I)}.$$

It is to be noted that quasi-uniformity condition on the finite element mesh is needed for the above estimates that is $\frac{h_j}{h} \geq c \ \forall h_j$ and for some positive constant c. The procedure remains same for the higher degree C^0-piecewise polynomial spaces. However, instead of looking at each element one concentrate

only on a master element on which the construction of local basis functions is relatively simpler. Then, appropriate nonsingular transformations which are easy to findout take the required estimates back to the individual elements.

With the finite dimensional space V_h, the Galerkin formulation is to seek a function $u_h \in V_h$ such that

$$a(u_h, v_h) = (f, v_h), \;\; v_h \in V_h. \tag{4.11}$$

Theorem 4.4 *Let u be a solution of (4.7) with $u \in H^2(I) \cap H_0^1(I)$. Let u_h be a solution of (4.11). Then there is a constant C independent of h such that the error $e = (u - u_h)$ satisfies*

$$\|e'\| \leq Ch\|u''\| \text{ and } \|e\| \leq Ch^2\|u''\|.$$

Proof. Note that

$$|a(u, v)| \leq M\|u\|_1\|v\|_1 \leq 2M\|u'\|\|v'\|$$

and

$$a(u, u) \geq \alpha_0\|u'\|^2.$$

For the boundedness we have used the Poincaré inequality $\|v\| \leq \|v'\|$ with $V = H_0^1(I)$. Now apply Cea's Lemma to obtain

$$\|u - u_h\|_V = \|(u - u_h)'\| \leq C(\alpha_0) \inf_{\chi \in V_h} \|(u - \chi)'\| \leq C\|(u - I_h u)'\| \leq Ch\|u''\|.$$

For estimate in L^2-norm, we use the Aubin -Nitsche duality arguments. Consider the following adjoint elliptic problem: For $g \in L^2(I)$, let ϕ be a solution of

$$-(a(x)\phi')' + a_0(x)\phi = g, \;\; x \in I, \tag{4.12}$$

with boundary conditions

$$\phi(0) = 0, \;\; \phi'(1) = 0.$$

The above problem satisfies the following regularity result

$$\|\phi\|_2 \leq C\|g\|. \tag{4.13}$$

Multiply the equation (4.12) by e and integrate with respect to x from 0 to 1. Use the Neumann boundary condition for ϕ and $e(0) = 0$ to have

$$(e, g) = a(e, \phi) = a(e, \phi - I_h\phi).$$

Here, we have used the Galerkin orthogonality condition $a(e, \chi) = 0 \; \forall \chi \in V_h$. A use of boundedness of the bilinear form yields

$$|(e, g)| \leq C\|e'\|\|(\phi - I_h\phi)'\| \leq Ch\|e'\|\|\phi\|_2 \leq Ch\|e'\|\|g\|.$$

In the last step we have employed the elliptic requlrity result (4.13). Now the result follows as $\|e\| = \sup_{\phi \in L^2(I)} |(e, g)|$.

Remark. We shall defer the computational procedure and the construction of finite element equations to the next section.

Example 4.5 *(The standard Galerkin method for elliptic equations). Let Ω be a domain in R^d with smooth boundary $\partial\Omega$ and consider the following homogeneous Dirichlet boundary value problem*

$$-\Delta u + u = f, \quad \text{in } \Omega, \tag{4.14}$$

$$u = 0, \ \text{ on } \partial\Omega, \tag{4.15}$$

where $\Delta = \Sigma_{j=1}^{d} \partial^2/\partial x_j^2$ is the Laplace operator.

For the Galerkin formulation to the boundary value problem (4.14), we first write this problem in weak or variational form as: find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = (f, v) \ \forall v \in H_0^1(\Omega), \tag{4.16}$$

where $a(u, v) = (\nabla u, \nabla v) + (u, v)$ This has been done in the last chapter.

Construction of finite element space. We consider the simplest finite element space V_h and then apply Galerkin method to the variational form (4.16) using V_h. For easy exposition, we shall assume that the domain Ω is a convex polygon in $\mathbb{R}^2$ with boundary $\partial\Omega$. Let $\mathcal{T}_h$ denote a partition of Ω into disjoint triangles K of diameter h_K such that

(i) $\cup_{K_i \in \mathcal{T}_h} \overline{K_i} = \overline{\Omega}, \ K_i \cap K_j = \emptyset, \ K_i, K_j \in \mathcal{T}_h$ with $K_i \neq K_j$.

(ii) No vertex of any triangle lies on the interior of a side of another triangle.

Let h_K be the longest side of the triangle $K \in \mathcal{T}_h$ and let $h = \max_{K \in \mathcal{T}_h} h_K$ denote the discretization parameter, i.e., the maximal length of a side of the triangulation $\mathcal{T}_h$. Therefore, the parameter h decreases as the triangulation is made finer. We assume that the angles of the triangulations are bounded below, independently of h. Sometimes (mainly for maximum error estimates), we also assume that the triangulations are quasi-uniform in the sense that the triangles of $\mathcal{T}_h$ are essentially of the same size. This may be expressed by demanding that the area of K in $\mathcal{T}_h$ is bounded below by ch^2 with $c > 0$ independent of h.

Let V_h denote the continuous functions on the closure $\bar{\Omega}$ of Ω which are linear on each triangle of $\mathcal{T}_h$ and vanish on $\partial\Omega$, i.e.,

$$V_h := \{\chi \in C^0(\overline{\Omega}) : \chi|_K \in \mathbb{P}_1(K) \ \forall K \in \mathcal{T}_h, \chi = 0 \text{ on } \partial\Omega\}.$$

Apparently, it is not quite clear that V_h is a subspace of $H_0^1(\Omega)$. However, an easy excercise shows that $V_h \subset H_0^1(\Omega)$. (Hints: What is needed to be shown that $\frac{\partial}{\partial x_i}(v_h|_K) \in L^2(\Omega), \ K \in \mathcal{T}_h$. Use distributional derivative and the contributions on the interelement boundaries—Try this !).

Let $\{P_j\}_1^{N_h}$ be the interior vertices of $\mathcal{T}_h$. A function in V_h is then uniquely determined by its values at the points P_j and thus depends on N_h parameters. Let Φ_j be the "pyramid function" in V_h which takes the value 1 at P_j but vanishes at the other vertices that is $\Phi_i(P_j) = \delta_{ij}$. Then $\{\Phi_j\}_1^{N_h}$ forms a basis for V_h, and every v_h in V_h can be expressed as

$$v_h(x) = \sum_{j=1}^{N_h} \alpha_j \Phi_j(x) \text{ with } \alpha_j = v_h(P_j).$$

Given a smooth function v on Ω which vanishes on $\partial\Omega$, we can now, for instance, approximate it by its interpolant $I_h v$ in S_h, which we define as the element of S_h that agrees with v at the interior vertices, *i.e.* $I_h v(P_j) = v(P_j)$ for $j = 1, ..., N_h$. The following error estimates for the interpolant just defined in the case of a convex domain in $\mathbb{R}^2$ are well known, namely;

$$\|I_h v - v\| \leq Ch^2 \|v\|_2,$$

and

$$\|\nabla(I_h v - v)\| \leq Ch\|v\|_2,$$

for $v \in H^2(\Omega) \cap H_0^1(\Omega)$. We shall briefly indicate the steps leading to the above estimates, and for a proof, we refer to an excellent text by Ciarlet [4]. These results may be derived by showing the corresponding estimate for each $K \in \mathcal{T}_h$ and then adding the squares of these estimates. For an individual $K \in \mathcal{T}_h$ the proof is achieved by means of the Bramble-Hilbert lemma, noting that $I_h v - v$ vanishes for all $v \in \mathbb{P}_1$. We remark here that in stead of finding the estimates for each individual elements, a master element is first constructed so that simplest affine transformations can be easily found to map to the individual elements. The local basis functions which play a crucial role in the error analysis can be easily derived for the master element. The Bramble-Hilbert Lemma is then applied only to the transformed problems on the master element.

We now return to the general case of a domain Ω in $\mathbb{R}^d$ and assume that we are given a family $\{V_h\}$ of finite-dimensional subspaces of $H_0^1(\Omega)$ such that, for some integer $r \geq 2$ and small h,

$$\inf_{\chi \in V_h} \{\|v - \chi\| + h\|\nabla(v - \chi)\|\} \leq Ch^s \|v\|_s, \text{for } 1 \leq s \leq r,\ v \in H^s \cap H_0^1. \quad (4.17)$$

The above example of piecewise linear functions corresponds to $d = r = 2$. In the case $r > 2, V_h$ most often consists of C^0-piecewise polynomials of degree $r - 1$ on a triangulation $\mathcal{T}_h$ as above. For instance, when $r = 4$, it is a case of piecewise cubic polynomial subspaces. In the general situation, estimates such as (4.17) with $s > d/2$ may often be obtained by exhibiting an interpolation operator I_h into V_h such that

$$\|I_h v - v\| + h\|\nabla(I_h v - v)\| \leq Ch^s \|v\|_s, \text{ for } 1 \leq s \leq r. \quad (4.18)$$

Note that in case $s \leq d/2$, the point-values of v are not well defined for $v \in H^s$ and the construction has to be somewhat modified, see Brennan and Scott [3]. When the domain Ω is curved and $r > 2$ there are difficulties near the boundary. However, the above situation may be accomplished by mapping a curved triangle onto a straight-edged one (isoparametric elements). But we shall not pursue it further and interested reader may refer to Ciarlet [4].

For our subsequent use, we remark that if the family of triangulations $\mathcal{T}_h$ is quasiuniform and V_h consists of piecewise polynomials, then one has the the following "inverse" inequality

$$\|\nabla\chi\| \leq Ch^{-1}\|\chi\|,\ \forall \chi \in V_h. \quad (4.19)$$

This inequality follows easily from the corresponding inequality for each triangle $K \in \mathcal{T}_h$, which in turn is obtained by a transformation to a fixed reference triangle, and using the fact that all norms on a finite dimensional space are equivalent, see, e.g., [4].

The optimal orders to which functions and their gradients may be approximated under our assumption (4.18) are $O(h^r)$ and $O(h^{r-1})$, respectively, and we shall try to obtain approximations of these orders for the solution of the Dirichlet problem.

Galerkin procedure. With V_h given as above, the Galerkin approximation $u_h \in V_h$ is determined as a function in V_h which satisfies

$$a(u_h, \chi) = (f, \chi), \ \ \forall \chi \in V_h. \tag{4.20}$$

Since $V_h \subset H_0^1$, the error $e = u - u_h$ satisfies the following Galerkin orthogonality condition

$$a(e, \chi) = a(u, \chi) - a(u_h, \chi) = (f, \chi) - (f, \chi) = 0 \ \ \forall \ \chi \in V_h.$$

Using a basis $\{\Phi_j\}_1^{N_h}$ for V_h, the problem (4.20) may be stated as : Find the coefficients α_i's in

$$u_h(x) = \sum_{i=1}^{N_h} \alpha_i \Phi_i(x),$$

such that

$$\sum_{i=1}^{N_h} \alpha_i [(\nabla \Phi_i, \nabla \Phi_j) + (\Phi_i, \Phi_j)] = (f, \Phi_j), \quad j = 1, ..., N_h.$$

In vector matrix form, this may be expressed as

$$A\alpha = F,$$

where $A = (a_{ij})$ is the so called global stiffness matrix, with $a_{ij} = (\nabla \Phi_i, \nabla \Phi_j) + (\Phi_i, \Phi_j)$, $F = (f_j)$ the global load vector with entries $f_j = (f, \Phi_j)$, and α the vector of unknowns α_i. The dimension of these arrays equals N_h, the dimension of V_h. Since the stiffness matrix A is positive definite, it is invertible, and hence, the discrete problem has a unique solution.

When V_h consists of piecewise polynomial functions, the elements of the matrix A may be calculated exactly. However, unless f has a particularly simple form, the elements (f, Φ_j) of F have to be computed by some quadrature formula.

We now prove the following estimate for the error $e = u - u_h$ between the exact or true solution of (4.16) and the solution of the discrete problem (4.20).

Theorem 4.6 *Let u_h and u be the solutions of (4.20) and (4.16), respectively. Then, for $1 \leq s \leq r$, the error $e = (u - u_h)$ satisfies*

$$\|e\| \leq Ch^s \|u\|_s,$$

and

$$\|\nabla e\| \leq Ch^{s-1} \|u\|_s.$$

Proof. In order to apply the Cea's Lemma, we note that

$$a(u,u) = \|\nabla u\|^2 + \|u\|^2 \geq \|\nabla u\|^2$$

is coercive in H_0^1-norm and also for $u, v \in H_0^1(\Omega)$

$$|a(u,v)| \leq M\|\nabla u\|\|\nabla v\|$$

is bounded. Here, we have used the Poincaré inequality. Now an appeal to Cea's Lemma yields

$$\|\nabla e\| = \|\nabla(u - u_h)\| \leq \inf_{\chi \in V_h} \|\nabla(u - \chi)\| \leq Ch^{s-1}\|u\|_s.$$

For the estimate in L_2 norm, we proceed by Aubin-Nitsche duality arguments. Let ϕ be an arbitrary in L_2, and $\psi \epsilon H^2 \cap H_0^1$ as the solution of

$$\begin{aligned} -\Delta\psi + \psi &= \phi, \text{ in } \Omega, \\ \psi &= 0, \text{ on } \partial\Omega, \end{aligned}$$

and recall the elliptic regularity condition

$$\|\psi\|_2 \leq C\|\phi\|. \tag{4.21}$$

Then,

$$(e,\phi) = (u - u_h, \Delta\psi + \psi) = a((u - u_h), \psi).$$

Using Galerkin orthogonality for $\chi \in V_h$, we have

$$(e,\phi) = a(e, (\psi - \chi)) \leq C\|\nabla e\| \cdot \|\nabla(\psi - \chi)\|,$$

and hence, using the approximation property (4.18) with $s = 2$, and the elliptic regularity (4.21), we obtain

$$(e,\phi) \leq Ch^{s-1}\|u\|_s h\|\psi\|_2 \leq Ch^s\|u\|_s\|\phi\|,$$

which completes the rest of the proof if we choose $\phi = e$.

Becasue of the variational formulation of the Galerkin method, the natural error estimates are expressed in L^2- based norms. Error analyses in other norms are also discussed in the literature. For example, we state, without proof, the following maximum-norm error estimate for C^0- piecewise linear elements. For a proof, we refer to Brenner and Scott [3].

Theorem 4.7 *Assume that V_h consists of C^0-piecewise linear polynomials and that the triangulations $\mathcal{T}_h$ are quasiuniform. Let u_h and u be the solutions of (4.20) and (4.16), respectively. Then the error e in the maximum norm satisfies*

$$\|e\|_{L^\infty(\Omega)} \leq Ch^2 \log\frac{1}{h}\|u\|_{W^{2,\infty}(\Omega)}.$$

4.3 Computational Issues

Although we shall discuss the formation of finite element equations only for the two examples discussed above, but the main theme remains the same for the higher order elements applied to different elliptic boundary value problems. We hope to being out the major steps leading to the linear algebraic equations.
Example 4.1 (Revisited). For C^0-linear elements, the global basis functions are givem by those hat functions described on section 4.2. Since we have imposed the boundary condition on the finite element sapce V_h, we obtain the global stiffness or Grammian matrix $A = [aij]_{1\le i,j\le N}$ with

$$a_{ij} = \int_0^1 [a(x)\phi_i'(x)\phi_j'(x) + a_0\phi_i\phi_j]\,dx = 0, \quad |i-j| > 1.$$

Therefore, each row has at most three non-zero entries,i.e., for $i = j-1, j, j+1$, the elements a_{ij} become

$$a_{ij} = \int_{x_{j-1}}^{x_{j+1}} [a(x)\phi_i'(x)\phi_j'(x) + a_0\phi_i\phi_j]\,dx.$$

On $[x_{j-1}, x_{j+1}]$, we have

$$\phi_j' = \begin{cases} \frac{1}{h_j}, & x_{j-1} \le x \le x_j, \\ -\frac{1}{h_{j+1}}, & x_j \le x \le x_{j+1}; \end{cases}$$

$$\phi_{j-1}' = \begin{cases} -\frac{1}{h_j}, & x_{j-1} \le x \le x_j \\ 0, & x_j < x < x_{j+1}, \end{cases}$$

and

$$\phi_{j+1}' = \begin{cases} 0, & x_{j-1} \le x \le x_j \\ \frac{1}{h_{j+1}}, & x_j \le x \le x_{j+1}. \end{cases}$$

Thus for $j = 2, \ldots, N-1$,

$$\begin{aligned} a_{j-1,j} &= \int_{x_{j-1}}^{x_j} a(x)\frac{1}{h_j^2}\,dx + \int_{x_{j-1}}^{x_j} a_0(x)(\frac{x_j - x}{h_j})(\frac{x - x_{j-1}}{h_j})\,dx \\ &= -\frac{1}{h_j^2}\int_{x_{j-1}}^{x_j} a(x)\,dx + \frac{1}{h_j^2}\int_{x_{j-1}}^{x_j} a_0(x)(x_j - x)(x - x_{j-1})\,dx, \end{aligned}$$

$$\begin{aligned} a_{j,j} &= \frac{1}{h_j^2}\int_{x_{j-1}}^{x_j} a(x)\,dx + \frac{1}{h_{j+1}^2}\int_{x_j}^{x_{j+1}} a(x)\,dx + \frac{1}{h_{j+1}^2}\int_{x_{j-1}}^{x_j} a_0(x)(x - x_{j-1})^2\,dx \\ &+ \frac{1}{h_{j+1}^2}\int_{x_j}^{x_{j+1}} a_0(x)(x_{j+1} - x)^2\,dx, \end{aligned}$$

and

$$a_{j+1,j} = \frac{-1}{h_{j+1}^2}\int_{x_j}^{x_{j+1}} a(x)\,dx + \frac{1}{h_{j+1}^2}\int_{x_j}^{x_{j+1}} a_0(x)(x_{j+1} - x)(x - x_j)\,dx.$$

When $j = 1$ and $j = N$. We have two entries like $a_{11},\ a_{21}$, and $a_{N-1,N},\ a_{N,N}$. For the right hand side, the load vector F is given by $[f_j]_{j=1,\ldots,N}$ with

$$\begin{aligned} f_j = (f, \phi_j) &= \int_0^1 f(x)\phi_j(x)\,dx = \int_{x_{j-1}}^{x_{j+1}} f(x)\phi_j(x)\,dx \\ &= \int_{x_{j-1}}^{x_j} f(x)(\frac{x - x_{j-1}}{h_j})\,dx + \int_{x_j}^{x_{j+1}} f(x)(\frac{x_{j+1} - x}{h_{j+1}})\,dx. \end{aligned}$$

Remarks.

(i) When a and a_0 are exactly integrable or a and a_0 are constants, we have, particularly, a simple form for the global stifness matrix A. In case $a = 1,\ a_0 = 0$ and $h = h_j,\ j{+}1, \ldots, N$, we obtain a triadiagonal matrix, which is similar to one derived by finite difference scheme (replacing the second derivative by central difference quotient)

(ii) For higher order elements say for C^0-piecewise cubic polinomials, it is difficult to obtain global basis functions ϕ_i's. But, on a master element, it is easy to construct local basis funcions. Therefore, the computation of the local stiffness matrix as well as the local load vector is done on this master element and then using affine transformations, these are obtained for each individual elements. However, the affine transformations are easy to construct and hence, make the computation of the local stiffness as well as the local load vector much simpler. Once, we have all the local sitffness matrices and local load vectors, we use assembly procedure by simply looking at the position of each element in the finite element mesh and then by adding contributions at the common vertex.

For the second examples, we shall, below, discuss the above mentioned issues.

Example 4.2 (Revisited). Since the supports of Φ_i's involve several elements in $\mathcal{T}_h$, it is difficult to derive these global basis functions and hence, it is not easy to compute the global stiffness matrix A. However, we may write

$$a(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \int_K (\nabla u_h \cdot \nabla v_h + u_h v_h)\,dx,$$

and

$$(f, v_h) = \sum_{K \in \mathcal{T}_h} \int_K f(x) v_h(x)\,dx.$$

On each element K, let $\{P_{iK}\}_{i=1}^3$ denote the threee vertices of K and let $\{\Phi_i^K\}$ denote the local basis functions with the property that $\Phi_i^K(P_{jK}) = \delta_{ij},\ 1 \le i, j \le 3$. Set

$$u_h^K = u_h\Big|_K = \sum_{i=1}^3 u_{iK}\Phi_i^K,$$

and $v_h^K = \Phi_j^K$, where $u_{iK} = u_h(P_{iK})$'s are unknows. On substitution for element K, we obtain

$$a(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \sum_{i,j=1}^e u_{iK} a_{ij}^K,$$

where

$$a_{ij}^K = \int_K (\nabla \Phi_i^K \cdot \nabla \Phi_j^K + \Phi_i^K \Phi_j^K)\, dx.$$

Note that on each element, we have the local stiffness matrix

$$A^K = (a_{ij}^K)_{1 \le i,j \le 3}.$$

Similary, the local vector $b^K = (b_j^K), 1 \le j \le 3$ is given by

$$b_j^K = \int_K f \Phi_j^K dx,\ 1 \le j \le 3.$$

Then, by looking at the position of the elements in the finite element mesh, i.e., by establishing a connection between the global node numbering and local node numbering (the relation between $\{P_j\}$'s and $\{P_{jK} : K \in \mathcal{T}_h,\ 1 \le j \le 3\}$, we assemble the contributions of the local stiffness matrices and the local load vectors to obtain the global stiffness matrix and the global load vector.

Assembling Algorithm.

- Set $A = 0$.
- For $K \in \mathcal{T}_h$ compute A^K and b^K.
- For $i, j = 1, 2, 3$, compute

$$a_{P_{jK}, P_{jK}} = a_{P_{iK}, P_{jK}} + a_{ij}^K$$

and

$$b_{P_{jK}} = b_{P_{jK}} + b^K.$$

Note that the contribution of a vertex, which is common to many elements is added up.

In general, computation of local basis function Φ_j^K may not be an easy task. Specially, for higher order elements like quadratic or cubic polynomial, it is cumbersome to find local basis functions. Therefore, one introduces a master element $\widehat{K}$ with vertices $(0,0), (1,0), (0,1)$ and the affine mapping $F : \widehat{K} \mapsto K,\ K \in \mathcal{T}_h$ which is of the form $F(\widehat{x}) = B\widehat{x} + d$ where, B is 2×2 matrix and $d \in \mathbb{R}^2$. Let $\widehat{u}_h(\widehat{x}) = u_h(F(\widehat{x}))$ and $\widehat{v}_h(\widehat{x}) = v_h(F(\widehat{x}))$ Then

$$\int_K u_h v_h dx = \int_{\widehat{K}} \widehat{u}_h \widehat{v}_h\, det(B)\, d\widehat{x}.$$

Further,

$$\nabla \widehat{u}_h \Big|_{\widehat{x}} = B^T \nabla u_h \Big|_{F(\widehat{x})},$$

where B^T = Jacobian of F. Thus,

$$\int_K \nabla u_h \cdot \nabla v_h dx = \int_{\widehat{K}} (B^{-T} \nabla \widehat{u})(B^{-T} \nabla \widehat{v}_h)\, det(B)\, d\widehat{x}.$$

On element $\widehat{K}$, it is easy to construct the local basis functions $\Phi_i^{\widehat{K}}$, that is, for the present case,

$$\Phi_i^{\widehat{K}} := \Phi_{(0,0)}^{\widehat{K}} = 1 - \widehat{x_1} - \widehat{x_2},\ \Phi_2^{\widehat{K}} := \Phi_{(1,0)}^{\widehat{K}} = \widehat{x_1},\ \Phi_3^{\widehat{K}} := \Phi_{(0,1)}^{\widehat{K}} = \widehat{x_2}.$$

Now, set

$$\widehat{u}_h(\widehat{x}) = \sum_{i=1}^{3} u_{iK}\Phi_i^{\widehat{K}} \text{ and } \widehat{v}_h = \Phi_j^{\widehat{K}}, j = 1, 2, 3,$$

and compute

$$\int_K \nabla u_h \cdot \nabla v_h \, dx = \sum_{i=1}^{3} \int_{\widehat{K}} u_{iK}(B^{-T}\nabla\Phi_i^{\widehat{K}})(B^{-T}\nabla\Phi_j^{\widehat{K}})\, det(B)\, d\widehat{x}.$$

Now, the computation only involves the evaluation of integrals like

$$\int_{\widehat{K}} \widehat{x_1}^i \widehat{x_2}^j d\widehat{x} = \frac{i!j!}{(i+j+2)!}.$$

Remarks.

(i) For simplicity of programming, the integrals are evaluated by using quadrature rules. One such simple rule is

$$\int_{\widehat{K}} f(\widehat{x})\, d\widehat{x} = \sum_{i=1}^{3} w_i f(P_i^{\widehat{K}}),$$

where w_i's are the quadrature weights. Then programming become easier as we have to calculate only the values of Φ_i^K and $\frac{\partial\Phi_i^{\widehat{K}}}{\partial\widehat{x_j}}$ at the vertices of $\widehat{K}$.

(ii) After performing assembly procedure, we can impose essential boundary conditions and then obtain a reduced nonsingular system

$$A\alpha = b.$$

The global stiffness matrix being sparse and large, one may use iterative procedures to solve the system. However, there are direct methods (elimination methods) like *frontal* and *skyline* methods which are used effectively along with the assembling procedure. But, we shall not dwell on this and refer the readers to Johnson [8].

Computation of Order of Convergence. The *a priori* error bound (the bound obtained *prior* to the actual computation) in section 4.2 dependes on the unknown solution, provided the exact solution has the required regularity properties. However, it tells us that the approximate solution converges as the mesh is refined. Therefore, we have confidence on the computed solution. Being an asymptotic estimate, i.e., u_h converges to the exact solution u in appropriate norm as $h \mapsto 0$, it may happen that after some refinements that is as $h \leq h_0$ for

some h_0, the approximate solution is close to the true solution and this h_0 is so small that the computers may consider it to be zero (otherwise very small h_0 may give rise to a large system of equations which is virtually difficult to solve). In such situations, the *a priori* error estimate looses its meaning.

Veryoften, the user community is confronted with the question

" *How does one get confidence on the computed solution ?*"

It is costumary to check the computed solutions for various decreasing mesh sizes. If the trend is such that up to some decimal places, the solutions for various mesh sizes remain same, then one has confidence on the computed numbers. But, this may go wrong! Therefore, alongwith this trend if one computes the order of convergence that matches with the theoretical order of convergence, then this may be a better way of ensuing the confidence. To compute the order of convergence, we note that for h_1 and h_2 with $0 < h_2 < h_1$

$$\|u - u_{h_1}\| \approx C(u)h_1^\alpha,$$

and

$$\|u - u_{h_2}\| \approx C(u)h_2^\alpha.$$

Then the order of convergence α is computed as

$$\alpha \approx \left(\log \frac{\|u - u_{h_1}\|}{\|u - u_{h_2}\|}\right) / \left(\log \frac{h_1}{h_2}\right).$$

In the abscence of the exact solution, we may replace u by a more refined computed solution.

4.4 Adaptive Methods.

Very often the user is confronted with the following problem

"*Given a tolerance (TOL) and a measurement say a norm $\|\cdot\|$, how to find an approximate solution u_h to the exact unknown solution u with minimal computational effort so that $\|u - u_h\| \leq TOL$?* "

For minimal computational effort, we mean that the computational mesh is not overly refined for the given accuracy. This is an issue in Adaptive methods. It is expected that successful computational algorithms will save substantial computational work for a given accuracy. These codes are now becoming a standard feature of the finite element software.

An adaptive algorithm consists of

- *a stopping criterion* which guarantees the error control within the preassigned error tolerance level,
- *a mesh modification strategy* in case the stopping criterion is not satisfied.

It is mostly based on sharp *a posteriori* error estimate in which the error is evaluated in terms of the computed solutions and the given data and *a priori* error estimate which predicts the error in terms of the exact (unknown) solutions

(the one which we have been discussing uptill now). While the stopping criterion is built solely on the *a posteriori* error estimate, the mesh modification strategy may depend on *a priori* error estimates.

A posteriori error estimates. We shall discuss *a posteriori* error estimates first for the example 4.1 with $a_0 = 0$ (for simplicity) and then take up the second example.

In order to measure the size of the error $e = u - u_h$, we shall use the following energy norm

$$\|v\|_E \equiv \|v'\|_a := \left(\int_0^1 a(x)|v'(x)|^2 \, dx \right)^{1/2}.$$

Since $a(x) \geq \alpha_0 > 0$ and v vanishes at $x = 0$, the above definition makes sense as a weighted norm. The *a priori* error estimate discussed in section 4.2 depends on the second derivative of the unknown solution u, where as the *a posteriori* error estimate, we shall describe below depends on the given data and the computed solution u_h.

Theorem 4.8 *Let u be the exact solution of (4.7) with $a_0 = 0$ and the corresponding Galerkin approximation be denoted by u_h. Then, the following a posteriori error estimate holds*

$$\|e\|_E \equiv \|u' - u_h'\|_a \leq C_i \|hR(u_h)\|_{\frac{1}{a}},$$

where $R(u_h)$, the residual on each interval I_i is defined as $R(u_h)|_{I_i} := f + (au_h')'$ and C_i is an interpolation constant depending on the maximum of the ratios $(\max_{I_i} a / \min_{I_i} a)$.

Proof. Note that using Galerkin orthogonality, we have

$$\begin{aligned} \|e'\|_a^2 &= \int_0^1 ae'(e - I_h e)' \, dx \\ &= \int_0^1 au'(e - I_h e)' \, dx - \int_0^1 au_h'(e - I_h e)' \, dx \\ &= \int_0^1 f(e - I_h e) \, dx - \sum_{i=1}^N \int_{I_i} au_h'(e - I_h e)' \, dx. \end{aligned}$$

Now integrate by parts the last term over each interval I_i and use the fact that $(e - I_h e)(x_i) = 0, \;\; i = 0, 1, \cdots, N$ to obtain

$$\begin{aligned} \|e'\|_a^2 &= \sum_{i=1}^N \int_{I_i} [f + (au_h')'](e - I_h e) \, dx \\ &= \int_0^1 R(u_h)(e - I_h e) \, dx \\ &\leq \|hR(u_h)\|_{\frac{1}{a}} \|h^{-1}(e - I_h e)\|_a. \end{aligned}$$

A slight modification of Theorem 4.2 yields an interpolation error in the weighted L^2- norm and thus, we obtain

$$\|h^{-1}(e - I_h e)\|_a \leq C_i \|e'\|_a,$$

where the interpolation constant C_i depends on the maximum of the ratios of $(\max_{I_i} a / \min_{I_i} a)$.

For the automatic control of error in the energy norm , we shall use the above *a posteriori* eror estimate and shall examine below an adaptive algorithm.

Algorithm.

- Chose an initial mesh $\mathcal{T}_{h^{(0)}}$ of the mesh size $h^{(0)}$.
- Compute the finite element solution $u_{h^{(0)}}$ in $V_{h^{(0)}}$.
- Given $u_{h^{(m)}}$ in $V_{h^{(m)}}$ on a mesh with mesh size $h^{(m)}$,
  - **Stop**, if
  $$C_i \|h^{(m)} R(u_{h^{(m)}})\|_{\frac{1}{a}} \leq TOL,$$
  - **otherwise** determine a new mesh $\mathcal{T}_{h^{(m)}}$ with maximal mesh size $h^{(m+1)}$ such that
  $$C_i \|h^{(m)} R(u_{h^{(m)}})\|_{\frac{1}{a}} = TOL. \tag{4.22}$$
- Continue.

By Theorem 4.6, the error is controlled to the TOL in the energy norm, if the stopping criterion is reached with solution $u_h = u_{h^{(m)}}$, otherwise, a new maximal mesh size is determined by solving (4.22). However, the maximality is obtained by the equidistribution of errors so that the contributions of error from each individual elements I_i are kept equal, i.e., the equidistribution give rise to

$$[a(x_i)]^{-1}(h_i^{(m)} R(u_{h^{(m-1)}}))^2 h_i^{(m)} = \frac{(TOL)^2}{N^{(m)}}, \;\; i = 1, 2, \cdots, N^{(m)},$$

where $N^{(m)}$ is the number of intervals in $\mathcal{T}_{h^{(m)}}$. In practice , the above nonlinear equation is simplified by replacing $N^{(m)}$ by $N^{(m-1)}$.

With slight modification of our *a priori* eror estimate in section 4.2, we may rewrite

$$\|u - u_h\|_E = \|u' - u_h'\|_a \leq \|(u - I_h u)'\|_a \leq C_i \|hu''\|_a.$$

Now it is a routine excercise to check that

$$\|hR(u_h)\|_{\frac{1}{a}} \leq CC_i \|hu''\|_a,$$

where $C \approx 1$. This indicates that the *a posteriori* eror estimate is optimal in the same sense as the *a priori* estimate.

As a second example, we recall the Poisson equation with homogeneous Dirichlet boundary condition

$$-\Delta u = f, \quad x \in \Omega \,\text{and}\; u = 0, \;\; x \in \partial\Omega.$$

For applying the adaptive methods, we shall define a mesh function $h(x)$ as $h(x) = h_K, \;\; x \in K$,where K is a triangle belonging to the family of triangulations $\mathcal{T}_h$, see the section 3 for this. With V_h consisting of continuous piecewise linear functions that vanish on the boundary $\partial\Omega$, the Galerkin orthogonality relation takes the form

$$(\nabla(u - u_h), \chi) = 0 \quad \forall \chi \in V_h.$$

We shall use the following interpolation error which can be obtained with some modifications of the standard interpolation error (see, Eriksson *et al.* [6]).

Lemma 4.9 *There is a positive constant C_i depending on the minimum angle condition such that the piecewise interpolant $I_h v$ in V_h of v satisfies*

$$\|h^r D^s (v - I_h v)\| \leq C_i \|h^{2+r-s} D^2 v\|, \;\; r = 0, 1, \quad s = 0, 1.$$

Now , we shall derive the *a posteriori* error estimate in the energy norm.

Theorem 4.10 *The Galerkin approximation u_h satisfies the following a posteriori error estimate for $e = u - u_h$*

$$\|\nabla(u - u_h)\| \leq C_i \|hR(u_h)\|,$$

Where $R(u_h) = R_1(u_h) + R_2(u_h)$ with

$$R(u_h)|_K = |f + \Delta u_h| \quad \text{on}\, K \in \mathcal{T}_h,$$

and

$$R_2(u_h)|_K = \max_{S \subset \partial K} |[\frac{\partial u_h}{\partial \nu_S}]|.$$

Here $[\frac{\partial u_h}{\partial \nu_S}]$ represents the jump in the normal derivative of the function u_h in V_h.

Proof. Using Galerkin orthogonality, we have

$$\begin{aligned} \|\nabla e\|^2 &= (\nabla(u - u_h), \nabla e) = (\nabla u, \nabla e) - (\nabla u_h, \nabla e) \\ &= (f, e) - (\nabla u_h, \nabla e) = (f, e - I_h e) - (\nabla u_h, \nabla(e - I_h e)). \end{aligned}$$

On integrating by parts the last term over each triangle K , it follows thar

$$\begin{aligned} \|\nabla e\|^2 &= \sum_{K \in \mathcal{T}_h} \int_K (f + \Delta u_h)(e - I_h e)\, dx + \sum_{K \in \mathcal{T}_h} \int_{\partial K} \frac{\partial u_h}{\partial \nu_K}(e - I_h e)\, ds \\ &= \sum_{K \in \mathcal{T}_h} \int_K (f + \Delta u_h)(e - I_h e)\, dx + \sum_{S \in \mathcal{T}_h} \int_S h_S^{-1} \frac{\partial u_h}{\partial \nu_K}(e - I_h e) h_S\, ds. \end{aligned}$$

An application of interpolation inequality completes the rest of the proof.

Note that if all the triangles are more or less isoscale and the trianglulation satisfies the minimum angle condition, then $C_i \approx 1$. Compared to one dimensional situation, in the present case we have a contribution fron the element sides S involving the jump in the normal derivative of u_h divided by the local mesh size that is R_2. In one dimensional case this term is zero, because the interpolation error vanishes at the nodal points.

Bibliography

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, New York, 1965.

[3] K. E. Brennan and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer Verlag, Berlin, 1994.

[4] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North Holland, Amsterdam, 1978.

[5] R. Dautray, and J. L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 1–6, Springer Verlag, Berlin, 1993.

[6] K. Eriksson, and C. Johnson, *Adaptive finite element methods for parabolic problems. I: A linear Model problem* SIAM J. Numer. Anal., 43–77, 1991.

[7] K. Eriksson, D. Estep, P. Hanso, and C. Johnson, *Introduction to daptive finite element methods for differential equations* Acta Numerica , 105–158, 1995.

[8] C. Johnson, *Numerical Solution of Partial Differential Equations by Finite Element Methods* Cambridge University Press, Cambridge, 1987.

[9] S. Kesavan *Topics in Functional Analysis and Applications*, Wiley Eastern Ltd. , New Delhi, 1989.

[10] B. Mercier, *Lectures on Topics in Finite Element Solution of Elliptic Problems*, TIFR Lectures on Mathematics, Vol. 63, Narosa Publi., New Delhi, 1979.

[11] R. Temam, *Navier Stokes Equation* Studies in Mathematics and Its Applications, Vol. 2, North Holland, (Revised Edition) 1984.

[12] V. Thomée *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Mathematics No. 1054, Springer Verlag, 1984.

[13] V. Thomée *Lectures on Approximation of Parabolic Problems by Finite Elements*, Lecture Notes No. 2, Department of Mathematics, Indian Institute of Technology, Bombay (India), 1994.

[14] M. F . Wheeler *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10, 723–759, 1973.