Hebron University

Faculty of Graduate studies

Mathematics Department

Finite Element Method for Elliptic Differential Equations

By

Alaa Taniniah

Supervisor

Dr. Hasan Almanasreh

This Thesis is Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Mathematics, Faculty of Graduate Studies, Hebron University, Hebron, Palestine.

Finite Element Method for Elliptic Differential Equation

By

Alaa Taniniah

Supervisor

Dr. Hasan Almanasreh

This thesis was defended successfully on 7/4/2021 and approved by:

Committee Members:		**Signature**
Dr. Hasan Almanasreh	Supervisor	....
Dr. Tareq Amro	Internal Examiner	....
Dr. Maher Qarawani	External Examiner	....

Dedication

I dedicate my thesis to my sons, parents, brothers, sisters and teachers who supported me on each step of the way.

Acknowledgment

I would like to express my sincere gratitude to my advisor Dr. Hasan Almanasreh for the continuous support of my master study and research, for his patience , motivation, enthusiasm and immense knowledge. His guidance helped me in all time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my master study.

Besides my advisor, I would like to thank doctors in the mathematics department at Hebron University. In particular my sincere thank to my thesis examiners, Dr. Tareq Amro and Dr. Maher Qarawani, for their encouragements, insightful comments and hard questions. Without their passionate participation and input, the validation survey could not have been successfully conducted.

Any attempt at any level can not be satisfactory completed without the support and guidance of my parent and friends.

Last but not the least, I would like to thank my family, my parents and my friends in particular Dr. Kamel Jebreen Who workes at Unité de Recherche Clinique (URC) Saint-Louis Lariboisière Fernand-Widal, Assistance Publique – Hôpitaux de Paris, and Abed Al-Qader Halman, who helped me a lot in gathering information, collecting data, and guiding me from time to time in making this study despite of their schedules. They gave me different ideas in making this project unique

الإقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان

طريقة العناصر المتهية لحل المعادلات التفاضلية الناقصة

FINITE ELEMENT METHOD FOR ELLIPTIC DIFFERENTIAL EQUATIONS

أقر بأن ما إشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص،
باستثناء ما تم الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل لم تقدم
من قبل لنيل أي درجة علمية أو بحث علمي أو بحث لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

The work provided in this thesis, unless otherwise referenced, is the result of the researcher's work, and has not been submitted elsewhere for any other degree or qualification.

Abstract

In this thesis we will discuss some basic and general theory of the finite element method. We will also discuss the variational formulation and discretization in order to assess the amount of error in the approximate solution applied to the space segmentation into triangles. For this purpose, first we are going to study the finite element method for second order elliptic problems in one and two dimensions and find a posteriori error estimates for Reaction-diffusion problems and Poisson equation. After that, we will review the modular solution method and the system of fragmentation of differential equations in different conditions on the limits of the definition range. Also illustrative examples will also be analyzed using the mathematical programming language 'Matlab'.

The a posteriori errors reviewed in this thesis are quantities that measure the rate of convergence of the numerical solutions of differential equations to the exact solution using a particular element method that can be estimated based on the approximate solution and the information available on differential equations. The advantage of the numerical errors of differential equations is to measure the size of the error in order to make it as small as possible and thus get the best approximation of the solution. To discuss these errors, there are basic concepts that will be addressed to explain, and then the errors will be reviewed in details for some partial differential equations.

الملخص

في هذا الرسالة سوف ندرس طريقة العناصر المحدودة العددية للمعادلات القطعية الناقصة ونتناول الأخطاء البعدية الناتجة من استخدامها وذلك لتقييم مقدار الخطأ في الحل التقريبي المطبق على تجزأة المجال . لهذا الغرض أولا سنناقش طريقة العناصر المحددة للحلول العددية للمعادلات التفاضلية أحادية وثنائية البعد، وسوف نستعرض طريقة الحل التبايني ونظام التجزأة للمعادلات التفاضلية بمختلف الشروط على حدود مجال التعريف. أيضا امثلة توضيحيه سوف يتم دراستها باستخدام لغة البرمجة الرياضية الماثلاب.

الأخطاء التي تتناولها هذه الرسالة هي كميات تقيس الأخطاء الناتجة من الحلول العددية للمعادلات التفاضلية باستخدام طريقة العناصر المحدودة التي يمكن تقديرها بالاعتماد على الحل التقريبي والمعلومات المتوفرة عن المعادلات التفاضلية. تتلخص فائدة الاخطاء البعدية للحلول العددية للمعادلات التفاضلية في معرفة مقدار الخطأ بالحل التقريبي وذلك لجعلة أقل ما يمكن و بالتالي الحصول على أفضل تقريب للحل . ولمناقشة هذة الأخطاء البعدية، هناك مفاهيم أساسية سيتم تناولها لتوضيحها أولا وبعد ذلك سيتم معالجتها بالتفصيل لعدد من المعادلات التفاضلية الجزئية الآنفة الذكر.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Finite Element Method (FEM) is a computational technique for solving problems defined by partial differential equations that occur in scientific and engineering applications [Wu (1996)] and [Ern and Guermond (2004)]. The FEM uses a variational form of the problem that involves an integral form of the differential equation over a given domain where this domain is divided into a number of subdomains called finite elements, or use the minimization method (Ritz method) that is equivalent to a variational form [Sun and Zhou (2016)] and [Braess and Verfürth (1996)].

We are interested in the existence and a posteriori estimations of weak solutions for linear elliptical differential equations. These problems arise in a variety of situations in biology, chemistry, or physics,$\cdots$ etc [Braess (2007a)] and [Hackbusch (2017)], [GIDAS (1981)] and [Grätsch and Bathe (2005)]. The goal of this thesis is to study the finite element method for second-order elliptic problems in one and two dimensions and to find a posteriori error estimates for reaction-diffusion problems and Poisson equations.

In this thesis, the Sobolev spaces that are used in the variational formulation of differential equations and some other required concepts are defined. The back ground of FEM, the classifications of the differential equations to elliptic, hyperbolic and parabolic, and according to the boundary conditions, Dirichlet, Neumann, Mix, and Robin problems are explained. We formulate the general theorems for existence and uniqueness in the Hilbert space context and state the conditions that spaces and bilinear form can satisfy [Larsson and Thomee (2003)] and [Gander and Kwok (2018)]. These results are used to investigate the solvability of particular partial differential equations.

The core of this work starts with the discussion of the variational (weak) formulation and the discretization of the problem with homogeneous and nonhomogeneous boundary conditions [Braess (2007b)] and [GIDAS (1981)]. We construct a variational formulation by multiplying the two sides of the differential equation by a test function $v(x) \in V$, V is some Sobolve space, and then integrate over a specified domain. The purpose of introducing a notation of weak wording is to provide access to the nature and uniqueness of solutions that are well suited to the numerical approximation of such problems [Houston and Süli (2001)] and [Yu and Zhao (2005)]. In the discretization process we create a finite dimensional space V_h of continuous linear functions on the partition $\mathcal{T}_h$, and find $u_h \in V_h$ that satisfies the variational formulation. Then we analyze the error calculation which is the difference between the approximate solution u_h and the exact solution u. Both types of error are a priori error and a posteriori estimates. The first type error bounds given by known information on the solution of the variational problem and the finite element function space, where the second type is error bounds given by information on the numerical solution obtained on the finite element function space. Two types of problems are studied: Reaction-Diffusion and Poisson Problems, where the main task is to discuss the a periori and a posteriori error estimates for these problems [Thomas et al. (2019)] and [Zhang and Yan (2001)].

The reaction-diffusion problem naturally occurs in systems consisting of several components interacting as chemical reactions [Brezis and Turner (1977)], and is widely used to explain pattern-forming phenomena in a number of biological [Courant (1943)], chemical and physical systems. The typical form is as follows:

$$-\omega\Delta u + cu = h(x), \quad x \in \Omega.$$

The Poisson equation as the model problem for elliptic partial differential equation. It arises, $e.g.$, in structural mechanics, theoretical physics as gravitation, electromagnetism, elasticity and in many other areas of science and engineering. The Poisson problem is defined as:

$$-\Delta u = h(x), \quad x \in \Omega.$$

This project consists of Four chapters. Chapter Two will be about the FEM in general Chapter Three talks about the variational formulation and discretization of differential equation. In Chapter Four we will explain the error estimation in its both types, a priori and a posteriori, for reaction-diffusion problem and Poisson equation.

Chapter 2

Differential equations and the FEM

The finite element method [Bathe (2014) and Izadi (2007)] is a numerical method for solving problems of engineering and mathematical physics. Typical problem areas of interest include structural analysis heat transfer, fluid flow, mass transports and electromagnetic potential.The finite element method formulation of the problem result in a system of algebraic equation. The method approximates the unknown function over the domain, that is divided into smaller parts called finite element. The simple equations that model these finite elements are then assembled into a large system of equations that models the entire problem. The FEM uses variational method from the calculus of variation to approximate a solution by minimizing an associated error function [Saad (2003)], [Ciarlet (2002)] and Wu (1996)].

The Finite Element Method is a numerical technique to find approximate solutions of differential equations. It was originated from the need of solving complex elasticity and structural analysis problems in Civil, Mechanical and Aerospace engineering.

2.1 History of the analysis of the finite element

- The finite element method was first proposed in 1909 to Ritz [Bathe (2014)], [Evans (2010)] and [Izadi (2007)], who developed an efficient method for approximate problem solving [Zeidler (2007)], which involves approximating the power function through known functions with unknown parameters.
- The study of the finite element can be traced back to the works of Alexander

Hrennikoff 1941 and Richard Currant 1942. Hrenikoff has created a frame method in which a flat, flexible medium is interpreted as a set of rails and girders. These pioneers share one important characteristic: the division of a continuous domain into a number of distinct subdomains, typically called elements.

- In 1943 German mathematician Richard Currant increased the probabilities of the Ritz method by introducing special linear functions defined via multiple-definition linear approximation in subareas [Kuo and Trudinger (1992)], and using the finite element model of the procedure to reduce the potential energy of the torsion strain function using values Grid point as unknown parameters.
- In 1950, solving a large number of equations simultaneously became possible with a digital computer.
- Ray W. Clough first published a paper in 1960 using the word "*Finite Element Method*".
- The first conference on "finite elements" was held at a price of US 1965.
- Zienkiewicz and Chung wrote their first book on "Operation Unique Elements" in 1967.
- In the late 1960 and early 1970, FEM was applied to a variety of engineering issues.
- Most commercial FEM software packages (ABAQUS, NASTRAN, ANSYS, etc.) appeared at 1970. Interactive finite element software on supercomputers is contributing to the rapid growth of CAD systems.
- In 1980 an algorithm was developed for electromagnetic, fluid flow, and thermal analysis applications using the finite element method.
- Engineers can analyze methods to manage vibration and extend the use of diversity, and to accelerate space structures using a finite and other methods of 1990. Trends to overcome additive solution to fluid flow are closely related to structural reactions and biomechanical problems. A higher degree of accuracy was observed in this decade [Langtangen and Mardal (2019)].

2.2 Advantages and disadvantages of the FEM

Advantages of the FEM:

1. Can comfortably manage the extremely complex geometry.
2. Can manage a variety of engineering problems (solid mechanics, Fluid, Dynamics, Electrostatic problems, Heat problems).
3. Can manage dynamic constraints (an undetermined structure can be resolved).

Disadvantages of the FEM:

(1) A general closed-form solution that would allow a system response to a change in different parameters to be examined is not generated.

(2) The FEM just obtain a "approximate" solution.

(3) The FEM has " an Internet " mistake [Gaeta and Rodríguez (2017)].

2.3 Sobolev spaces

Definition 2.3.1. L_p-spaces, For $p \in [1, \infty)$,

$$L_p(\Omega) : = \Big\{ v : \Omega \to \mathbb{R} \text{ measurable and } \int_\Omega |v(x)|^p \, dx < \infty \Big\}. \tag{2.1}$$

$$\|v\|_{L_p(\Omega)} := \Big(\int_\Omega |v(x)|^p \, dx \Big)^{\frac{1}{p}}. \tag{2.2}$$

For $p = \infty$,

$$L_\infty(\Omega) := \Big\{ v : \Omega \to \mathbb{R} \text{ measurable and } |v(x)| < \infty \; a.e. \Big\}$$

$$\|v\|_{L_\infty(\Omega)} := \inf \Big\{ k > 0, \; |v(x)| \le k \;\; a.e. \Big\}$$

The integral (2.2) is called Lebesgue integral and "a.e" means "almost every where" [Quarteroni and Valli (2008)], i.e. $\forall x \in \Omega \backslash \mathbb{N}$, for null sets $\mathbb{N}$.

Important properties

1. Banach space is $(L_p(\Omega), \| \cdot \|_{L_p})$, for $p \in \mathbb{N}$.
2. The space $(L_2(\Omega), \langle \cdot \rangle_{L_2(\Omega)})$ is a Hilbert space [Braess (2007a)], where the inner product in L_2 is defined as

$$\langle \varphi, \psi \rangle_{L_2(\Omega)} = \int\limits_\Omega \varphi(x) \, \psi(x) \, dx.$$

Notation 1. The space C_c^∞ denoted the infinitely differentiable space functions $\psi : \Omega \to \mathbb{R}$ with compact support in Ω, the function $\psi \in C_c^\infty(\Omega)$ is called a test function [Evans (2010)].

Definition 2.3.2. Assume a function $u \in C^1(\Omega)$. If $\psi \in C_c^\infty$ we give the formula of integration by parts

$$\int_{\Omega} u\psi_{x_i}\, dx = -\int_{\Omega} u_{x_i}\psi\, dx \tag{2.3}$$

there is no boundary term since ψ is with compact support in Ω.If k is a positive integer, $u \in C^k(\Omega)$, and $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_d)$ is a multi-index of order $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d = k$, then

$$\int_{\Omega} u\, D^{\alpha}\psi\, dx = (-1)^{|\alpha|} \int_{\Omega} D^{\alpha}\, u\, \psi\, dx, \quad \forall \psi \in C_c^k(\Omega),$$

where

$$D^{\alpha}\psi = \frac{\partial^{\alpha_1}}{\partial {x_1}^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial {x_n}^{\alpha_n}}\,(\psi).$$

Remark 1. Given a domain Ω, a set of locally integrable functions is defined by [Brenner and Scott (2008)]

$$L^1_{loc}(\Omega) := \left\{ g:\ g \in L^1(\Gamma), \quad \forall\, \text{compact}(\Gamma) \subset \text{interior}(\Omega) \right\}.$$

2.3.1 (Weak derivative)

Suppose $f, g \in L^1_{loc}(\Omega)$ and α is a multi-index, we say that g is the $\alpha^{th}-$ weak partial derivative of f, written $g = D^{\alpha} f$, if [Evans (2010)] and [Brenner and Scott (2008)]

$$\int_{\Omega} f D^{\alpha}\psi\, dx = (-1)^{|\alpha|} \int_{\Omega} g\psi\, dx, \quad \forall \psi \in C_c^{\infty}(\Omega),$$

or equivalently

$$\langle f, D^{\alpha}\psi\rangle_{L_2(\Omega)} = (-1)^{|\alpha|} \langle g, \psi\rangle_{L_2(\Omega)}, \quad \forall \psi \in C_c^{\infty}(\Omega).$$

Definition 2.3.3. Given a function $g \in L^1_{loc}(\Omega)$ we say that $h \in L^1_{loc}(\Omega)$ has a weak derivative $D^{\alpha}h$ if

$$\int_{\Omega} h(x) D^{\alpha}\psi(x)\, dx = (-1)^{|\alpha|} \int_{\Omega} D^{\alpha}h(x)\psi(x)\, dx, \quad \forall\, \psi \in C_c^{\infty}(\Omega).$$

Remark 2.

- **Uniqueness** : If a locally integrable function has a weak derivative, then it is unique, *i.e.*, if $v = D^\alpha u \in L^1_{loc}(\Omega)$ and $\tilde{v} = D^\alpha u \in L^1_{loc}(\Omega)$ both are weak partial derivatives of u, then $v = \tilde{v}$ *a.e.*,[Wait (1631)].

- **Consistency in the definition:** If $u \in C^1(\Omega) \bigcap C(\overline{\Omega})$, then the weak derivative matches the classical derivative, [Braess (2007a)].

Definition 2.3.4. Let k be a non-negative integer, and let $\psi \in L^1_{loc}$ be assumed to have a weak derivative $D^\alpha(\psi)$ for all $|\alpha| \leq k$, [Larson and Bengzon (2013)]. Define the sobolev space W^k_p

$$W^k_p := \Big\{ \psi \in L^1_{loc} : \ \|\psi\|_{W^k_p} < \infty \Big\},$$

where for $1 \leq p < \infty$,

$$\|\psi\|_{W^k_p(\Omega)} := \Big(\sum_{|\alpha| \leq k} \|D^\alpha \psi\|^p_{L_p(\Omega)} \Big)^{\frac{1}{p}},$$

and for $p = \infty$,

$$\|\psi\|_{W^k_\infty(\Omega)} := \max_{|\alpha| \leq k} \|D^\alpha \psi\|_{L_\infty(\Omega)}.$$

Remark 3.

(1) If $p = 2$ we usually write

$$W^k_2(\Omega) = H^k(\Omega) = \Big\{ \psi \in L_2(\Omega) : \sum_{|\alpha| \leq k} D^\alpha \psi \in L_2(\Omega) \Big\}, \quad k = 0, 1, \cdots.$$

We use the letter H since as will as $H^k(\Omega)$ is Hilbert space with the inner product.

$$\langle u, v \rangle_{W^k_2} = \sum_{|\alpha| \leq k} \langle D^\alpha u, D^\alpha v \rangle.$$

(2) The special case when $k = 1$ and $p = 2$ the space is

$$H^1 = \Big\{ \psi \in L_2 : \frac{\partial \psi}{\partial x_i} \in L_2, \ i = 1, \cdots, n \Big\}. \tag{2.4}$$

Note that

$$\|\psi\|_{H^1(\Omega)} = \Big(\|\psi\|^2_{L^2(\Omega)} + \|D\psi\|^2_{L^2(\Omega)} \Big)^{\frac{1}{2}}$$

Definition 2.3.5. [Braess (2007a)] The Sobolev space H^k_0 is the completion of the C_c^∞ with respect to the norm $\|\cdot\|_{H^k}$, i.e.,

$$u \in H_0^k(\Omega) \Longleftrightarrow \exists\, v_n \in C_c^\infty(\Omega) \text{ such that } \lim_{n\to\infty} \|u - v_n\|_{H^k(\Omega)} = 0.$$

Note that $H_0^k(\Omega)$ is a closed subspace of H^k. If the boundary Γ is C^1, then it is assumed that $v \in C(\overline{\Omega}) \bigcap H_0^k(\Omega)$ implies that $v(x) = 0$ for all $x \in \Gamma$.
Finally, the special Sobolev space H_0^1, defined as the closure of C_0^∞ in $H^1(\Omega)$

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

Definition 2.3.6. Let $(V, (\,\cdot\,,\,\cdot\,))$ be an inner product space, if the associated normed linear space $(V, \|\cdot\|)$ is complete,then $(V, (\,\cdot\,,\,\cdot\,))$ is called a Hilbert space.

Notation 2. H_0^1 is a Hilbert space have the same norm and same inner product as H^1.

Theorem 2.3.1. [Larson and Bengzon (2013)]. The Sobolev space $H_p^k \equiv W_p^k$ with regard to the norm $\|\cdot\|_{H_p^k}$ is called a Banach space.

Notation 3. With the Hilbert space V, the dual space V' can be defined as the space of all linear functional $L(v)$, where $L(v)$ is bounded if $L(v) \leq C\,\|v\|_V \;\forall v \in V$.

Lemma 2.3.1 (Poincaré-Frederic's inequality). [Quarteroni (2014)].
Let Ω be a bounded set of $\mathbb{R}^n$ for any n, then a constant C_Ω exists such that

$$\|u\|_{L^2(\Omega)} \leq C_\Omega\,\|u\|_{H^1(\Omega)} \quad \forall u \in H_0^1(\Omega).$$

2.4 Classification of the PDE

Partial differential equations can be divided into three distinct families: *elliptical*, *parabolic* and *hyperbolic* equations, for each of which suitable unique computational methods are considered. For the sake of brevity, here we shall restrict ourselves to the case of a linear second-order PDE, of the form. [Quarteroni (2014)] and [Renardy and Rogers (2004)]

$$A(x,y)\,U_{xx} + B(x,y)\,U_{xy} + C(x,y)\,U_{yy} = F(x,y,U,U_x,U_y) \tag{2.5}$$

with assigned function F.The classification shall be based on the sign of the discriminant, $\Delta = B^2 - 4AC$. In particular :

1. **Elliptic** equation, if $\Delta < 0$.
2. **Parabolic** equation, if $\Delta = 0$.
3. **Hyperbolic** equation, if $\Delta > 0$.

Examples form for the PDE types:

1. Elliptic DEs as the **Poisson equation**

$$\nabla^2 U = h(x,y) \quad \text{or} \quad U_{xx} + U_{yy} = h.$$

If $h = 0$ we introduce **Laplace equation.**

$$\nabla^2 U = 0 \quad \text{or} \quad U_{xx} + U_{yy} = 0$$

2. Parabolic DEs as the **Heat equation** or diffusion equation

$$U_t = \alpha^2 U_{xx}.$$

3. Hyperbolic DEs as the **Wave equation**

$$U_{tt} - \alpha^2 U_{xx} = 0.$$

Remark 4. Some important elliptic PDE in 2D: [Zhang and Yan (2001)]

- $U_{xx} + U_{yy} = 0$ (Laplace Equation).
- $-(U_{xx} + U_{yy}) = h(x,y)$ (Poisson Equation).
- $-(U_{xx} + U_{yy}) + aU = h$ (General Helmholtz Equation).
- $U_{xxxx} + 2U_{xxyy} + U_{yyyy} = 0$ (Bi-harmonic Equation).

Another form to Second-order elliptic PDE is

$$-\nabla \cdot (A\nabla U) + BU = h(x,y).$$

Now, we are going to discus the numerical methods FEM solving PDEs, as their empirical solutions are typically difficult to find. First, we begin with the formulation of variation with boundary conditions.

2.5 Abstract FEM Variational forms

The "weak" formula must be chosen based on the type of partial differential equation. For this purpose, we must first discuss the classification of second-order differential equations. Let us assume $\Omega \subset \mathbb{R}^n$, $n \geq 2$ is an open-connected

Lipschitz boundary set Γ. The general linear elliptic differential equation of the second order is of the form:

$$-\nabla\cdot(\sigma\nabla u)+\beta\cdot\nabla u+\mu u=h, \tag{2.6}$$

where $\sigma:\Omega\to\mathbb{R}^{n\times n}$ is a matrix of real-valued functions, $\beta:\Omega\to\mathbb{R}^n$ is a vector of real-valued functions, and $\mu:\Omega\to\mathbb{R}$ is a real-valued function, [Burden et al. (2015)], [Cao et al. (2019)] and [Quarteroni (2014)]
Since we are considering partial differential equations, correct initial and boundary values must be defined, depending on the form of differential equation. We now have a differential equation for boundary values:

- **Dirichlet boundary condition**
- **Neumann boundary condition**
- **Mixed Dirichlet-Neumann boundary condition**
- **Robin boundary condition**

We derive a weak formula from the equation (2.6), using each one of these boundary conditions. First proceed formally and then define the mathematical structure for the weak formulation, [Braess (2013)].

1. **Dirichlet boundary condition,(homogeneous and Non-homogeneous)**

   (a) **Homogeneous Dirichlet boundary condition** $u=0\ on\ \Gamma$, [Eriksson (1996)].
   Multiply equation (2.6) with the test function v, which disappears with Γ, integrate over Ω and use the Green formula. to obtain

$$\int\limits_{\Omega}\Big(-\nabla\cdot(\sigma\nabla u)+\beta\cdot\nabla u+\mu u\Big)v\,dx=\int\limits_{\Omega}hv\,dx, \tag{2.7}$$

$$\int\limits_{\Omega}(-\nabla\cdot(\sigma\nabla u)v\,dx+\int\limits_{\Omega}(\beta\cdot\nabla u)v\,dx+\int\limits_{\Omega}(\mu u)v\,dx=\int_{\Omega}hv\,dx. \tag{2.8}$$

But

$$\int\limits_{\Omega}-\nabla\cdot(\sigma\nabla u)v\,dx=\int\limits_{\Omega}\sigma\nabla u\cdot\nabla v\,dx.\underbrace{-\int\limits_{\Gamma}v(n\cdot\sigma\nabla u)\,d\Gamma}_{=0}=\int\limits_{\Omega}\sigma\nabla u\cdot\nabla v\,dx \tag{2.9}$$

substituting equation (2.9) in equation (2.8) gives

$$\int\limits_{\Omega}\Big(\sigma\nabla u\cdot\nabla v+(\beta\cdot\nabla u)v+(\mu u)v\Big)\,dx=\int\limits_{\Omega}hv\,dx,\quad\forall u,v\in H_0^1(\Omega). \tag{2.10}$$

Notation 4. Since $u \in H^1(\Omega)$, then u has a boundary condition trace, because of the boundary condition $u|_\Gamma = 0$, the solution is searched for in $H_0^1(\Omega)$ and the test function is also $H_0^1(\Omega)$.

$$\text{The weak formulation: } \begin{cases} \text{seek} & u \in H_0^1(\Omega) \\ \quad\text{such that} & \\ a(u,v) = L(v), & \forall v \in H_0^1(\Omega) \end{cases}$$

where the bilinear form

$$a(u,v) = \int\limits_\Omega \Big(\sigma \nabla u \cdot \nabla v + (\beta \cdot \nabla u)\, v + (\mu u)\, v\Big)\, dx, \tag{2.11}$$

and linear form

$$L(v) = \int\limits_\Omega hv\, dx.$$

(b) **Non-homogeneous Dirichlet boundary condition:** $u = g$ on Γ, where $g : \Gamma \to \mathbb{R}$ [Eriksson (1996)].
We assume that g is sufficiently smooth function so that u_g of g in $H^1(\Omega)$ is achieved.
The function $u_g \in H^1(\Omega)$ is such that $u_g = g$ on Γ.

$$\text{The weak formulation : } \begin{cases} \text{seek} & u \in H^1(\Omega) \\ \quad\text{such that} & \\ u = u_g + \phi, & \phi \in H_0^1(\Omega), \\ \quad\text{and} & \\ a(\phi, v) = \int\limits_\Omega hv\, dx - a(u_g, v), & v \in H_0^1(\Omega). \end{cases}$$

2. **Neumann boundary conditions, homogeneous and non-homogeneous**

(a) **Homogeneous Neumann boundary condition:** $n \cdot \sigma \nabla u = 0$ on Γ.
The weak formulation is similar to the below case, Non-homogeneous, where it is given for arbitrary g.

(b) **Non-homogeneous Neumann boundary condition:** $n \cdot \sigma \nabla u = g$ on Γ, where $g : \Gamma \to \mathbb{R}$.
Then the Neumann condition define the normal derivative of u :
$n \cdot \nabla u =: \partial_n u =: \dfrac{\partial u}{\partial n}$.

For smooth v:

$$\int_{\Omega}(-\nabla\cdot(\sigma\nabla u)+\beta\cdot\nabla u+\mu u)v\,dx=\int_{\Omega}hv\,dx \tag{2.12}$$

$$\int_{\Omega}(\sigma\nabla u.\nabla v+(\beta\cdot\nabla u)v+\mu uv)\,dx=\int_{\Omega}hv\,dx+\int_{\Gamma}\sigma\frac{\partial u}{\partial n}v\,d\Gamma \tag{2.13}$$

$$=\int_{\Omega}hv\,dx+\int_{\Gamma}gv\,dx. \tag{2.14}$$

The bilinear form is therefore

$$a(u,v)=\int_{\Omega}(\sigma\nabla u.\nabla v+(\beta\cdot\nabla u)v+\mu uv)\,dx,$$

and the linear form

$$L(v)=\int_{\Omega}hv\,dx+\int_{\Gamma}gv\,d\Gamma,\quad \forall v\in H^1(\Omega),$$

where $h,g\in L^2(\Omega)$.

$$\text{The weak formulation:}\quad\begin{cases}\text{seek} & u\in H^1(\Omega)\\ \qquad\text{such that} & \\ a(u,v)=\int_{\Omega}hv\,dx+\int_{\Gamma}gv\,dx, & \forall v\in H^1(\Omega).\end{cases}$$

3. **Mixed (Dirichlet- Neumann) boundary condition** [Thomas (1998)]. Consider a boundary partition in the form $\Gamma=\Gamma_D\bigcup\Gamma_N$. Impose Dirichlet on Γ_D and Neumann on Γ_N where $\Gamma_D\bigcap\Gamma_N=\phi$. In this case,

$$u=g_1\ \text{ on }\ \Gamma_D,$$
$$n\cdot\sigma\nabla u=g_2\ \text{ on }\ \Gamma_N.$$

Suppose that $\Gamma_D\neq\phi$ is used to ensure the uniqueness of a solution to the strong problem without conditions of data compatibility. The weak formulation is obtained by multiplying equation (2.6) by a test function v

which disappears on Γ_D, and integrates over Ω

$$\int\limits_\Omega (-\nabla \cdot (\sigma \nabla u) + \beta \cdot \nabla u + \mu u) v \, dx = \int\limits_\Omega hv \, dx \tag{2.15}$$

$$\int\limits_\Omega \Big(\sigma \nabla u . \nabla v + (\beta \cdot \nabla u) v + \mu uv \Big) \, dx = \int\limits_\Omega hv \, dx \tag{2.16}$$

$$+ \underbrace{\int\limits_{\Gamma_D} \sigma \frac{\partial u}{\partial n} v \, d\Gamma_D}_{=0} + \int\limits_{\Gamma_N} \sigma \frac{\partial u}{\partial n} v \, d\Gamma_N \tag{2.17}$$

$$= \int\limits_\Omega hv \, dx + \int\limits_{\Gamma_N} g_2 v \, d\Gamma. \tag{2.18}$$

Having denoted the spaces by V_1 and V_2 as :

$$V_1 = H^1_{\Gamma_D} = \Big\{ v \in H^1(\Omega) : v|_{\Gamma_D} = g_1 \Big\}.$$
$$V_2 = \Big\{ v \in H^1(\Omega) : v\Big|_{\Gamma_D} = 0 \Big\}$$

Hence the bilinear form is

$$a(u, v) = \int\limits_\Omega \Big(\sigma \nabla u \cdot \nabla v + (\beta \cdot \nabla u) v + \mu uv \Big) \, dx,$$

and the linear form is

$$L(v) = \int\limits_\Omega hv \, dx + \int\limits_{\Gamma_N} g_2 v \, d\Gamma.$$

$$\text{The weak formulation: } \begin{cases} \text{seek} & u \in H^1_{\Gamma_D} \\ \quad\text{such that} & \\ a(u, v) = \int\limits_\Omega hv + \int\limits_{\Gamma_N} g_2 \, v & \forall v \in V_2. \end{cases}$$

4. **Robin boundary condition :** $\sigma \dfrac{\partial u}{\partial n} + \gamma u = g$ on Γ, where $g, \gamma : \Gamma \to \mathbb{R}$.
Now, multiply equation (2.6) by a test function v and integrate over Ω

$$\int\limits_\Omega (-\sigma \nabla u + \beta \cdot \nabla u + \mu u) v \, dx = \int\limits_\Omega hv \, dx \tag{2.19}$$

Using Green's formula and the boundary conditions yields

$$\int\limits_{\Omega}\Big(\sigma\nabla u\cdot\nabla v+(\beta\cdot\nabla u)v+\mu uv\Big)\,dx+\int\limits_{\Gamma}\gamma uv\,d\Gamma=\int\limits_{\Omega}hv\,dx+\int\limits_{\Gamma}gv\,d\Gamma \tag{2.20}$$

Hence, the bilinear form

$$\tilde{a}(u,v)=\int\limits_{\Omega}\Big(\sigma\nabla u\cdot\nabla v+(\beta\cdot\nabla u)v+\mu uv\Big)\,dx+\int\limits_{\Gamma}\gamma uv\,d\Gamma,$$

and the linear form is

$$L(v)=\int\limits_{\Omega}hv\,dx+\int\limits_{\Gamma}gv\,d\Gamma.$$

$$\text{The weak formulation :}\begin{cases}\text{seek} & u\in H^1(\Omega)\\ \quad\text{such that} & \\ \tilde{a}(u,v)=L(v) & \forall v\in H^1(\Omega).\end{cases}$$

We can summarize all cases of weak formulation in table (2.1).

Problem	**Trial space**	**Bilinear form**	**Linear form**
Homogeneous Dirichlet	$H_0^1(\Omega)$	$a(u,v)$	$\int\limits_{\Omega} hv$
Non-homogeneous Dirichlet	$H^1(\Omega)$	$a(u,v)$	$\int\limits_{\Omega} hv$
Neumann	$H^1(\Omega)$	$a(u,v)$	$\int\limits_{\Omega} hv+\int\limits_{\Gamma} gv$
Mixed (Dirichlet-Neumann)	$H^1(\Omega)$	$a(u,v)$	$\int\limits_{\Omega} hv+\int\limits_{\Gamma_N} gv$
Robin	$H^1(\Omega)$	$a(u,v)+\int\limits_{\Gamma}\gamma uv$	$\int\limits_{\Omega} hv+\int\limits_{\Gamma} gv$

Table 2.1 – The weak formulation corresponding to the different boundary conditions for the second-order PDE (2.6).

2.6 Existence and Uniqueness theorems

We will address the general theorem for existence and uniqueness in Hilbert space and set the conditions that spaces and bilinear forms should satisfy. The

Lax-Milgram theorem is the basic and most important result to prove the existence and uniqueness of the solution to the elliptical problems. Assuming that H is a real Hilbert space with norm $\|\cdot\|$ and inner product $\langle\cdot\,.\,\cdot\rangle$.

Theorem 2.6.1. ***(Riesz representation Theorem)***
If H is a Hilbert space with scalar product $\langle u, v\rangle$ and norm $\|u\| = \sqrt{\langle u, u\rangle}$, and if $L(v)$ is a linear bounded function on H, then there is a unique $u \in H$, such that $L(v) = \langle u, v\rangle, \quad \forall v \in H,$ [Gustafson (2012)].

Theorem 2.6.2. Suppose a bilinear form a is a symmetric, i.e. $a(v, w) = a(w, v)$ $\forall v, w \in H$, then
$(Minimization\ problem) \Longleftrightarrow (Variational\ Formula)$ with

- (Var) Find $v \in H$ such that $a(v, w) = L(w), \quad \forall w \in H.$
- (Min) Find $v \in H$ such that $F(v) \leq F(w), \ \ \forall w \in H,$
  where
  $F(w) = \dfrac{1}{2}a(w, w) - L(w), \quad \forall w \in H$ [Øksendal (2003)].

Proof. Take $\gamma \in \mathbb{R}$, then

$$\begin{aligned}
(\Leftarrow) \quad F(v + \gamma w) &= \frac{1}{2}a(v + \gamma w, v + \gamma w) - L(v + \gamma w)\\
&= \Big(\frac{1}{2}a(v, v) - L(v)\Big) + \gamma a(v, w) - \gamma L(w) + \frac{1}{2}\gamma^2 a(w, w)\\
&\geq \Big(\frac{1}{2}a(v, v) - L(v)\Big) \qquad \Big(since\ \frac{1}{2}\,\gamma^2\, a(w, w) \geq 0\ \ and\ \ a(v, w) = L(w)\Big)\\
&= F(v)\\
&\therefore F(v) \leq F(v + \gamma w)\\
&\Rightarrow \ \ F(v) \leq F(w), \quad \forall w \in H, \quad \text{which is the minimization problem.}
\end{aligned}$$

$(\Rightarrow)$ $\ \ Let\ g(\gamma) = F(v + \gamma w),$ where $g : \mathbb{R} \to \mathbb{R}.$

Since $\Gamma = 0$ is a minimization value of g, then g′(0)=0, hence

$$0 = g'(0) = 0\ \cdot\ a(w, w) + a(v, w) - L(w) = a(v, w) - L(w)$$

$\therefore a(v, w) = L(w) \quad \forall w \in H,$

which is the variational problem.

□

Proposition 2.6.1 (Parallelogram Law)**.** Let $\|\cdot\|$ be a norm associated to a scalar product $\langle\ \cdot\ ,\ \cdot\ \rangle$. The following equivalence holds

$$\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Theorem 2.6.3. (Young's Inequality)
Suppose p and q are conjugate. For $a, b \in \mathbb{R}$, if $a \geq 0$, $b \geq 0$, then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

We use the Young's Inequality in the corollary

Corollary 2.6.1. (Hölder's Inequality).
Suppose p and q are conjugate. Then for $x = [x_1\, x_2\, \cdots\, x_n]^\top$ and $y = [y_1\, y_2\, \cdots\, y_n]^\top$ in $\mathbb{F}^n$, there holds

$$\sum_{k=1}^{n} |x_k y_k| \leq \left(\sum_{k=1}^{n} |x_k|^p \right)^{\frac{1}{p}} \left(\sum_{k=1}^{n} |y_k|^q \right)^{\frac{1}{q}}$$

or equivalently, using the Hadamard product $x * y = [x_1 y_1 \;\; x_2 y_2 \;\; \cdots \;\; x_n y_n]$,

$$\|x * y\|_1 \leq \|x\|_p \|y\|_q.$$

When $p = 1$ and $q = \infty$, we have

$$\sum_{k=1}^{n} |x_k\, y_k| \leq \left(\sum_{k=1}^{n} |x_k| \right) \sup \{|y_1|,\ |y_2|,\ \cdots,\ |y_n|\},$$

or equivalently,

$$\|x * y\|_1 \leq \|x\|_1\, \|y\|_\infty.$$

Theorem 2.6.4. (**Lax-Milgram**)
Let H be a Hilbert space with norm $\| \cdot \|_H$ and scalar product $\langle \cdot\, , \cdot \rangle_H$, and assume that a is a bilinear functional satisfies:

(1) a is symmetric, i.e. $a(v, w) = a(w, v), \quad \forall v, w \in H$,

(2) a is H-elliptic, i.e. $\exists\, \alpha > 0$ such that $a(v, v) \geq \alpha \|v\|_H^2, \quad \forall v \in H$,

(3) a is continuous, i.e. $\exists\, C \in \mathbb{R}$ such that $|a(v, w)| \leq C \|v\|_H \|w\|_H$,

also let $L \in H \to \mathbb{R}$ be a bounded linear functional on H, i.e. $\exists\, M \in \mathbb{R}$ such that

$$|L(v)| \leq M \|v\|_H, \quad \forall v \in H,$$

then there is a unique function $v \in H$ such that $a(v, w) = L(w), \;\; \forall w \in H$, and the stability estimate $\|v\|_H \leq \dfrac{M}{\alpha}$ holds, [Yu and Zhao (2005)].

Proof. The goal is to construct $v \in H$ to solve the minimization problem $F(v) \leq F(w)$ for all $w \in H$ which is equivalent to the variational problem by the previous theorem. The energy norm $\|v\|^2 = a(v,v),$ is equivalent to the norm of H, because by condition (2) and (3)

$$\alpha \, \|v\|_H^2 \leq a(v,v) = \|v\|^2 \leq |a(v,v)| \leq C \ \|v\|_H^2 \tag{2.21}$$

Let $\beta = \inf_{v \in H} F(v),$ then $\beta \in \mathbf{R}$, since

$$\begin{aligned} F(v) &= \frac{1}{2}\|v\|^2 - L(v) && (2.22)\\ &\geq \frac{1}{2}\|v\|^2 - M\|v\| && (2.23)\\ &\geq \frac{M^2}{2} - M^2 && (2.24)\\ &= -\frac{M^2}{2} && (2.25) \end{aligned}$$

We want to find a solution to the minimization problem $\min_{v \in H} F(v)$. It is therefore natural to study a minimizing sequence v_k, such that

$$F(v_k) \rightarrow \beta = \inf_{v \in H} F(v). \tag{2.26}$$

The next step is to conclude that the v_k infact converges to a limit:

$$\begin{aligned} \left\|\frac{v_k - v_l}{2}\right\|^2 &= \frac{1}{2}\|v_k\|^2 + \frac{1}{2}\|v_l\|^2 - \left\|\frac{v_k + v_l}{2}\right\|^2 \quad \text{(by the parallelogram law)}\\ &= \frac{1}{2}\|v_k\|^2 + \frac{1}{2}\|v_l\|^2 - \left\|\frac{v_k + v_l}{2}\right\|^2 - L(v_k) - L(v_l) + 2L(\frac{v_k + v_l}{2})\\ &= \frac{1}{2}\|v_k\|^2 - L(v_k) + \frac{1}{2}\|v_l\|^2 - L(v_l) - \left(\left\|\frac{v_k + v_l}{2}\right\|^2 - 2L(\frac{v_k + v_l}{2})\right)\\ &= F(v_k) + F(v_l) - 2F\left(\frac{v_k + v_l}{2}\right)\\ &\leq F(v_k) + F(v_l) - 2\beta \quad (\text{ by } \ (2.21))\\ &\rightarrow 0 \quad (\text{by } \ (2.26)) \end{aligned}$$

Thus v_k is a Cauchy sequence in H and since H is a Hilbert space (in particular H is a complete space) we have $v_k \rightarrow v \in H$.

Finally $F(v) = \beta$, since

$$\begin{aligned} |F(v_k) - F(v)| &= \left|\frac{1}{2}(\|v_k\|^2 - \|v\|^2) - L(v_k - v)\right| \\ &= \left|\frac{1}{2}a(v_k - v, v_k + v) - L(v_k - v)\right| \\ &\leq \left(\frac{C}{2}\|v_k + v\|_H + M\right)\|v_k + v\|_H \\ &\to 0 \end{aligned}$$

There is therefore a unique function $u \in H$ such that $F(u) \leq F(v) \quad \forall v \in H$.

★ The uniqueness of v can also be verified from the stability estimate. If v_1 and v_2 are two variation problem solutions we have $a(v_1 - v_2, w) = 0$ for all $w \in V$. Thus, the stability estimate is $\|v_1 - v_2\|_H = 0$ i.e. $v_1 = v_2$ and therefore the solution is unique. □

Remark 5. The Lax-Milgram theorem is a general version of the Riesz theorem, [Gustafson (2012)] and [Øksendal (2003)].

Lemma 2.6.1. (Céa′s lemma)
Let H be a Hilbert space, $a : H \times H \to \mathbb{R}$ a bilinear form and L be a linear form that satisfy the assumptions of the Lax-Milgram theorem. Let V_h be a closed subspace of H, then there is a unique $v_h \in V_h$ such that

$$a(v_h, w_h) = L(w_h), \quad \forall w_h \in V_h,$$

and

$$\|v - v_h\|_H \leq \frac{M}{\alpha} \inf_{w_h \in V_h} \|v - w_h\|_H,$$

where M is a continuity constant of a and α its H-ellipticity constant, [Brenner and Scott (2008)] and [Le Dret and Lucquin (2016)].

Proof. Since V_h is a closed subspace of H, the Lax-Mligram hypotheses for the variation problem on V_h are therefore satisfied,thus the existence and uniqueness of v_h is assured. Now we have

$$a(v, w) = L(w), \quad \forall w \in H, \tag{2.27}$$

in particular for $w = w_h^* \in V_h$ so

$$a(v_h, w_h^*) = L(w_h^*), \quad \forall w_h^* \in V_h. \tag{2.28}$$

Substract (2.28) from (2.27) to get

$$a(v - v_h, w_h^*) = 0, \quad \forall w_h^* \in V_h. \quad \text{(H-ellipticity)}$$

Next,

$$\begin{aligned}
\alpha\|v - v_h\|_H^2 &\leq a(v - v_h, v - v_h) \\
&= a(v - v_h, v - w_h + w_h - v_h) \\
&\leq a(v - v_h, v - w_h) + a(v - v_h, w_h - v_h) \\
&= a(v - v_h, v - w_h) + 0 \quad \text{(by Galerkin orthogonality)} \\
&= a(v - v_h, v - w_h) \\
&\leq M\|v - v_h\|_H\|v - w_h\|_H \\
\Longrightarrow \|v - v_h\|_H &\leq \frac{M}{\alpha}\|v - w_h\|_H, \quad \forall w_h \in V_h. \\
Thus, \\
\|v - v_h\|_H &\leq \frac{M}{\alpha} \inf_{w_h \in V_h} \|v - w_h\|_H \\
&= \frac{M}{\alpha} \min_{w_h \in V_h} \|v - w_h\|_H \quad (\text{ } V_h \text{ is closed }).
\end{aligned}$$

□

Remark 6.

(i) Céa′s theorem shows that v_h is quasi-optimal in the sense that the error $\|v - v_h\|_H$ is proportional to the best that the subspace V_h can be used.

(ii) In the symmetrical case, we have shown that

$$\|v - v_h\|_H = \min_{w_h \in V_h} \|v - w_h\|_H.$$

This means that there is no "better" approximation in the finite element space V_h than the finite element solution itself, [Brenner and Scott (2008)].

Chapter 3

Approximation of elliptical problems

This chapter explores the construction of the FEM for elliptic problems and describes some of their key properties. Unlike finite difference schemes, which are constructed more or less by replacing derivatives in differential equations with divided differences, the derivation of the FEM is very systematic, [Kunert (2001)] and [Li (2017)]. The finite element method has been developed to solve complex engineering problems, particularly in the field of modeling flexibility and structural mechanical engineering, including elliptical and complex engineering. In this chapter, we will solve the problems in $1D$ and $2D$ that these models represent.

3.1 FEM for 1D boundary value problems

Consider the one-dimensional problem

$$\begin{cases} -u''(x) = h(x), & x \in \Omega = (0,1), \\ & u(0) = 0 = u(1). \end{cases} \tag{3.1}$$

we will use the following two methods to approximate the solution.

- The Galerkin method (variational method).
- The Ritzs method (Minimization method).

Definition 3.1.1. (The Galerkin Method)
Is the method which is used to rewrite the differential equation in a variational form and then discertize the system. So, when we talk about Galerkin method,

which uses piecewise polynomial as approximation functions, we mean the finite element method, [Rauch (1997)].

Definition 3.1.2. (The Ritzs method)
Is a straightforward method for finding an approximate solution to boundary value problems. In the Ritz method, we solve a boundary value problem by approximating a solution with a linear approximation of basis functions. The method is based on a component of mathematics called variation calculus.

3.1.1 The Galerkin method or variational form method

Now we will use the Galerkin method to solve equation (3.1) by the following steps, [Larsson and Thomee (2003)] and [Strauss (2007)].

1. *Construct a variational or weak formulation.*
2. *Generate a mesh, e.g., a Cartesian mesh uniform.*
3. *Build a set of basic functions.*
4. *Reflect the approximate finite element solution by a linear combination of basic functions.*
5. *Solve the linear system of equations for the coefficients and thus obtain the approximate solution.*
6. *Conduct the error analysis (A periori and A posteriori error analysis).*

Recall the variation formulation of equation (3.1) by seeking $u \in H_0^1(\Omega)$ such that

$$\underbrace{a(u,v) = L(v),}_{\text{weak form}} \qquad \forall v \in H_0^1(\Omega), \tag{3.2}$$

$$\text{where,} \qquad a(u,v) = \int\limits_{\Omega} u'\, v'\, dx \qquad \text{and} \qquad L(v) = \int\limits_{\Omega} h\, v\, dx.$$

Next, we join $\mathcal{T}_h = \Big\{0 = x_0 \ < x_1 \ < \cdots \ < x_M \ < x_{M+1} = 1\Big\}$. The partition of $\Omega = [0,1]$ with subinterval $I_i = [x_{i-1}, x_i]$, see Figure (3.1), and $h_i = x_i - x_{i-1}$ for $i = 1, \cdots, M+1$. Define the partwise constant function $g(x) = x_i - x_{i-1} = h_i$

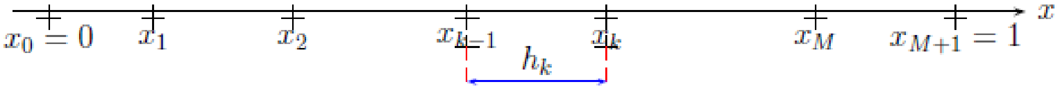


Figure 3.1 – subdivision of Ω = [0,1]

Remark 7. In 1D problem

(1) x_i is called node, or nodal point.

(2) (x_{i-1}, x_i) is called an element and we denote it by Ω_i, for example $(x_1, x_2) \equiv \Omega_2$.

(3) $h = \max\limits_{1 \le i \le M+1} h_i$.

Now, a discrete solution will be found in the finite dimensional function space, so let $\mathcal{C}(\Omega) = \mathcal{C}(0,1)$ denote the set of all continuous piecewise linear function on $\mathcal{T}_h$ (continuous in the whole interval Ω, linear on each sub-interval I_i) with zero boundary conditions and define

$$V_h^{(0)} = \Big\{ v : v \in \mathcal{C}(0,1)\ ,\ v(0) = v(1) = 0 \Big\}.$$

Remember that $V_h^{(0)}$ is a finite dimensional subspace of H_0^1, (dim $V_h^{(0)} = M+1$), where

$$H_0^1 = \Big\{ w(x) : \int\limits_0^1 (w^2(x) + (w'(x))^2)\, dx < \infty \ \ and \ \ w(0) = w(1) = 0 \Big\}.$$

The finite element formulation for our Dirichlet boundary value problem is given by: find $u_h \in V_h^{(0)}$ in such a way that the following variation formulation holds true

$$\int\limits_\Omega u_h'\, v'\, dx = \int\limits_\Omega h\, v\, dx, \quad \forall v \in V_h \tag{3.3}$$

$$\text{or} \quad a(u_h, v) = (h, v), \quad \forall v \in V_h. \tag{3.4}$$

Notice that the method of the finite element is a finite dimensional variant of the weak formulation.Let us introduce the basis functions $\Big\{\varphi_i\Big\}_{i=1}^{M} \subset V_h^{(0)}$ of hat function which are linearly independent and has the property

$$\varphi_i(x_j) = \begin{cases} 1, & \text{if} \quad i = j \\ 0, & \text{if} \quad i \ne j \end{cases}$$

After we introduce the basis function we rewrite $V_h^{(0)}$ as

$$V_h^{(0)} = \text{Span}\Big\{\ \varphi_0, \varphi_1, \varphi_2, \cdots, \varphi_{M+1}\ \Big\}.$$

Any test function $v \in V_h^0$ can be one of φ_i, $i = 0, 1, \cdots, M+1$. Also, in terms of the basis functions, $u_h = \sum\limits_{j=0}^{M+1} \xi_j\, \varphi_j(x)$, with $\xi_j = u_h(x_j)$.

Substitute v and u_h in (3.4) and taking into account that $\xi_0 = u_h(x_0) = 0$ and $\xi_{M+1} = u_h(x_{M+1}) = 0$ this gives

$$\begin{aligned} &\sum_{j=1}^{M} \xi_j \int_{\Omega} \varphi_j' \, \varphi_i' \, dx = \int_{\Omega} h\varphi_i \, dx \\ \text{or} \quad &\sum_{j=1}^{M} \xi_j \, a(\varphi_j, \varphi_i) = (h, \varphi_i), \quad \text{for} \quad i = 1, \cdots, M. \end{aligned}$$

This linear system of equations can be represented as

$$A\,\xi = b, \tag{3.5}$$

where $\xi = (\xi_j)$, $A = (a_{ij})$, is the `stiffness matrix` with elements $a_{ij} = a(\varphi_j, \varphi_i)$, and $b = (b_i)$, the `load vector` with $b_i = (h, \varphi_i)$, $i, j = 1, 2, \cdots, M$.

Now, we would like to determine $\xi_j = u_h(x_j)$, the estimated values of $u(x)$ at the nodal points x_j, $1 \le j \le M$. To proceed in the calculation, we have

$$\sum_{j=1}^{M} \xi_j \left(\int_0^1 \varphi_i' \, \varphi_j' \, dx \right) = \int_0^1 h \, \varphi_i \, dx, \quad i = 1, \cdots .M,$$

where

$$A = \left\{ a_{ij} \right\}_{i,j=1}^{M}, \quad a_{ij} = \int_0^1 \varphi_i' \, \varphi_j' \, dx,$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{bmatrix}, \quad \text{with} \quad b_i = \int_0^1 h \, \varphi_i \, dx, \quad \text{and} \quad \xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_M \end{bmatrix}$$

To compute the entries of the stiffness matrix A, we recall the set of basis functions φ_i

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i}, & x \in [x_{i-1}, x_i] \\ \dfrac{x_{i+1} - x}{h_{i+1}}, & x \in [x_i, x_{i+1}] \\ 0, & O.W \end{cases} \implies \varphi_i'(x) = \begin{cases} \dfrac{1}{h_i}, & x \in (x_{i-1}, x_i) \\ \dfrac{-1}{h_{i+1}}, & x \in (x_i, x_{i+1}) \\ 0, & O.W \end{cases}$$

The Stiffness Matrix A :

(1) If $|i - j| > 1$, then φ_i and φ_j have disjoint supports, see Figure (3.2), and

obviously $a_{ij} = \int\limits_0^1 \varphi_i' \, \varphi_j' \, dx = 0.$

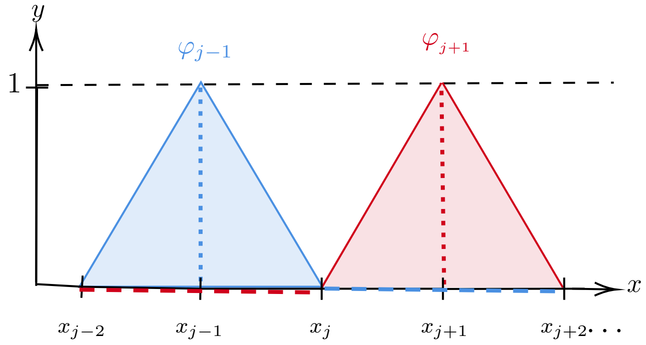


Figure 3.2 – φ_{j-1}, φ_{j+1}

(2) For $i = j$,

$$a_{ii} = \int\limits_{x_{i-1}}^{x_i} \left(\frac{1}{h_i}\right)^2 dx + \int\limits_{x_i}^{x_{i+1}} \left(\frac{-1}{h_{i+1}}\right)^2 dx = \frac{\overbrace{x_i - x_{i-1}}^{h_i}}{h_i^2} + \frac{\overbrace{x_{i+1} - x_i}^{h_{i+1}}}{h_{i+1}^2} = \frac{1}{h_i} + \frac{1}{h_{i+1}}.$$

(3) For $j = i \pm 1$ see figure (3.3),

$$a_{i,i+1} = \int\limits_{x_i}^{x_{i+1}} \frac{-1}{h_{i+1}} \cdot \frac{1}{h_{i+1}} \, dx = -\frac{\overbrace{x_{i+1} - x_i}^{h_{i+1}}}{h_{i+1}^2} = -\frac{1}{h_{i+1}}.$$

It is clear that $a_{i+1,i} = a_{i,i+1} = -\dfrac{1}{h_{i+1}}.$

To sum up, we have,

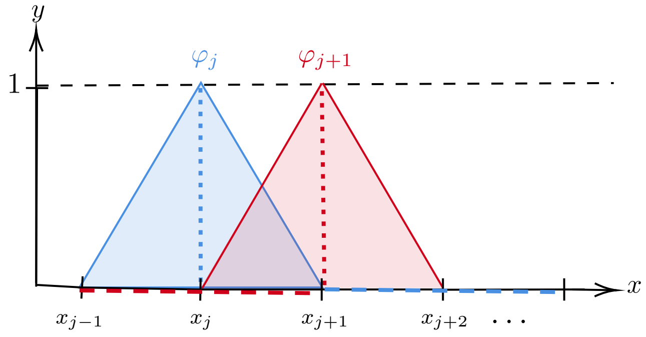


Figure 3.3 – φ_j,φ_{j+1}

$$\begin{cases} a_{ij} = 0, & \text{if } |i-j| > 1, \\ a_{ii} = \dfrac{1}{h_i} + \dfrac{1}{h_{i+1}}, & \text{for } i = 1, 2, \cdots, M, \\ a_{i+1,i} = a_{i,i+1} = -\dfrac{1}{h_i}, & \text{for } i = 1, 2, \cdots, M. \end{cases}$$

The stiffness matrix A is symmetric and has the form:

$$A = \begin{bmatrix} \dfrac{1}{h_1} + \dfrac{1}{h_2} & \dfrac{-1}{h_2} & 0 & \cdots & 0 \\ \dfrac{-1}{h_2} & \dfrac{1}{h_2} + \dfrac{1}{h_3} & \dfrac{-1}{h_3} & 0 & 0 \\ 0 & \ddots & \ddots & & \\ & & \ddots & \ddots & \dfrac{-1}{h_M} \\ 0 & & 0 & \dfrac{-1}{h_M} & \dfrac{1}{h_M} + \dfrac{1}{h_{M+1}} \end{bmatrix}.$$

With a uniform mesh, we put $h_i = h$ for all $i = 1, 2, \cdots, M,$ so the stiffness matrix will be

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

For the load vector b, we have

$$b_i = \int_0^1 h(x)\, \varphi_i\, dx = \int_{x_{i-1}}^{x_i} h(x) \frac{x - x_{i-1}}{h_i}\, dx + \int_{x_i}^{x_{i+1}} h(x) \frac{x_{i+1} - x}{h_{i+1}}\, dx.$$

Computing local Stiffness Matrix L_i^e ***and local Load Vector*** H_i^e **:**
For the element (x_i, x_{i+1}) there are only two nonzero hat functions

$$\psi_i^e(x) = \frac{x_{i+1} - x}{x_{i+1} - x_i}, \qquad \psi_{i+1}^e(x) = \frac{x - x_i}{x_{i+1} - x_i}.$$
$$(\psi_i^e)' = \frac{-1}{h_{i+1}}, \qquad (\psi_{i+1}^e)' = \frac{1}{h_{i+1}}.$$

When ψ_i^e and ψ_{i+1}^e are described only by the specific element, see Figure (3.4), a stiffness matrix and a load vector from the two hat functions can be easily explained

$$\int_{x_i}^{x_{i+1}} (\psi_i')^2 \, dx = \int_{x_i}^{x_{i+1}} (\frac{-1}{h_{i+1}})^2 \, dx = \int_{x_i}^{x_{i+1}} \frac{1}{h_{i+1}^2} \, dx = \frac{1}{h_{i+1}}.$$

$$\int_{x_i}^{x_{i+1}} \psi_i'\psi_{i+1}' \, dx = \int_{x_i}^{x_{i+1}} \frac{-1}{h_{i+1}}\frac{1}{h_{i+1}} \, dx = \int_{x_i}^{x_{i+1}} \frac{-1}{h_{i+1}^2} \, dx = \frac{-1}{h_{i+1}}.$$

$$\int_{x_i}^{x_{i+1}} (\psi_{i+1}')^2 \, dx = \int_{x_i}^{x_{i+1}} (\frac{1}{h_{i+1}})^2 \, dx = \int_{x_i}^{x_{i+1}} \frac{1}{h_{i+1}^2} \, dx = \frac{1}{h_{i+1}}.$$

Thus, the local stiffness matrix is

$$L_i^e = \begin{bmatrix} \frac{1}{h_{i+1}} & \frac{-1}{h_{i+1}} \\ \frac{-1}{h_{i+1}} & \frac{1}{h_{i+1}} \end{bmatrix}.$$

and the local load vector is

$$H_i^e = \begin{bmatrix} \int_{x_i}^{x_{i+1}} h(x) \, \psi_i \, dx \\ \int_{x_i}^{x_{i+1}} h(x) \, \psi_{i+1} \, dx \end{bmatrix}.$$

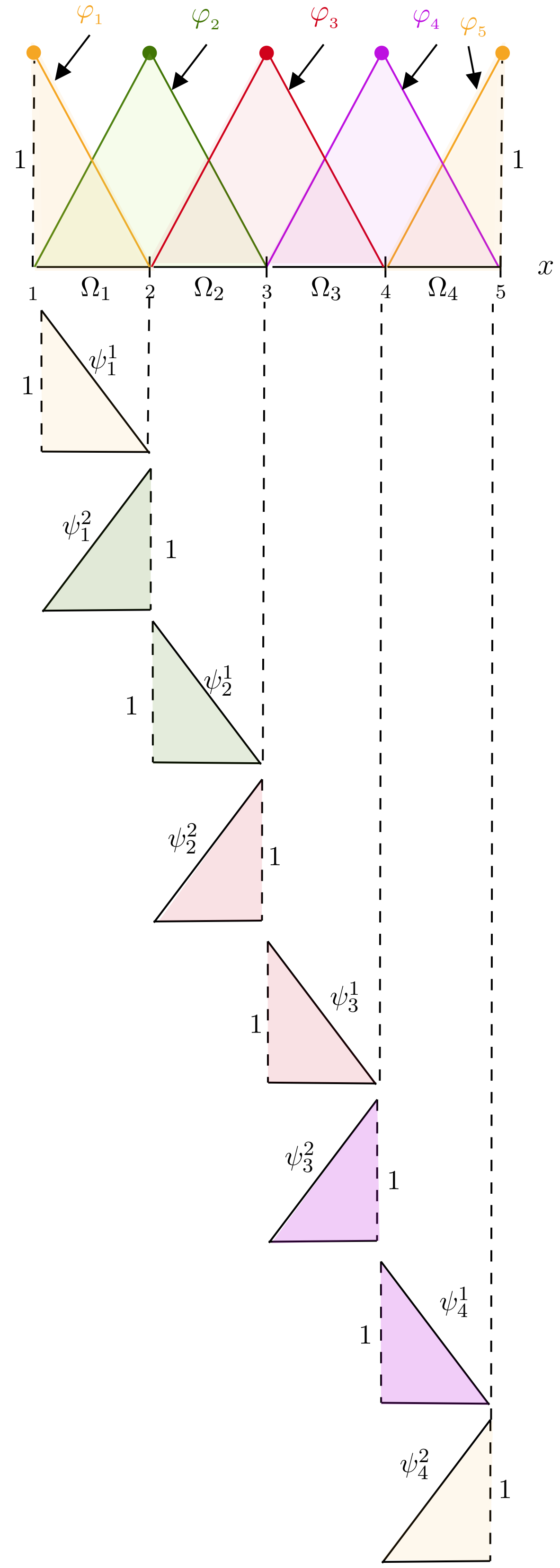


Figure 3.4 – Continuous piecewise linear basis function φ_i for a four-element mesh generated by linear shape function ψ_i^e,ψ_{i+1}^e defined over each element.

3.1.2 The Ritz Method (Minimization method)

Even though not every problem has a minimization type, the Ritz method is an earlist method that has been shown to be efficient for the model problem (3.1), [Rauch (1997)]. The minimization type is

$$\min_{v\in H_0^1(0,1)} F(v) \quad \text{where} \quad F(v) = \frac{1}{2}\int_0^1 (v')^2\, dx - \int_0^1 hv\, dx \tag{3.6}$$

Substitute the approximate solution $u_h(x) = \sum_{j=1}^{M} \xi_j \varphi_j(x)$ in $F(v)$ we get

$$F(u_h) = \frac{1}{2}\int_0^1 \left(\sum_{j=1}^{M} \xi_j\, \varphi_j'(x)\right)^2 dx - \int_0^1 h(x)\left(\sum_{j=1}^{M} \xi_j\, \varphi_j(x)\right) dx.$$

This is a multivariate function of $\{\xi_1, \xi_2, \cdots, \xi_M\}$ written as $F(u_h)$. The requisite conditions for a global minimum are

$$\frac{\partial F}{\partial \xi_1} = 0, \quad \frac{\partial F}{\partial \xi_2} = 0, \cdots, \frac{\partial F}{\partial \xi_i} = 0, \cdots, \frac{\partial F}{\partial \xi_M} = 0.$$

Thus, with regard to the partial derivative,

$$\begin{aligned}
\frac{\partial F}{\partial \xi_1} &= \int_0^1 \left(\sum_{j=1}^{M} \xi_j\, \varphi_j'\right) \varphi_1'\, dx - \int_0^1 h(x)\, \varphi_1\, dx = 0 \\
&\vdots \\
\frac{\partial F}{\partial \xi_i} &= \int_0^1 \left(\sum_{j=1}^{M} \xi_j\, \varphi_j'\right) \varphi_i'\, dx - \int_0^1 h\, \varphi_i\, dx = 0, \quad i = 1, 2, \cdots, M.
\end{aligned}$$

By rewriting the last equation,we arrive at

$$\sum_{j=1}^{M} \left(\int_0^1 \varphi_j'\, \varphi_i'\, dx\right) \xi_j = \int_0^1 h\, \varphi_i\, dx \quad i = 1, 2, \cdots, M.$$

These are the same equations that we get from Galerkin method.

Comparison of the methods of Galerkin and Ritz in the FEM :

(i) Ritz and Galerkin are theoretically similar to several issues.

(ii) The Ritz method is based on the techniques of minimization and optimization of the model that can be used to solve the problem.

(iii) The Galerkin method typically has a lower criterion than the Ritz method.

3.2 The process of the finite element for 2D

The procedure of the finite element method to solve 2D problems is the same as that of the following flow chart, which also has 1D problems, demonstration

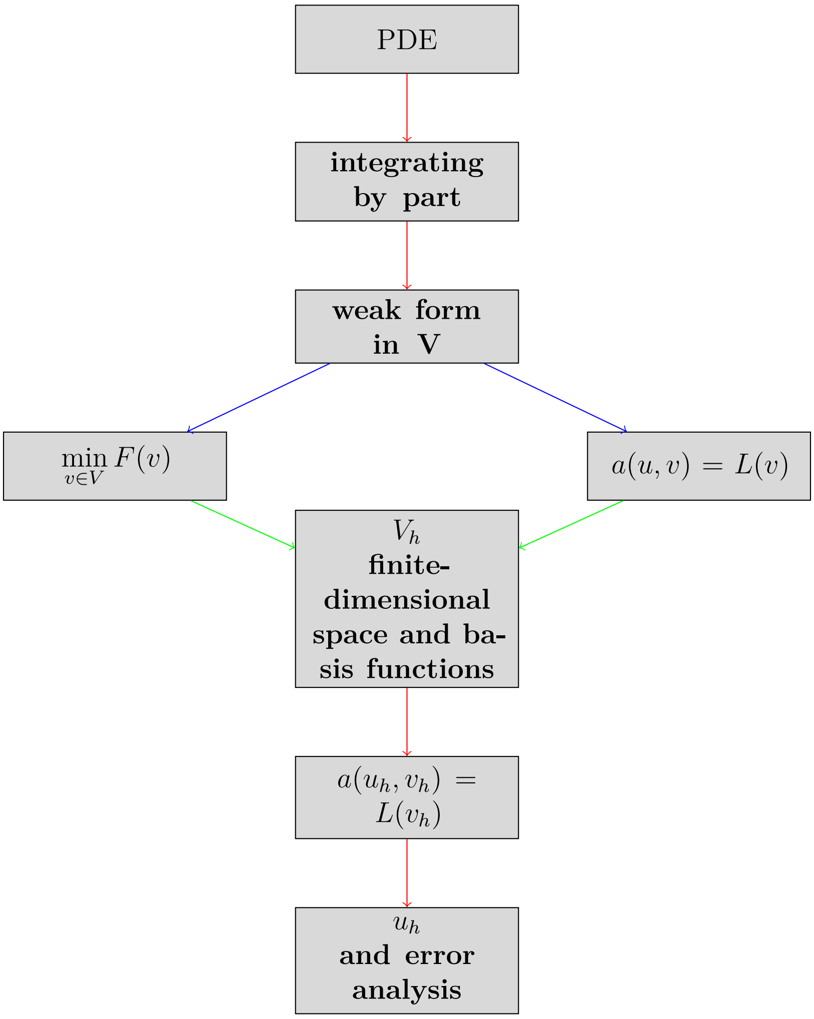


Figure 3.5 – Procedure of the finite element process for solving 2D problems

3.2.1 Green's theorem in 2D

In this section it is important to know the divergence theorem in the Cartesian coordinates.

Theorem 3.2.1. If $F \in H^1(\Omega) \times H^1(\Omega)$ is a $2D$ vector then

$$\iint\limits_{\Omega} \nabla \cdot F \, dx \, dy = \int\limits_{\Gamma} F \cdot n \, ds, \tag{3.7}$$

where n is the normal direction unit pointing out to the boundary Γ with the line element ds and ∇ is the gradient operator see Figure (3.6), i.e. $\nabla = \left[\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right]^\top$.

Secondly Green's theorem is a corollary of the divergence theorem if we set $F = v\nabla u = \left[v\dfrac{\partial u}{\partial x}, v\dfrac{\partial u}{\partial y}\right]^\top$.

Remark 8.

1. For $F = v\nabla u$ we have

$$\begin{aligned}
\nabla \cdot F &= \frac{\partial}{\partial x}\left(v\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(v\frac{\partial u}{\partial y}\right) \\
&= \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + v\,\frac{\partial^2 u}{\partial x^2} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} + v\,\frac{\partial^2 u}{\partial y^2} \\
&= \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} + v\,\frac{\partial^2 u}{\partial x^2} + v\,\frac{\partial^2 u}{\partial y^2} \\
&= \nabla u \cdot \nabla v + v\Delta u.
\end{aligned}$$

2. $\Delta u = \nabla \cdot \nabla u = u_{xx} + u_{yy}$.

3. The normal derivative $\dfrac{\partial u}{\partial n}$ is defined by

$$\begin{aligned}
\frac{\partial u}{\partial n} &= n \cdot \nabla u \\
&= n_x\frac{\partial u}{\partial x} + n_y\frac{\partial u}{\partial y}
\end{aligned}$$

where

$$n = (n_x, n_y) \quad \text{and} \quad (n_x^2 + n_y^2 = 1)$$

4. The normal derivative $\dfrac{\partial u}{\partial n}$ can be abbreviated by u_n.

Lemma 3.2.1. (Green′s formula)
Let Ω be a bounded domain and let $u(x,y) \in H^2(\Omega)$ and $v(x,y) \in H^1(\Omega)$ then

$$\iint\limits_{\Omega} \Delta uv \, dx \, dy = \int\limits_{\Gamma} u_n v \, ds - \iint\limits_{\Omega} \nabla u \cdot \nabla v \, dx \, dy. \tag{3.8}$$

Proof. By using equation (3.7)

$$\begin{aligned}\iint\limits_{\Omega} \nabla \cdot F \, dx \, dy &= \iint\limits_{\Omega} \Big(\nabla u \cdot \nabla v + \Delta uv\Big) \, dx \, dy \\ &= \iint\limits_{\Omega} \nabla u \cdot \nabla v \, dx \, dy + \iint\limits_{\Omega} \Delta uv \, dx \, dy\end{aligned}$$

and

$$\begin{aligned}\int\limits_{\Gamma} F \cdot n \, ds &= \int\limits_{\Gamma} (\nabla u \cdot n) v \, ds \\ &= \int\limits_{\Gamma} \frac{\partial u}{\partial n} v \, ds = \int\limits_{\Gamma} u_n v \, ds.\end{aligned}$$

Therefore
$$\begin{aligned}\iint\limits_{\Omega} \Big(\nabla u \cdot \nabla v + \Delta uv\Big) \, dx \, dy &= \int\limits_{\Gamma} u_n v \, ds \\ \iint\limits_{\Omega} \nabla u \cdot \nabla v \, dx \, dy + \iint\limits_{\Omega} \Delta uv \, dx \, dy &= \int\limits_{\Gamma} u_n v \, ds \\ \iint\limits_{\Omega} \Delta uv \, dx \, dy &= \int\limits_{\Gamma} u_n v \, ds - \iint\limits_{\Omega} \nabla u \cdot \nabla v \, dx \, dy.\end{aligned}$$

□

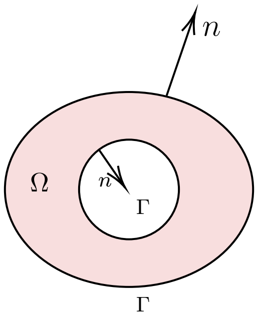


Figure 3.6 – 2D Domain diagram Ω, its boundary Γ, and its normal path unit n.

3.2.2 Weak form with variation method in 2D

Consider the Tow-dimensional problem

$$\begin{cases} -\Delta u = h(x,y), & (x,y) \in \Omega, \\ & u(x,y)\Big|_{\Gamma} = 0. \end{cases} \tag{3.9}$$

Let Ω be a domain bounded in $\mathbb{R}^2$ with a polygon boundary Γ, the region Ω can be precisely filled by a finite number of triangles. It will be assumed that any pair of triangles in a triangulation of intersect along a complete edge, at a vertex, or not at all, we will denote the diameter (longest side) of h_k, the k triangle, by h, i.e., $h = \max_k h_k$, [Gustafson (2012)].

In order to construct an approximation of the finite element of the problem, we begin by considering its weak formulation. Multiply both sides of equation (3.9) with a test function $v \in H_0^1(\Omega)$, integrate over Ω, and apply Green formula to get

$$\begin{aligned} \int\limits_{\Omega} hv\,dx &= \int\limits_{\Omega} -\Delta u\, v\, dx, \\ &= \int\limits_{\Omega} \nabla u \cdot \nabla v\, dx - \underbrace{\int_{\Gamma} (n \cdot \nabla u) v\, ds}_{=0} \\ &= \int\limits_{\Omega} \nabla u \cdot \nabla v\, dx. \end{aligned}$$

The weak formulation is

$$\int\limits_{\Omega} \nabla u \cdot \nabla v\, dx = \int\limits_{\Omega} h\, v\, dx, \tag{3.10}$$

$$\text{or} \quad a(\nabla u, \nabla v) = (h, v) \equiv L(v), \qquad v \in H_0^1(\Omega). \tag{3.11}$$

Let $V_h \subset H_0^1$ be the subspace of the finite elements consisting of continuous linear functions on the partition satisfying the boundary condition $v = 0$ on Γ. Then Galerkin method for equation (3.9) is formulated as,

$$\begin{aligned} &\text{Find } u_h \in V_h \text{ such that} \\ &\qquad \int\limits_{\Omega} \nabla u_h \cdot \nabla v\, dx = \int\limits_{\Omega} h\, v\, dx. \\ &\qquad \textit{or} \quad a(\nabla u_h, \nabla v) = L(v), \quad \forall v \in V_h. \end{aligned}$$

Making the Ansatz. Let M be the number of interior nodes. Using the basis functions $\Big\{\varphi_j(x)\Big\}_{j=1}^{M} \subset V_h$, each function $u_h \in V_h$ can be written

$$u_h = \sum_{j=1}^{M} \xi_j \, \varphi_j(x), \tag{3.12}$$

where ξ_j is the value of u_h at node j, and $\varphi_j(x)$ is the basis function, then the FEM can be reset to the following:
substituting equation (3.23) in equation (3.10) yield that

$$\sum_{j=1}^{M} \xi_j \int\limits_{\Omega} \nabla\varphi_j \cdot \nabla v \, dx = \int\limits_{\Omega} h \, v \, dx \quad \forall v \in V_h.$$

Since the equation hold for all $v \in V_h$, in particular it is hold for $v = \varphi_i$, $i = 1, 2, \cdots, M$.

$$\sum_{j=1}^{M} \xi_j \int\limits_{\Omega} \nabla\varphi_j \cdot \nabla\varphi_i \, dx = \int\limits_{\Omega} h \, \varphi_i \, dx, \quad i = 1, \cdots, M. \tag{3.13}$$

Problem (3.13) is i system of linear equation in the coefficients ξ_j, $j = 1, 2, \cdots, M$, that is,

$$S \, \xi = F,$$

where

$$(\texttt{ The stiffness matrix}) \quad S = \Big(s_{ij}\Big) \in \mathbb{R}^{n\times n}, \quad s_{ij} = \int\limits_{\Omega} \nabla\varphi_j \cdot \nabla\varphi_i \, dx, \quad i, j = 1, 2, \cdots, M.$$

$$(\texttt{The load vector }) \quad F = \Big(F_1, \cdots, F_M\Big)^{\top} \in \mathbb{R}^{n}, \quad F_i = \int_{\Omega} h \, \varphi_i \, dx, \quad i = 1, 2, \cdots, M.$$

$$\text{and} \quad \xi = \Big[\xi_1, \cdots, \xi_M\Big]^{\top} \in \mathbb{R}^{n}.$$

3.2.3 Weak form with minimization method in 2D

Define the quadratic function $J : H_0^1 \to \mathbb{R}$ as:

$$J(w) = \frac{1}{2} \, a(w, w) - L(w), \quad w \in H_0^1.$$

Lemma 3.2.2. Suppose u is the weak solution for form (3.2) in H_0^1, and suppose that $a(\,\cdot\,,\,\cdot\,)$ is a symmetric bilinear functional in H_0^1, then u is the unique minimizer of $J(\,\cdot\,)$ over H_0^1.

Lemma 3.2.3. Suppose that $u \in H_0^1$ minimizes $J(\ \cdot\ )$ over H_0^1, then u is the unique solution of problem (3.2).

The two lemmas together convey the equivalence of the weak formulation:

$$\text{find } u \in H_0^1 \quad \text{such as} \quad a(u,w) = L(w), \qquad \forall w \in H_0^1, \tag{3.14}$$

of the elliptic boundary value problem (3.9) to the related minimization problem:

$$\text{find } u \in H_0^1 \quad \text{such as} \quad J(u) \leq J(w), \qquad \forall w \in H_0^1.$$

We can now use this equivalence to provide a variable description to approximate the finite element solution u_h. Given that V_h is a finite-dimensional subspace of H_0^1 consisting of continuous polynomials of a fixed degree, the finite element approximation of (3.14) is

$$\text{find } u_h \in V_h \quad \text{such as} \quad a(u_h,w_h) = L(w_h), \qquad \forall w_h \in V_h. \tag{3.15}$$

You can repeat the argument above (or simply replace H_0^1 with V_h all the time) to show the equivalence of equation (3.15) to the following minimization problem:

$$\text{find } u_h \in V_h \quad \text{such as} \quad J(u_h) \leq J(w_h), \qquad \forall w_h \in V_h.$$

Thus, u_h can be defined as a unique functional minimizer of $J(w)$ in V_h. As w_h extends over the finite element space V_h, this means that the finite element solution u_h inherits the energy minimization properties of the weak solution $u \in H_0^1$ in the sense that: [Houston and Süli (2001)]

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

Remark 9. In general, of course $J(u) < J(u_h)$.

3.3 Numerical Examples

Example 1. Consider the B.V.P:

$$\begin{cases} \Big((x+2)u'\Big)' + u = 3x, & x \in (0,1), \\ & u^{'}(0) = u(1) = 0. \end{cases} \tag{3.16}$$

Let $I = (0,1)$ be divided into a uniform mesh, calculate the finite element approximation u_h for $n = 2$.

Solution :

Variation Formulation Multiply equation (3.16) by a test function $v(x)$ such that $v(1) = 0$ and integrate over $\Omega = (0,1)$

$$\int_0^1 \Big((x+2)u'\Big)' v\, dx + \int_0^1 uv\, dx = \int_0^1 3xv\, dx \tag{3.17}$$

$$(x+2)u'v\Big|_0^1 + \int_0^1 (x+2)u'v'\, dx + \int_0^1 uv\, dx = \int_0^1 3xv\, dx \tag{3.18}$$

$$\int_0^1 (x+2)u'v'\, dx + \int_0^1 uv\, dx = \int_0^1 3xv\, dx. \tag{3.19}$$

find $u \in H^1$ such that $u(1) = 0$ and (3.19) holds for all $v \in H^1$ and $v(1) = 0$. Let V_h be a finite dimensional subspace of H^1 consists of continuous linear polynomials spanned by $\varphi_i, \;\; i = 0,1,2$ on the partition $x_0 = 0, x_1 = \frac{1}{2}, x_2 = 1$.

Discertization: Let u_h be an approximation of u from the space V_h, then, using $\xi_i = u(x_j), \forall j = 0,1,2$

$$u_h = \sum_{j=0}^{2} \xi_j \varphi_j = \sum_{j=0}^{1} \xi_j \varphi_j + \xi_2 \varphi_2 = \sum_{j=0}^{1} \xi_j \varphi_j,$$

since $\xi_2 = u(x_2) = u(1) = 0$. Let $v \in V_h$, then $v = \Big\{\varphi_i\Big\}_{i=0}^{1}$, thus (3.19) can be written as :

$$\int_0^1 (x+2) \sum_{j=0}^{1} \xi_j \varphi_j' \varphi_i'\, dx + \int_0^1 \sum_{j=0}^{1} \xi_j \varphi_j \varphi_i\, dx = \int_0^1 3x\varphi_i\, dx, \;\; i = 0,1 \tag{3.20}$$

$$\sum_{j=0}^{1} \xi_j \int_0^1 (x+2)\varphi_j' \varphi_i'\, dx + \sum_{j=0}^{1} \xi_j \int_0^1 \varphi_j \varphi_i\, dx = \int_0^1 3x\varphi_i\, dx, \;\; i = 0,1 \tag{3.21}$$

When $i = 0$,

$$\Big(\int_0^1 (x{+}2)\varphi_0'\varphi_0' dx\Big)\xi_0 + \Big(\int_0^1 (x{+}2)\varphi_1'\varphi_0' dx\Big)\xi_1 + \Big(\int_0^1 \varphi_0\varphi_0 dx\Big)\xi_0 + \Big(\int_0^1 \varphi_1\varphi_0 dx\Big)\xi_1 = \int_0^1 3x\varphi_0 dx.$$

When $i = 1$,

$$\Big(\int_0^1 (x{+}2)\varphi_0'\varphi_1' dx\Big)\xi_0 + \Big(\int_0^1 (x{+}2)\varphi_1'\varphi_1' dx\Big)\xi_1 + \Big(\int_0^1 \varphi_0\varphi_1 dx\Big)\xi_0 + \Big(\int_0^1 \varphi_1\varphi_1 dx\Big)\xi_1 = \int_0^1 3x\varphi_1 dx.$$

Note that

$$\varphi_0 = \begin{cases} 1-2x, & \text{for } 0<x<\frac{1}{2}, \end{cases} \text{ and } \varphi_1 = \begin{cases} 2x, & \text{for } 0<x<\frac{1}{2}, \\ 2-2x, & \text{for } \frac{1}{2}<x<1. \end{cases}$$

After performing the integrals above, we arrive at

$$\begin{aligned} -\frac{13}{3}\xi_0 + \frac{55}{12}\xi_1 &= \frac{1}{8} \\ \frac{55}{12}\xi_0 + -\frac{29}{3}\xi_1 &= \frac{3}{4} \end{aligned}$$

which can be written in matrix form as :

$$\begin{bmatrix} -\frac{13}{3} & \frac{55}{12} \\ \frac{55}{12} & -\frac{29}{3} \end{bmatrix} \begin{bmatrix} \xi_0 \\ \xi_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{8} \\ \frac{3}{4} \end{bmatrix}$$

$$\xi = \begin{bmatrix} -\frac{13}{3} & \frac{55}{12} \\ \frac{55}{12} & -\frac{29}{3} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{8} \\ \frac{3}{4} \end{bmatrix} = \begin{bmatrix} -\frac{1392}{3007} & -\frac{660}{3007} \\ -\frac{660}{3007} & -\frac{624}{3007} \end{bmatrix} \begin{bmatrix} \frac{1}{8} \\ \frac{3}{4} \end{bmatrix} = \begin{bmatrix} -\frac{669}{3007} \\ -\frac{1101}{6014} \end{bmatrix}.$$

Hence, $\xi = \begin{bmatrix} -\frac{669}{3007} & -\frac{1101}{6014} & 0 \end{bmatrix}^T$.

Therefore, $u_h = \sum_{j=0}^{2} \xi_j \varphi_j = -\frac{669}{3007}\varphi_\circ - \frac{1101}{6014}\varphi_1 + 0\varphi_2 = -\frac{669}{3007}\varphi_\circ - \frac{1101}{6014}\varphi_1.$

Example 2. Consider the following BVP:

$$\begin{cases} -w''(x) + 2w(x) = 0, & 0 \;<\; x \;<\; 1 \;, \\ w(0) = \alpha \neq 0, & w(1) = \beta \neq 0 \end{cases} \tag{3.22}$$

on a partition $\mathcal{T}_h$ of the interval $[0,1]$ into subintervals $n+1$ of the length $h = \dfrac{1}{n+1}$.

Solution: The goal is to create an approximate solution w_h in a finite dimensional space spanned by the hat functions $\varphi_j,\ j = 0, 1, \cdots, n+1$ on the partition $\mathcal{T}_h$.

The continuous solution is assumed to be in the Hilbert space

$$H^1 = \Big\{ v : \int\limits_0^1 \Big(v(x)^2 + v'(x)^2 \Big)\, dx < \infty \Big\}$$

Since both $w(0) = \alpha$ and $w(1) = \beta$ are provided, we need to use the trial functions in

$$V := \Big\{ v : v \in H^1, \quad v(0) = \alpha,\ v(1) = \beta \Big\},$$

and the test function in

$$V^0 := H_0^1 = \Big\{ v : v \in H^1, \quad v(0) = \ v(1) = 0 \Big\}.$$

We multiply (3.22) with a test function $v \in V^0$ and integrate over $(0, 1)$

$$\begin{aligned} -w'(1)v(1) + w'(0)v(0) + \int\limits_0^1 w'v'\, dx + 2\int\limits_0^1 wv\, dx = 0 \Longleftrightarrow \\ \text{find} \quad w \in V \quad \text{so that} \quad \int\limits_0^1 w'v'\, dx + 2\int\limits_0^1 wv\, dx = 0 \quad v \in V^0. \end{aligned}$$

The partition $\mathcal{T}_h$ of $[0, 1]$ into $n+1$ subintervals yields the uniform subintervals $I_1 = [0, h],\ I_2 = [h, 2h], \cdots, I_{n+1} = [nh, (n+1)h]$, that are described by the nodes $x_0 = 0, x_1 = h, \cdots, x_n = nh,\ x_{n+1} = (n+1)h = 1$. The corresponding discrete space for the trail function is

$$V_h := \Big\{ w_h : w_h \text{ is piecewise linear and continuous on } \mathcal{T}_h,\ w(0) = \alpha,\ w(1) = \beta \Big\},$$

and for the test function

$$V_h^0 := \Big\{ v_h : v_h \text{ is piecewise linear and continuous on } \mathcal{T}_h,\ v(0) = v(1) = 0 \Big\}.$$

Figure (3.7) shows φ_j, for $j = 0, 1, \cdots, n+1$.

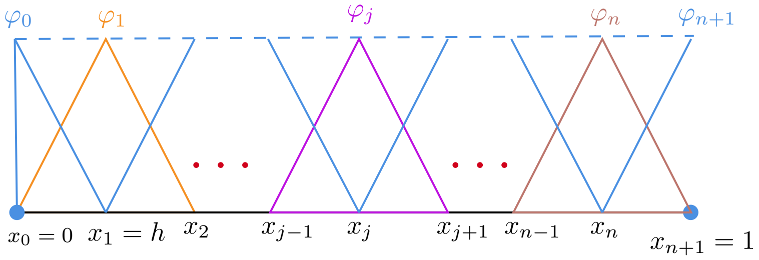


Figure 3.7 – Hat function φ_j including two half-hat functions φ_0 and φ_{n+1}

Now, the finite element solution is : find $w_h \in V_h$ such that

$$\text{(FEM)}: \quad \int_0^1 w_h' v' \, dx + 2\int_0^1 w_h v \, dx = 0, \quad v \in V_h^0.$$

Since $\left\{\varphi_j\right\}_{j=0}^{n+1}$ are basis for V_h, then

$$w_h(x) = \sum_{j=1}^{n+1} \xi_j \varphi_j = \xi_0\varphi_0 + \sum_{j=1}^{n} \xi_j\varphi_j + \xi_{n+1}\varphi_{n+1}(x)$$
$$= \alpha\varphi_0 + \beta\varphi_{n+1} + \sum_{j=1}^{n} \xi_j\varphi_j$$

where

$$\varphi_0(x) = \frac{1}{h}\begin{cases} h-x, & 0 \le x \le h, \\ 0, & O.W \end{cases}, \quad \varphi_j(x) = \frac{1}{h}\begin{cases} x-x_{j-1}, & x_{j-1} \le x \le x_j, \\ x_{j+1}-x, & x_j \le x \le x_{j+1}, \\ 0, & x \notin [x_{j-1}, x_{j+1}] \end{cases}$$

and

$$\varphi_{n+1}(x) = \frac{1}{h}\begin{cases} x-x_n, & nh \le x \le (n+1)h, \\ 0, & O.W \end{cases}.$$

Inserting w_h into (FEM) and choosing $v = \varphi_i, \quad i = 1, \cdots, n$ yield

$$\sum_{j=1}^{n}\Big(\int_0^1 \varphi_j'(x)\varphi_i'(x)\,dx + 2\int_0^1 \varphi_j(x)\varphi_i(x)\,dx\Big)\xi_j$$
$$= -\Big(\int_0^1 \varphi_0'(x)\varphi_i'(x)\,dx + 2\int_0^1 \varphi_0(x)\varphi_i(x)\,dx\Big)\alpha +$$
$$-\Big(\int_0^1 \varphi_{n+1}'(x)\varphi_i'(x)\,dx + 2\int_0^1 \varphi_{n+1}(x)\varphi_i(x)\,dx\Big)\beta$$

This corresponds to $A\xi = b$ with $A = S + 2M$ where S is the stiffness matrix, and M is the mass matrix. The stiffness matrix S, which is completed before is given by:

$$\textbf{stiffness matrix} \quad S = \frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}$$

The mass matrix M, for uniform mesh, is completed as follows

$$M = \left\{m_{ij}\right\}_{i,j=1}^{n}, \quad m_{ij} = \int\limits_0^1 \varphi_j(x)\varphi_i(x)\,dx.$$

(1) If $|i - j| \ > 1$, then

$$\int\limits_0^1 \varphi_j(x)\varphi_i(x)\,dx = 0.$$

(2) for i = j,

$$\begin{aligned} m_{ii} &= \int\limits_0^1 \varphi_j(x)^2\,dx = \frac{1}{h^2}\Big(\int\limits_{x_{i-1}}^{x_i} (x - x_{i-1})^2\,dx + \int\limits_{x_i}^{x_{i+1}} (x_{i+1} - x)^2\,dx \Big) \\ &= \frac{1}{h^2}\left[\frac{(x - x_{i-1})^3}{3}\right]_{x_{i-1}}^{x_i} - \frac{1}{h^2}\left[\frac{(x_{i+1} - x)^3}{3}\right]_{x_i}^{x_{i+1}} \\ &= \frac{1}{h^2}\cdot\frac{h^3}{3} + \frac{1}{h^2}\cdot\frac{h^3}{3} \\ &= \frac{2}{3}h, \quad i = 1, \cdots, n. \end{aligned}$$

(3) for i = j+1,

$$
\begin{aligned}
m_{j+1,j} &= \int_0^1 \varphi_j(x)\varphi_{j+1}(x)\,dx = \frac{1}{h^2}\int_{x_j}^{x_{j+1}} (x_{j+1}-x)(x-x_j) \\
&= \frac{1}{h^2}\Big[(x_{j+1}-x)\frac{(x-x_j)^2}{2}\Big]_{x_j}^{x_{j+1}} - \frac{1}{h^2}\int_{x_j}^{x_{j+1}} -\frac{(x-x_j)^2}{2}\,dx \\
&= \frac{1}{h^2}\Big[\frac{(x-x_j)^3}{6}\Big]_{x_j}^{x_{j+1}} \\
&= \frac{1}{6}h, \quad j = 1,\cdots,n-1.
\end{aligned}
$$

Note that $m_{ij} = m_{ji}$, $\forall i, j$, this implies $m_{j,j+1} = m_{j+1,j}$, i.e., the mass matrix is symmetric.
Thus, for uniform mesh, the entries of the mass matrix M are

$$
m_{ij} = m_{ji} = \begin{cases} 0, & \text{for } |i-j| > 1, \\ \frac{2}{3}h, & \text{for } i = j, \\ \frac{1}{6}h, & \text{for } |i-j| = 1. \end{cases}
$$

Hence, the mass matrix is

$$
M = h\begin{bmatrix} \frac{2}{3} & \frac{1}{6} & 0 & 0 & \cdots & 0 \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & 0 & \cdots & 0 \\ 0 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} & \cdots & 0 \\ & \ddots & \ddots & \ddots & & \\ & & & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ & & & & \frac{1}{6} & \frac{2}{3} \end{bmatrix} = \frac{h}{6}\begin{bmatrix} 4 & 1 & 0 & 0 & \cdots & 0 \\ 1 & 4 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 4 & 1 & \cdots & 0 \\ & \ddots & \ddots & \ddots & & \\ & & & 1 & 4 & 1 \\ & & & & 1 & 4 \end{bmatrix}.
$$

Back to our example, for $i, j = 1,\cdots,n$, the coefficient matrix $A = S + 2M$ is given as

$$
\big[A\big]_{ij} = \int_0^1 \varphi_i'\varphi_j'\,dx + 2\int_0^1 \varphi_i\varphi_j\,dx = \begin{cases} \frac{2}{h} + \frac{4h}{3}, & i = j, \\ -\frac{1}{h} + \frac{h}{3}, & |i-j| = 1, \\ 0, & O.W \end{cases}.
$$

Finally, with $\xi_0 = \alpha$ and $\xi_{n+1} = \beta$, the load vector is given by

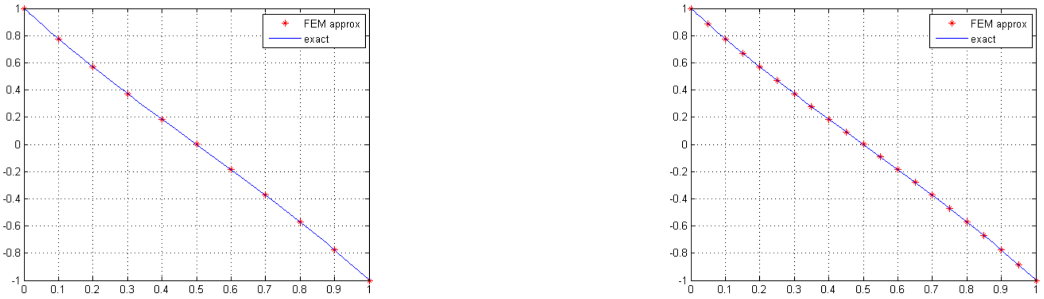
$$b = \begin{bmatrix} -\Big(-\dfrac{1}{h} + \dfrac{h}{3}\Big)\xi_0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ -\Big(-\dfrac{1}{h} + \dfrac{h}{3}\Big)\xi_{n+1} \end{bmatrix} = \begin{bmatrix} \alpha\Big(\dfrac{1}{h} - \dfrac{h}{3}\Big) \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \beta\Big(\dfrac{1}{h} - \dfrac{h}{3}\Big) \end{bmatrix}.$$

If we take $n = 3$ then $h = \dfrac{1}{4}$, and thus the stiffness and mass matrices and the load vector becomes

$$S = \begin{bmatrix} 8 & -4 & 0 & 0 & \cdots & 0 & \\ -4 & 8 & -4 & 0 & \cdots & 0 & \\ 0 & -4 & 8 & -4 & \cdots & 0 & \\ & \ddots & \ddots & \ddots & & & \\ & & & -4 & 8 & -4 & \\ & & & & -4 & 8 & \end{bmatrix},$$

$$M = \frac{1}{24}\begin{bmatrix} 4 & 1 & 0 & 0 & \cdots & 0 & \\ 1 & 4 & 1 & 0 & \cdots & 0 & \\ 0 & 1 & 4 & 1 & \cdots & 0 & \\ & \ddots & \ddots & \ddots & & & \\ & & & 1 & 4 & 1 & \\ & & & & 1 & 4 & \end{bmatrix}, \text{and } b = \begin{bmatrix} \frac{47}{12}\alpha \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \frac{47}{12}\beta \end{bmatrix}.$$

Now, using the Matlab software we get the below finite element approximation for $\alpha = 1$, $\beta = -1$. The exact solution in this case is $w(x) = C_1\, e^{\sqrt{2}x} + C_2\, e^{-\sqrt{2}x}$, where $C_1 = 1 - C_2$, and $C_2 = \dfrac{-(1 + e^{\sqrt{2}})}{e^{-\sqrt{2}} - e^{\sqrt{2}}}$.



(a) Finite Element approximation at $n = 10$

(b) Finite Element approximation at $n = 20$

Figure 3.8 – Finite element approximation with the exact solution for n=10 and n=20

From figure 3.8 , it seems that the approximation is the same as the exact solution at the nodal points, but when we zoom in, the error becomes clear.

Example 3. Consider the problem of the elliptical boundary value

$$\begin{cases} -\Delta u = 8\pi^2 \sin 2\pi x \ \sin 2\pi y, & x, y \in \Omega = [0,1] \times [0,1] \ , \\ u = 0, & \text{on } \Gamma \ . \end{cases} \tag{3.23}$$

where the exact solution is $u(x, y) = \sin 2\pi x \sin 2\pi y$, and $g(x, y) = 8\pi^2 \sin 2\pi x \sin 2\pi y$

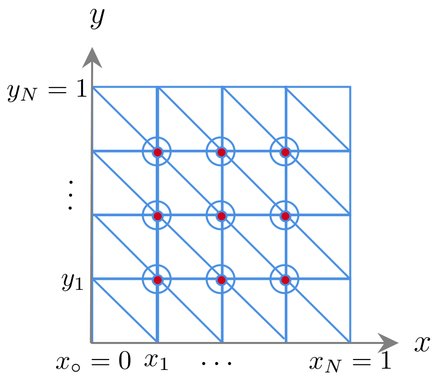


Figure 3.9 – Triangulation of $\bar{\Omega} = [0, 1] \times [0, 1]$.

$$\varphi_{ij} = \begin{cases} 1 - \frac{x-x_i}{h} - \frac{y-y_j}{h}, & (x,y) \in 1, \\ 1 - \frac{y-y_j}{h}, & (x,y) \in 2\ , \\ 1 - \frac{x_i-x}{h}, & (x,y) \in 3\ , \\ 1 - \frac{x_i-x}{h} - \frac{y_j-y}{h}, & (x,y) \in 4\ , \\ 1 - \frac{y_j-y}{h}, & (x,y) \in 5\ , \\ 1 - \frac{x-x_i}{h}, & (x,y) \in 6\ , \\ 0, & \text{else where.} \end{cases}$$

where each of 1 , 2 , 3 , 4 , 5 and 6 denotes the triangle number around the node (x_i, y_j) as shown in Figure (3.10).

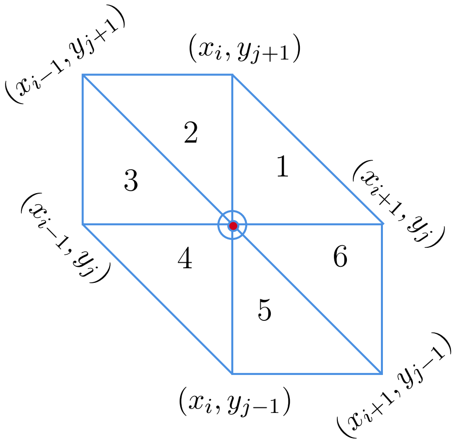


Figure 3.10 – Triangles surrounding the node (x_i, y_j)

The partial derivatives of this basis functions are

$$\frac{\partial \varphi_{ij}}{\partial x} = \begin{cases} \frac{-1}{h}, & (x,y) \in 1\ , \\ 0, & (x,y) \in 2\ , \\ \frac{1}{h}, & (x,y) \in 3\ , \\ \frac{1}{h}, & (x,y) \in 4\ , \\ 0, & (x,y) \in 5\ , \\ \frac{-1}{h}, & (x,y) \in 6\ , \\ 0, & \text{else where.} \end{cases} \quad \text{and} \quad \frac{\partial \varphi_{ij}}{\partial y} = \begin{cases} \frac{-1}{h}, & (x,y) \in 1\ , \\ \frac{-1}{h}, & (x,y) \in 2\ , \\ 0, & (x,y) \in 3\ , \\ \frac{1}{h}, & (x,y) \in 4\ , \\ \frac{1}{h}, & (x,y) \in 5\ , \\ 0, & (x,y) \in 6\ , \\ 0, & \text{else where.} \end{cases}$$

To find the variational formulation of the problem multiply (3.23) by a test function

v, and integrate over Ω, we get

$$\int\limits_{\Omega} -\Delta uv\, dx = \int\limits_{\Omega} gv\, dx,$$

$$\int\limits_{\Omega} \nabla u \cdot \nabla v\, dx - \int\limits_{\Gamma} \nabla u \cdot n\, ds = \int\limits_{\Omega} gv\, dx,$$

Since $u = 0$ on Γ, then

$$\int\limits_{\Omega} \nabla u \cdot \nabla v\, dx = \int\limits_{\Omega} gv\, dx,$$

$$\equiv \quad a(\nabla u, \nabla v) = L(\nabla v).$$

Find u_h such that

$$a(\nabla u_h, \nabla v) = L(\nabla v),$$

uses the basis function, and let

$$u_h = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \xi_{ij}\varphi_{ij}, \quad v = \varphi_{kr}, \quad k, r = 1, 2, \cdots, N-1.$$

where $N-1$ is the number of internal nodes,

$$\begin{aligned}
&\sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \xi_{ij} \int\limits_{\Omega} \nabla\varphi_{ij} \cdot \nabla\varphi_{kr}\, d\Omega \\
&= \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \xi_{ij} \int\limits_{\Omega} \left(\frac{\partial\varphi_{ij}}{\partial x}\frac{\partial\varphi_{kr}}{\partial x} + \frac{\partial\varphi_{ij}}{\partial y}\frac{\partial\varphi_{kr}}{\partial y} \right) dx\, dy \\
&= \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} \xi_{ij} \int\limits_{\sup \varphi_{kr}} \left(\frac{\partial\varphi_{ij}}{\partial x}\frac{\partial\varphi_{kr}}{\partial x} + \frac{\partial\varphi_{ij}}{\partial y}\frac{\partial\varphi_{kr}}{\partial y} \right) dx\, dy \\
&= 4\xi_{kr} - \xi_{k-1,r} - \xi_{k+1,r} - \xi_{k,r-1} - \xi_{k,r+1} \quad k, r = 1, 2, \cdots, N-1.
\end{aligned}$$

the approximation of the finite element is equal to

$$\begin{aligned}&-\frac{\xi_{k+1,r}+\xi_{k-1,r}-2\xi_{k,r}+\xi_{k,r+1}+\xi_{k,r-1}-4\xi_{k,r}}{h^2}\\&=\frac{1}{h^2}\iint\limits_{\sup\varphi_{kr}} g(x,y)\varphi_{kr}(x,y)\,dx\,dy \quad k,l=1,2,\cdots,N-1,\end{aligned}$$

$$\text{and} \quad \xi_{kr}=0 \quad \text{on} \quad \Gamma.$$

Thus, in this particular triangulation, the rounding of the selected element gives rise to the familiar 5-point finite difference diagram with the computation of the mean of the power function g in a special way.

Figure (3.11) shows the basis function at a specific nodal point, this function is also called tent function.

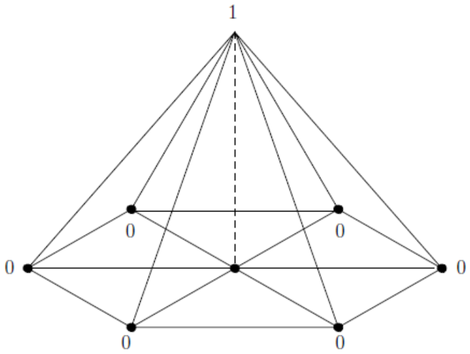


Figure 3.11 – φ is considered to be continuous in $\bar{\Omega}$ and linear in each triangle.

Using the Matlab software, the following non-uniform triangulation is generated, (Figure 3.12).

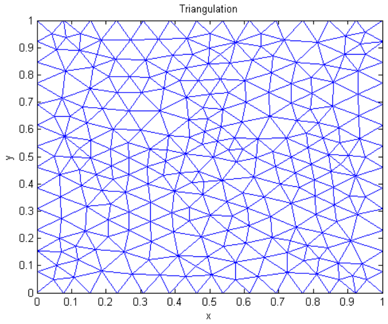
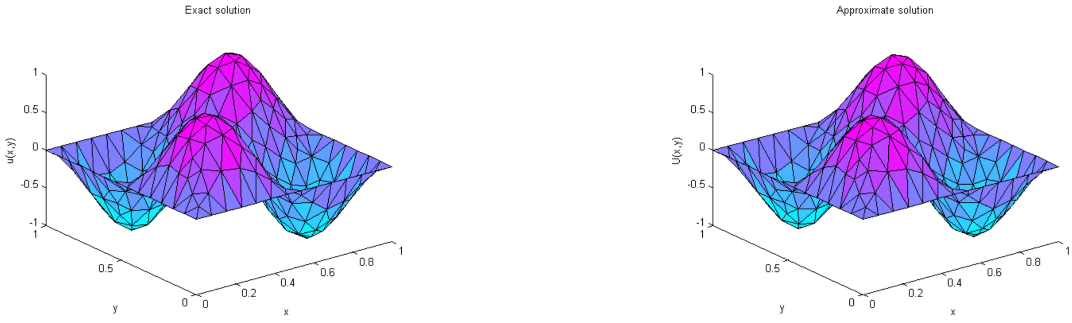


Figure 3.12 – The triangulation of randomly generated mesh of Ω at $h_{\max}=0.08$

The exact solution and the approximation are shown in Figure 3.13.



(a) The Exact solution

(b) Th Approximate solution

Figure 3.13 – The exact and the approximate solutions

Here, the triangulation is generated in such a way that the maximum edge size is equal to 0.08.

Chapter 4

Error estimation

4.1 Introduction

Errors are very important in numerical analysis, there are many types of errors including rounding error, truncation error, data error and model instability. One of the most important error problems is the ability to control error which can be very helpful in solving an error estimation problem because it can help us evaluate the solution or the model itself.

Mathematical theory of error estimation is one of the main and fundamental considerations of computational science. It was noticed that many of the analytical findings used in mathematical models involved numerical errors. This will actually help us to determine the efficiency of the numerical process. The use of error measurements to control temporal steps in the numerical solution of ordinary differential equations is perhaps the first use of posteriori-estimates to control error estimation in numerical solutions of prime or boundary value problems.

In fact, the error in numerical solution is defined as the difference between exact solution u and estimated solutions u_h, [Braess (2007a)]. The aim of the error estimation is to prevent inaccuracies in the numerical solution. Briefly,the error is defined as $e = u - u_h$, where u is the exact solution to the variational problem

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

and u_h is the approximation to the variation problem

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

Error analysis of the Method of Finite Element

Error analysis for finite element method usually includes two parts

1. Error estimates for an intermediate function in V_h, often the interpolation function, and
2. Convergence analysis, a restricted method that shows the finite element solution converges to the true solution of the weak form in some norm as the mesh size h reaches zero, [Rauch (1997)].

Methodology 4.1.1.

a) Given the weak form $a(u, v) = L(v)$ and the space V, which normally has an infinite dimension, the problem is to find $u \in V$ so that the weak model is satisfied with every $v \in V$. Then u is called the weak form solution.

b) The finite-dimensional subspace of V denoted as V_h (i.e., $V_h \subset V$) is used for the finite element method.

c) The solution to the weak form in the subspace V_h is denoted by u_h, i.e., we need $a(u_h, v_h) = L(v_h)$ for any $v_h \in V_h$.

d) The general error is defined by $e_h = u(x) - u_h(x)$, and we are looking for a sharp upper bound for $\|e_h\|$ using certain norm.

Notation 5.
We use the FEM to obtain the approximate solution u_h of a PDE, the key question is " How large is the error $e = u - u_h$?", [Schopf (2014)]. Some ingredients are required to be able to estimate the error:

1. **Galerkin Orthogonality**.
2. **Interpolation Estimate**.
3. **Coercivity** .

4.1.1 Galerkin Orthogonality

Theorem 4.1.1. ***(Galerkin Orthogonality)***

(1) u_h is a projection of u to V_h through the inner product $a(u, w)$, [Rauch (1997)] see Figure (4.1)

$$u - u_h \perp V_h, \quad i = 1, 2, \cdots, N. \quad \textbf{(Galerkin Orthogonality)}$$

(2) u_h is the best approximation in terms of energy norm,that is

$$\|u - u_h\|_a \leq \|u - w_h\|_a, \quad \forall w_h \in V_h.$$

Proof. (1), Note that

$$\begin{aligned} a(u,w) &= L(w), \qquad \forall w \in V, \\ a(u_h,w_h) &= L(w_h), \qquad \forall w_h \in V_h. \end{aligned}$$

By subtracting these two equations and noting that $V_h \subset V$ we get

$$\begin{aligned} a(u-u_h,w) &= 0, \quad \forall w \in V_h, \\ or \quad a(e,w_h) &= 0, \qquad \forall w_h \in V_h. \end{aligned}$$

The second part is the Galerkin Orthogonality proves (2), We want to prove that u_h is the best approximation in V_h

$$\begin{aligned} \|u-u_h\|_a^2 &= a(u-u_h,u-u_h) \\ &= a(u-u_h,u-w_h+w_h-u_h) \\ &= a(u-u_h,u-w_h) + a(u-u_h,w_h-u_h) \\ &= a(u-u_h,u-w_h) + 0 \quad \text{(by Galerkin Orthogonality)} \\ &\le \|u-u_h\|_a \|u-w_h\|_a \\ \Longrightarrow \|u-u_h\|_a &\le \|u-w_h\|_a, \qquad \forall w_h \in V_h. \end{aligned}$$

Thus, the second assertion is proved. □

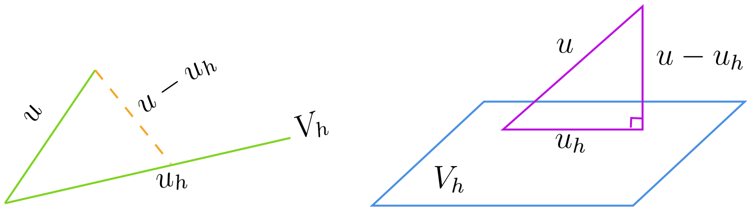


Figure 4.1 – Finite element approximation properties diagram.

4.1.2 Interpolation Estimate

First of all, let us remember that

$$\|u-u_h\| \ge \inf_{w \in V_h} \|u-w\| \quad \text{for some norm.}$$

We need to be able to estimate the value of $\inf\limits_{w \in V_h} \|u-w\|$ or at least get a sharp upper bound. We will do this by estimating $\|u-w\|$ for a particular choice of w.

Suppose $\pi_h u$ is a constant approximation of $u(x)$, then for $x \in [x_{k-1}, x_k]$ of Taylor's expansion theorem we have

$$u(x) = \underbrace{u\left(\overbrace{\frac{x_{k-1}+x_k}{2}}^{\hat{x}_k}\right)}_{\pi_h u} + \int_{\hat{x}_k}^{x} u'(y)\,dy.$$

This leads to

$$\Big|u - \pi_h u\Big| = \Big|\int_{\hat{x}_k}^{x} u'(y)\,dy\Big|.$$

Let us consider the L^2- norm, then

$$\begin{aligned}
\|u - \pi_h u\|^2 &= \int_a^b (u - \pi_h u)^2\,dx = \sum_k \int_{x_{k-1}}^{x_k} (u - \pi_h u)^2\,dx \\
&= \sum_k \int_{x_{k-1}}^{x_k} \Big(\int_{\hat{x}_k}^{x} u'(y)\,dy\Big)^2\,dx \\
&= \sum_k \int_{x_{k-1}}^{x_k} \Big(\int_{\hat{x}_k}^{x} 1 \cdot u'(y)\,dy\Big)^2\,dx \\
&\le \sum_k \int_{x_{k-1}}^{x_k} \left(\Big(\int_{\hat{x}_k}^{x} 1^2\,dy\Big)^{\frac{1}{2}} \cdot \Big(\int_{\hat{x}_k}^{x} (u'(y))^2\,dy\Big)^{\frac{1}{2}}\right)^2\,dx \\
&= \sum_k \int_{x_{k-1}}^{x_k} \Big|\int_{\hat{x}_k}^{x} 1^2\,dy\Big| \cdot \Big|\int_{\hat{x}_k}^{x} (u'(y))^2\,dy\Big|\,dx \\
&= \sum_k \int_{x_{k-1}}^{x_k} \left|x - \frac{x_{k-1}+x_k}{2}\right| \cdot \int_{\hat{x}_k}^{x} (u'(y))^2\,dy\,dx \\
&\le \sum_k \frac{h_k}{2} \int_{x_{k-1}}^{x_k} \int_{\hat{x}_k}^{x} (u'(y))^2\,dy\,dx \\
&= \sum_k \frac{h_k^2}{2} \int_{x_{k-1}}^{x_k} (u'(y))^2\,dy \\
&\le \frac{1}{2} \int_a^b (hu'(y))^2\,dy \\
&= \frac{1}{2} \|hu'\|_{L^2}^2,
\end{aligned}$$

where $h_k = x_k - x_{k-1}$ and $h = \max_k h_k$. Therefore we have found an interpolation estimate

$$\|u - \pi_h u\|_{L^2} \leq \frac{1}{\sqrt{2}} \|h\, u'\|_{L^2}. \tag{4.3}$$

In general, the following estimate holds

$$\|D^\alpha(u - \pi_h u)\|_{L^2} \leq C(\alpha, \beta) \|h^{\beta+1-\alpha} D^{\beta+1}\, u\|_{L^2}, \tag{4.4}$$

where $\pi_h u$ is an interpolation of u of degree β, and $C(\alpha,\beta)$ is a constant depending only on α and β, [Kirby and Logg (2012)].

Remark 10.

$$\pi_h^s v \in V_h^s \ : \ \pi_h^s v\Big|_{k_j} = \pi_{k_j}^s (v\Big|_{k_j}) \quad \forall k_j \in \mathcal{T}_h. \tag{4.5}$$

Theorem 4.1.2. Let $v \in H^{s+1}(I)$, for $s \geq 1$, and let $\pi_h^s(v) \in V_h^s$ be its interpolating function defined in (4.5). The following estimate of the interpolation error hold

$$\left| v - \pi_h^s v \right|_{H^k(I)} \leq C_{k,s}\, h^{s+1-k} \left| v \right|_{H^{s+1}(I)}, \quad \text{for} \quad k = 0, 1$$

The constant $C_{k,s}$ are independent of v and h. We recall that $H^0(I) = L^2(I)$ and that $| \cdot |_{H^0(I)} = \| \cdot \|_{L^2(I)}$, [Quarteroni (2014)].

Proposition 4.1.1. The following interpolation error estimate in 2D hold : [Larson and Bengzon (2013)]

$$\|v - \pi v\|_{L^2(k)} \leq C h_k^2 \|D^2 v\|_{L^2(k)},$$
$$\|D(v - \pi v)\|_{L^2(k)} \leq C h_k \|D^2 v\|_{L^2(k)},$$

where C : a constant independent of v and h,
k : the element of partition $\mathcal{T}_h$,
h_k : the diameter of element k,
$\pi_h v$: is linear interpolate operator,
D : is the derivative operator which equal $\left((\frac{\partial}{\partial x})^2 + (\frac{\partial}{\partial y})^2 \right)^2$.

4.1.3 *Coercivity*

Definition 4.1.1. A bilinear form $a : V \times V \longrightarrow \mathbb{R}$ is called coercive if there is a constant $\alpha > 0$

$$a(w, w) \geq \alpha \|w\|_V^2, \qquad \forall w \in V.$$

4.2 *Error Estimation*

In this section, we have two kinds of error estimation, [Kirby and Logg (2012)],

1. A priori : $e = e(u)$.
2. A posteriori : $e = e(u_h)$.

Error estimates generally have the form $\underbrace{\|u - u_h\| \leq C(h)}_{*}$, where

u : is the exact solution
u_h : is the approximation,
h : is a mesh parameter,
$C(h)$: is an h function.

The difference between a posteriori and a priori estimates is simply as follows, [John (2016)] and [Kawecki et al. (2018)].

- A priori estimates: the right hand-side of $(*)$ depends on h and u, but not on u_h.
- A posteriori estimations : the right hand-side of $(*)$ depends on h and u_h but not on u.

4.2.1 A Priori error estimates

A priori estimate (also called a priori bound) is a Latin expression meaning before and denote the fact that an estimate of a solution is derived before a solution was present. They are known before the solution is formed, and called a priori estimates. Predicting errors in numerical methods has always been a project of numerical analysis. These estimates give information about the convergence and stability of the various solutions and give approximate information about the proximal behavior of errors in the calculations where the network parameters vary appropriately.
Priori error estimators offer behavioral information that approximates prediction

errors but is not meant to provide an actual error estimation for a given network, and this plays an important role in demonstrating the existence of a solution. These estimates also provide us with an excellent tool for dealing with a very realistic problem.
There are various methods that provide advance estimates of solutions to elliptical problems, [Kelemen and Quittner (2010)]. The first method known as blow-up was first introduced by B. Gidas and J. Spruck in [GIDAS (1981)]. Another method is the Rellich-Pohozaev Identity and Moving Planes method, implemented by D.G.de Turkmenistan, P. -L. Lions and R.D. Nussbaum [Nussbaum (1975)]. Moreover, we have the Hardy - Sobolev inequalities method by H.Brezis and R.E.L Turner [Brezis and Turner (1977)]. Finally, the boot procedure by Ph. Souplet and P. Quittner, [Quittner and Souplet (2004)].

Assume that $a(\cdot, \cdot)$ is a symmetric and coercive form. The form $a(\cdot, \cdot)$ is an inner product and $\|w\|_E = \sqrt{a(w,w)}$ is called energy norm. Let us look at the energy norm error

$$\begin{aligned}
\|e\|_E^2 &= a(e,e) = a(e, u-u_h) \\
&= a(e, u-v+v-u_h), \\
&= a(e,u-v) + a(e, \underbrace{v-u_h}_{\in V_h}), \\
&= a(e,u-v) + 0, \quad \text{(from Galerkin Orthogonality)} \\
&= a(e,u-v) \\
\|e\|_E^2 &\leq \|e\|_E \|u-v\|_E. \quad \text{(by Cauchy-Schwartz Inequality)}
\end{aligned}$$

Thus, Divide both side by $\|e\|_E$ to get

$$\begin{aligned}
\|e\|_E &\leq \|u-v\|_E \\
\|u-u_h\|_E &\leq \|u-v\|_E, \quad \forall v \in V_h.
\end{aligned}$$

As a result, the finite element solution is the optimum solution in the energy norm!. We combine this with the interpolation estimate (4.3) by setting $v = \pi_h u$, then

$$\begin{aligned}
\|u-u_h\|_E &\leq \|u-\pi_h u\|_E \\
&\leq C(\alpha,\beta)\|h^{\beta+1-\alpha} D^{\beta+1} u\|, \quad \text{(by (4.4))}
\end{aligned}$$

Special case, put $\beta = 1$, $\alpha = 0$ in a linear interpolation $\pi_h u$, then the a priori estimate becomes

$$\|u-u_h\|_E \leq C\|h^2\, D^2\, u\|.$$

Also, the calculation of the a priori error in the V-norm for bilinear that are not symmetric follows from the coersivity property as follows

$$\begin{aligned}
\|e\|_V^2 &\leq \frac{1}{\eta} a(e,e), \quad (\eta > 0) \\
&= \frac{1}{\eta} a(e, u - u_h) \\
&= \frac{1}{\eta} a(e, u - v + v - u_h), \quad v \in V_h \\
&= \frac{1}{\eta} \left(a(e, u - v) + a(e, v - u_h) \right) \\
&= \frac{1}{\eta} a(e, u - v), \quad \text{(by Galerkin Orthogonality)} \\
&\leq \frac{C}{\eta} \|e\|_V \|u - v\|_V.
\end{aligned}$$

Divide both sides by $\|e\|_V$ to get the next inequality which called **Cea′s lemma**,

$$\|e\|_V \leq \frac{C}{\eta} \|u - v\|_V, \quad \forall v \in V_h. \tag{4.6}$$

Also, we can use the interpolation estimation $(v = \pi_h u)$

$$\|e\|_V \leq \frac{C \cdot C(\alpha, \beta)}{\eta} \|h^{\beta+1-\alpha} D^{\beta+1}\, u\|. \tag{4.7}$$

4.2.2 *A posteriori estimation of error*

A posteriori error estimates aim to set an error bound between the known numerical approximation and the unknown exact solution that can be determined in practice until the estimated solution is known. Usually, they take the form of a fixed problem,

$$\|u - u_h\| \leq \Big\{ \sum_{k \in \mathcal{T}_h} \eta_k^2 \Big\}^{\frac{1}{2}}, \tag{4.8}$$

where $\eta_k = \eta_k(u_h)$ is a quantity linked to the mesh element k, compatible with u_h. This quantity is called an element estimator or (the estimator variable). In this method, the bilinear form $a(\,\cdot\,,\,\cdot\,)$ does not have to be divided into symmetrical and non-symmetrical parts, here we deduce residual based a posterior error estimation expressed in residual and therefore computable term.

4.3 Error Estimator for Poisson Equation with Homogeneous condition

The Poisson equation is the model problem for elliptical partial differential equation, a lot like that the heat and wave equations are for parabolic and hyperbolic PDEs, [James et al. (2013)].
One of the most important mathematical equations of physical phenomenon models is the Poisson equation. Just like an example, the solution of this equation gives the electrostatic potential of the distribution charge of the equation. It also appears frequently in structural mechanics, theoretical physics such as gravitation, electromagnetism, elasticity and many other fields of research and engineering.
The Poisson equation is named after Siméen-Denis Poisson, a French mathematician. The Poisson equation is

$$-\Delta u = h(x), \quad x \in \Omega,$$

Where Ω is the n-dimensional space. The unknown function u, e.g., as electrostatic potential data of h, is the distribution of the charge. Note that when the potential data $h = 0$, then the equation becomes $-\Delta u = 0$ and is called the Laplace equation.

Theorem 4.3.1 **(Clément approximation operator)**
Let $\mathcal{T}$ be a semi-regular triangle, then there is a linear map $\pi_h : H^1 \to V_h$ with the following properties : [Braess and Verfürth (1996)).

$$\|v - \pi_h v\|_{L^2(k)} \leq c h_k \|v\|_{H^1(\tilde{k})}, \quad \forall v \in H^1(\Omega),$$
$$\|v - \pi_h v\|_{L^2(\gamma)} \leq c h_k^{\frac{1}{2}} \|v\|_{H^1(\tilde{k})}, \quad \forall v \in H^1(\Omega).$$

where

h : the diameter of the element k,

c : an interpolation constant that depends on the shape of the element for our model problem,

$\tilde{k}$: Sub-domain of k that shares starting edge with k,

and $\tilde{k} = \Big\{ \bigcup k', \, k' \in \mathcal{T}_h : \bar{k}' \bigcap \bar{k} \neq \phi \Big\}$.

4.3.1 A posteriori error estimator

Consider the problem

$$\begin{cases} -\Delta u = h & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases} \tag{4.9}$$

where $\Omega \subset \mathbb{R}^2$ is the polygonal domain with continuous lipschitz boundaries Γ and $h \in L^2(\Omega)$. The weak formulation of equation (4.9) is to find $u \in H_0^1(\Omega)$ such that

$$a(u,v) = L(v), \quad v \in V. \tag{4.10}$$

where, V is a subspace of $H_0^1(\Omega)$ defined below, and

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

$$L(v) = \int_\Omega hv \, dx.$$

The form $a(u,v)$ is V- elliptic bilinear from $V \times V \to \mathbb{R}$ and the linear functional $L(v)$ is an element of the double space V', where

- $*$ $V = \Big\{ v : v$ continuous on $\overline{\Omega}$, $v = 0$ on $\Gamma \Big\}$.
- $*$ $V' = \Big\{$is the dual space of the vector space V : is the set of linear functional on $V\Big\}$.

Consider the finite element solution $u_h \in V_h$ satisfying

$$a(u_h, v_h) = L(v_h), \quad v_h \in V_h \subset V. \tag{4.11}$$

The quantity

$$\begin{aligned} a(e,v) &= a(u - u_h, v) \\ &= a(u,v) - a(u_h, v) \\ &= L(v) - a(u_h, v) \\ &= \mathcal{R}(v), \quad \forall v \in V. \end{aligned}$$

Notation 6.

(A) $\mathcal{R}(\ \cdot\ )$ is called the weak residual value.

(B) $\mathcal{R}(u_h) = 0, \quad \forall v_h \in V_h$, because the error $e \in V$ fulfills the following residual equation.

$$a(e,v) = \mathcal{R}(v), \quad \forall v \in V. \tag{4.12}$$

By the Galerkin orthogonal

$$a(e, v_h) = \mathcal{R}(v_h) = 0, \quad \forall v_h \in V_h. \tag{4.13}$$

(C) The residual function norm is equivalent to the energy error norm.

$$\|\mathcal{R}\|_{V'} = \sup_{v \in V} \frac{|\mathcal{R}(v)|}{\|v\|_E} = \sup_{v \in V} \frac{|a(e,v)|}{\|v\|_E} = \|e\|_E.$$

i.e., $\|\mathcal{R}\|_{V'} \equiv \|e\|_E$.

We will divide the integral into sum of integrals over the triangulation $k \in \mathcal{T}_h$, [Hackbusch (2017)] and [Hicks et al. (2014)].
That is, we can write $\mathcal{R}(u)$ in terms of the interior and edge contribution.

$$\begin{aligned}
\mathcal{R}(v) &= L(v) - a(u_h, v) \\
&= \int_\Omega hv \, dx - \int_\Omega \nabla u_h \cdot \nabla v \, dx \\
&= \sum_{k \in \mathcal{T}_h} \int_k \Big(hv - \nabla u_h \cdot \nabla v \Big) \, dx \\
&= \sum_{k \in \mathcal{T}_h} \int_k \Big(hv + \Delta u_h v \Big) \, dx + \sum_{\gamma \in \partial\mathcal{T}_h} \int_\gamma \mathcal{J}(\nabla u_h) v \, ds \\
&= \sum_{k \in \mathcal{T}_h} \int_k \Big(h + \Delta u_h \Big) v \, dx + \sum_{\gamma \in \partial\mathcal{T}_h} \int_\gamma \mathcal{J}(\nabla u_h) v \, ds.
\end{aligned}$$

where $\mathcal{T}_h$: *is triangulation family on* $\overline{\Omega}$,
k : *is an element in* $\mathcal{T}_h$,
$\mathcal{J}(\nabla u_h)$: *is the move of* u_h *over the edge of* γ,
i.e., $\mathcal{J}(\nabla u_h)$= $(\nabla u_h^+ - \nabla u_h^-) \cdot n_\gamma$,
n_γ : is the normal unit at the edge of γ.

Notation 7. The direction of the normal unit n_γ to the edge γ is irrelevant and $\nabla u_h^\pm = \lim_{s \to o_+} \nabla u_h(x \pm s n_\gamma)$.

Next, evaluate the residual equation (4.12) at e and use (4.13) to get

$$a(e, e) = \mathcal{R}(e - w_h), \quad \forall w_h \in V_h. \tag{4.14}$$

Thus,

$$\|e\|_E^2 = a(e,e) = \sum_{k \in \mathcal{T}_h} \int_k \Big(h + \Delta u_h \Big) e \, dx + \sum_{\gamma \in \partial\mathcal{T}_h} \int_\gamma \mathcal{J}(\nabla u_h) e \, ds. \tag{4.15}$$

Now, we need the interpolation factor π_h, substituting $w_h = \pi_h e$ in (4.14) and using the Galerkin orthogonal condition to insert the interpolation $\pi_h e$ in (4.15) to get

$$\|e\|_E^2 = a(e,e) = \sum_{k\in\mathcal{T}_h} \int_k \Big(h+\Delta u_h\Big)(e-\pi_h e)dx + \sum_{\gamma\in\partial\mathcal{T}_h} \int_\gamma \mathcal{J}(\nabla u_h)(e-\pi_h e)ds. \quad (4.16)$$

Since $r = h + \Delta u_h,$ applying Cauchy- Schwartz inequality to get

$$\|e\|_E^2 \leq \sum_{k\in\mathcal{T}_h} \|r\|_{L^2(k)} \ \|e-\pi_h e\|_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} \|\mathcal{J}(\nabla u_h)\|_{L^2(\gamma)} \ \|e-\pi_h e\|_{L^2(\gamma)}. \quad (4.17)$$

Next, to overwrite the properties of Theorem (4.3.1) in (4.17)

$$\|e\|_E^2 \leq \sum_{k\in\mathcal{T}_h} \|r\|_{L^2(k)} \ Ch\|e\|_{H^1(\tilde{k})} + \sum_{\gamma\in\partial\mathcal{T}_h} \|\mathcal{J}(\nabla u_h)\|_{L^2(\gamma)} \ Ch^{\frac{1}{2}}\|e\|_{H^1(\tilde{k})}. \quad (4.18)$$

$$\|e\|_E^2 \leq C\|e\|_{H^1(\tilde{k})}\Bigg\{ \sum_{k\in\mathcal{T}_h} h\|r\|_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h^{\frac{1}{2}}\|\mathcal{J}(\nabla u_h)\|_{L^2(\gamma)}\Bigg\}. \quad (4.19)$$

Using Poincaré inequity $\|e\|_{H^1(\tilde{k})} \leq \|e\|_E$ implies

$$\|e\|_E^2 \leq C\|e\|_E\Bigg\{ \sum_{k\in\mathcal{T}_h} h\|r\|_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h^{\frac{1}{2}}\|\mathcal{J}(\nabla u_h)\|_{L^2(\gamma)}\Bigg\}. \quad (4.20)$$

This is equivalents to

$$\|e\|_E \leq C\Bigg\{ \sum_{k\in\mathcal{T}_h} h\|r\|_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h^{\frac{1}{2}}\|\mathcal{J}(\nabla u_h)\|_{L^2(\gamma)}\Bigg\}. \quad (4.21)$$

Squaring both sides yields

$$\|e\|_E^2 \leq C\Bigg\{ \sum_{k\in\mathcal{T}_h} h\|r\|_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h^{\frac{1}{2}}\|\mathcal{J}(\nabla u_h)\|_{L^2(\gamma)}\Bigg\}^2. \quad (4.22)$$

Applying Young's inequality gives

$$\|e\|_E^2 \leq C\Bigg\{ \sum_{k\in\mathcal{T}_h} h^2\|r\|_{L^2(k)}^2 + \sum_{\gamma\in\partial\mathcal{T}_h} h\|\mathcal{J}(\nabla u_h)\|_{L^2(\gamma)}^2\Bigg\}. \quad (4.23)$$

Now, break the constant C into two constants, C_1 and C_2, in equality (4.23) can be written as

$$\|e\|_E^2 \leq \sum_{k\in\mathcal{T}_h} \Big(C_1 h^2\|r\|_{L^2(k)}^2 + C_2 h\|\mathcal{J}(\nabla u_h)\|_{L^2(\partial k)}^2\Big).$$

Let

$$\eta_k^2 = C_1 h^2 \|r\|_{L^2(k)}^2 + C_2 h \|\mathcal{J}(\nabla u_h)\|_{L^2(\partial k)}^2,$$

then

$$\|e\|_E^2 \leq \sum_{k \in \mathcal{T}_h} \eta_k^2.$$

4.4 Error Estimator for Poisson equation with mixed (Dirichlet- Neumann) boundary condition

Consider the problem

$$\begin{cases} -\Delta u = h, & \text{on } \Omega, \\ u = 0, & \text{in } \Gamma_D, \\ n \cdot \nabla u = g, & \text{in } \Gamma_N. \end{cases} \tag{4.24}$$

with domain $\Omega \subset \mathbb{R}^2$ and lipschitz boundary $\Gamma = \Gamma_D \bigcup \Gamma_N$, where Γ_D : the boundary of Dirichlet, and Γ_N : the boundary of Neumann. The data is assumed to be sufficiently smooth, i.e., $h \in L^2(\Omega)$, $g \in L^2(\Omega)$, and n is the external normal vector to Γ. The variational formulation of the boundary problem is to find $u \in V$ such that

$$a(u, v) = L(v), \quad \forall v \in V, \tag{4.25}$$

where the trail and test space V is a sobolev space from $H^1(\Omega)$ whose trace disappears at the Dirichlet boundary,i.e.,

$$V = \Big\{ v \in H^1(\Omega) : \;\; v = 0 \text{ on } \Gamma_D \Big\}. \tag{4.26}$$

The form $a(u, v)$ is assumed to be the bilinear V-coercive form of $V \times V$, and the linear function $L(v)$ is an element of the dual space V' like as

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx, \tag{4.27}$$

$$L(v) = \int_\Omega hv \, dx + \int_{\Gamma_N} gv \, ds. \tag{4.28}$$

Associated with the bilinear form is the energy norm defined by $\|v\|_E = (a(v,v))^{\frac{1}{2}}$.

Remark 11. As well known, the existence and uniqueness of the variational solution is given by the lax-Milgram theorem as the bilinear form $a(\,\cdot\,,\,\cdot\,)$ satisfied

$$\begin{aligned} |a(v,w)| &\leq M\|v\|_V\,\|w\|_V, \quad \forall v,w \in V,\\ a(v,v) &\geq \alpha\|v\|_V^2 \quad \forall v \in V, \end{aligned}$$

where M and α are positive constants independent of v and w.

4.4.1 Finite element approximation

Let V_h be a finite subspace of V on the mesh $\mathcal{T}_h$, then the finite element approximation requires finding a function $u_h \in V_h$

$$a(u_h, v_h) = L(v_h), \quad \forall v_h \in V_h. \tag{4.29}$$

The finite element approximation error denoted by $e = u - u_h$ achieves the error representation

$$\begin{aligned} a(e,v) &= a(u,v) - a(u_h,v) && (4.30)\\ &= L(v) - a(u_h,v) && (4.31)\\ &= \mathcal{R}(v), \quad \forall v \in V. && (4.32) \end{aligned}$$

It is the basis for a large class of error estimation using the energy norm, where $\mathcal{R}(v)$ is called residual functional or weak residual. Now, if we replace the test function v by v_h in (4.30), then we have fulfilled the Galerkin Orthogonality condition

$$a(e, v_h) = \mathcal{R}(v_h) = 0, \quad \forall v_h \in V_h. \tag{4.33}$$

Assuming that the bilinear form is positive, then it follows

$$\|\mathcal{R}\|_{V'} = \sup_{v \in V(\Omega)} \frac{|\mathcal{R}(v)|}{\|v\|_E} = \sup_{v \in V(\Omega)} \frac{|a(e,v)|}{\|v\|_E} = \|e\|_E. \tag{4.34}$$

where $\|\mathcal{R}\|_{V'}$ is the residual norm in the dual space.

4.4.2 A posteriori Error Estimator

The weak formulation of the problem (4.24) is to find $u \in H_D^1$ such that

$$a(u,v) = \int\limits_{\Omega} \nabla u \cdot \nabla v = \int\limits_{\Omega} hv + \int\limits_{\Gamma_N} gv = L(v), \quad \forall v \in H_D^1. \tag{4.35}$$

where

$$H_D^1 = \Big\{ u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_D \Big\}$$

The abstract form corresponding to the Galerkin method to the problem (4.24) is to find $u_h \in V_h$ such that

$$a(u_h, v) = L(v), \quad \forall v \in V_h. \tag{4.36}$$

We want to find the error formula $e = u - u_h$ with the help of (4.35) and (4.36), it get

$$\begin{aligned} a(e,v) &= a(u - u_h, v) \\ &= a(u,v) - a(u_h, v) \\ &= L(v) - a(u_h, v), \quad \forall v \in V. \end{aligned}$$

This is equivalent to

$$\int\limits_{\Omega} \nabla(u - u_h) \cdot \nabla v \, dx = \int\limits_{\Omega} hv \, dx + \int\limits_{\Gamma_N} gv \, ds - \int\limits_{\Omega} \nabla u_h \cdot \nabla v \, dx, \quad \forall v \in V.$$

Now,

$$\begin{aligned}
a(e,v) &= L(v) - a(u_h,v) && (4.37)\\
&= \int\limits_{\Omega} hv\,dx + \int\limits_{\Gamma_N} gv\,ds - \int\limits_{\Omega} \nabla u_h\cdot\nabla v\,dx && (4.38)\\
&= \sum_{k\in\mathcal{T}_h}\Big\{\int\limits_{k} hv\,dx + \int\limits_{\partial k\bigcap\Gamma_N} gv\,ds - \int\limits_{k}\nabla u_h\cdot\nabla v\,dx\Big\} && (4.39)\\
&= \sum_{k\in\mathcal{T}_h}\Big\{\int\limits_{\Omega} hv\,dx + \int\limits_{\partial k\bigcap\Gamma_N} gv\,ds\Big\} + \sum_{k\in\mathcal{T}_h}\Big\{\int\limits_{k}\Delta u_h v\,dx - \int\limits_{\partial k\backslash\Gamma_N}\frac{\partial u_h}{\partial n_k}v\,ds\Big\} && (4.40)\\
&= \sum_{k\in\mathcal{T}_h}\left\{\int\limits_{k}\big(h+\Delta u_h\big)v\,dx + \int\limits_{\partial k\bigcap\Gamma_N}\big(g-\frac{\partial u_h}{\partial n_k}\big)v\,ds - \int\limits_{\partial k\backslash\Gamma_N}\frac{\partial u_h}{\partial n_k}v\,ds\right\}. && (4.41)
\end{aligned}$$

$\forall v\in V$, integrating by parts over each element

$$a(e,v) = \sum_{k\in\mathcal{T}_h}\left\{\int\limits_{k} rv\,dx + \int\limits_{\partial k\bigcap\Gamma_N} Rv\,ds - \int\limits_{\partial k\backslash\Gamma_N}\frac{\partial u_h}{\partial n_k}v\,ds\right\}. \qquad (4.42)$$

where

(i) The residual interior is r and his equals

$$r = h + \Delta u_h \quad \text{in} \quad k\in\mathcal{T}_h.$$

(ii) The residual boundary is R and his equals

$$R = g - \frac{\partial u_h}{\partial n_k} \quad \text{on} \quad \partial k\bigcap\Gamma_N. \qquad (4.43)$$

(iii) The normal outward unite vector for ∂k is n_k.

Next, the contribution from the final term in equation (4.42) can be rewritten by observing the function v continuous along an edge

$$a(e,v) = \sum_{k\in\mathcal{T}_h}\left\{\int\limits_{k} rv\,dx + \int\limits_{\partial k\bigcap\Gamma_N} Rv\,ds\right\} - \underbrace{\sum_{\gamma\in\partial\mathcal{T}_h\backslash\partial\Omega}\int\limits_{\gamma}\frac{\partial u_h}{\partial n}v\,ds}_{*}\,. \qquad (4.44)$$

the sum $(*)$ is over all the inter-edge γ on the interior of the mesh and the quantity

$$\frac{\partial u_h}{\partial n} = \frac{\partial u_h}{\partial n}\bigg|_{k_1} - \frac{\partial u_h}{\partial n}\bigg|_{k_2}, \quad \text{for } k_1 \bigcap k_2 \in \partial\mathcal{T}_h.$$

is defined on the node γ, which separate the element k_1,k_2 represents a jump discontinuity in the approximation to the flux.Identity (4.44) can be written more compactly by extending the definition of the boundary residual to include the jump discontinuity in the flux, on interior the definition (4.43) is augmented by $R = -\frac{1}{2}\left[\frac{\partial u_h}{\partial n}\right]$ so that equation (4.44) becomes

$$a(e,v) = \sum_{k\in\mathcal{T}_h} \int_k rv\,dx + \underbrace{\sum_{\gamma\in\partial\mathcal{T}_h} \int_\gamma Rv\,ds}_{**}, \quad \forall v \in V. \tag{4.45}$$

where the sum ($**$)extended the entire node of the partition $\mathcal{T}_h$, [Gaeta and Rodríguez (2017)] and [Wait (1631)]. Using the property (4.33), for $v \in V$, let $\pi_h v$ be an interpolation to v from V_h, put $v = \pi_h v$ in identity (4.45) there hold

$$0 = a(e,v_h) = a(e,v) \qquad = \sum_{k\in\mathcal{T}_h} \int_k r\pi_h v\,dx + \sum_{\gamma\in\partial\mathcal{T}_h} \int_\gamma R\pi_h v\,ds. \tag{4.46}$$

substract equation (4.46) from equation (4.45) to get

$$a(e,v) = \sum_{k\in\mathcal{T}_h} \int_k r(v-\pi_h v)\,dx + \sum_{\gamma\in\partial\mathcal{T}_h} \int_\gamma R(v-\pi_h v)\,ds, \quad \forall v \in V. \tag{4.47}$$

The identity (4.47) plays an significant role, either indirectly or directly, in a posteriori error analysis, using the Cauchy-Schwartz inequality that we have

$$a(e,v) \le \sum_{k\in\mathcal{T}_h} \|r\|_{L^2(k)}\,\|v-\pi_h v\|_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} \|R\|_{L^2(\gamma)}\,\|v-\pi_h v\|_{L^2(\gamma)}. \tag{4.48}$$

Now, using theorem (4.3.1) in equation (4.48)

$$\begin{aligned} a(e,v) &\le \sum_{k\in\mathcal{T}_h} Ch_k\|r\|_{L^2(k)}\,|v|_{H^1(\tilde{k})} + \sum_{\gamma\in\partial\mathcal{T}_h} Ch_k^{\frac{1}{2}}\|R\|_{L^2(\gamma)}\,|v|_{H^1(\tilde{k})} \\ &\le C\,|v|_{H^1(\tilde{k})}\left\{\sum_{k\in\mathcal{T}_h} h_k\|r\|_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h_k^{\frac{1}{2}}\|R\|_{L^2(\gamma)}\right\}, \end{aligned}$$

applying the Cauchy-Schwartz inequality

$$a(e,v) \leq C\,|v|_{H^1(\tilde{k})}\left\{\sum_{k\in\mathcal{T}_h} h_k^2\|r\|_{L^2(k)}^2 + \sum_{\gamma\in\partial\mathcal{T}_h} h_k\|R\|_{L^2(\gamma)}^2\right\}^{\frac{1}{2}}.$$

using the coircivity of the bilinear form over the global space V it follows $\|v\|_{H^1(\tilde{k})} \leq C\|v\|$, then replaces v with e which becomes

$$a(e,e) \leq C\,\|e\|\left\{\sum_{k\in\mathcal{T}_h} h_k^2\|r\|_{L^2(k)}^2 + \sum_{\gamma\in\partial\mathcal{T}_h} h_k\|R\|_{L^2(\gamma)}^2\right\}^{\frac{1}{2}} \tag{4.49}$$

$$\|e\|^2 \leq C\,\|e\|\left\{\sum_{k\in\mathcal{T}_h} h_k^2\|r\|_{L^2(k)}^2 + \sum_{\gamma\in\partial\mathcal{T}_h} h_k\|R\|_{L^2(\gamma)}^2\right\}^{\frac{1}{2}}, \tag{4.50}$$

divided both sides by $\|e\|$ the squaring to get **A posterior error equation**

$$\|e\|^2 \leq C\left\{\sum_{k\in\mathcal{T}_h} h_k^2\|r\|_{L^2(k)}^2 + \sum_{\gamma\in\partial\mathcal{T}_h} h_k\|R\|_{L^2(\gamma)}^2\right\}.$$

Both the consistency of the R.H.S can be easily determined from the data and the approximation of the finite element, so we can write the terms as a single sum.

$$\|e\|^2 \leq C\sum_{k\in\mathcal{T}_h}\left\{h_k^2\|r\|_{L^2(k)}^2 + \frac{1}{2}h_k\|R\|_{L^2(\partial(k))}^2\right\}.$$

Here, $\partial\mathcal{T}_h$ shows the set of the inner edges (the edges do not lie on the boundary),and **local error** estimators can now be defined as follows

$$\eta^2 = h_k^2\|r\|_{L^2(k)}^2 + \frac{1}{2}h_k\|R\|_{L^2(\partial(k))}^2$$

On which we can construct a **global error** estimator, [Ainsworth and Oden (2000)].

$$\eta = \left(\sum_{k\in\mathcal{T}}\eta_k^2\right)^{\frac{1}{2}} = \left\{\sum_{k\in\mathcal{T}_h} h_k^2\|r\|_{L^2(k)}^2 + \sum_{\gamma\in\partial\mathcal{T}_h} h\gamma\|R\|_{L^2(\gamma}^2\right\}^{\frac{1}{2}}.$$

Notation 8.

$\mathcal{T}_h$:= quasi-uniform triangulation,

Γ_N:= boundary, defined under the terms of Neumann,

$\mathcal{T}_{\Gamma_N}$:= inner edges with Neumann conditions,

$\mathcal{T}_\Omega$:= inner edges with Dirichlet conditions,

γ := the common edge of two inner triangles,

n_k := normal outward unit for element $k \in \mathcal{T}_h$,

n_γ := outward normal vector, perpendicular to each $\gamma \in \partial\mathcal{T}_h$,

$\partial\mathcal{T}_h$:= set of all edges within Ω,

$\tilde{k}$:= subdomain of elements that share a common nodal with k.

4.5 Error Estimator for Reaction-Diffusion Problem

Suppose $\Omega \subset \mathbb{R}^2$ is a domain bounded by Lipschitz boundaries Γ. Consider the elliptic boundary problem of the model, [Ainsworth and Oden (1997)].

$$\begin{cases} -\Delta u + c\,u = h, & \text{in } \Omega, \\ u = 0, & \text{on } \Gamma_D, \\ \dfrac{\partial u}{\partial n} = g, & \text{on } \Gamma_N, \end{cases} \tag{4.51}$$

where $h \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$, $c \geq 0$ and the boundary Γ_D , Γ_N thought to be $\bar{\Gamma}_D \bigcup \bar{\Gamma}_N = \Gamma$, $\Gamma_D \bigcap \Gamma_N = \phi$. The outward natural vector of Γ is denoted as n, where $n \in \Big[L^\infty(\Gamma)\Big]^n$.

Notation 9. In mathematics, a Lipschitz domain (or domain with Lipschitz boundary) is a domain in Euclidean space whose boundary is "sufficiently regular" in the sense that it can be thought of as locally being the graph of a Lipschitz continuous function.

- The variational form of this problem is

$$\text{find } u \in V \quad \text{like that} \quad a(u,v) = L(v), \quad \forall v \in V.$$

- V is the space define as

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

- and

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

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

Suppose $V_h \subset V$ is a finite element subspace, then the finite element approximation is to find $u_h \in V_h$ such that

(i) The bilinear and linear forms are

$$a(u_h, v_h) = L(v_h), \quad \forall v_h \in V_h.$$

(ii) The error $e = u - u_h$ belongs to the space V and is satisfied

$$a(e, v) = L(v) - a(u_h, v), \quad \forall v \in V.$$

(iii) The normal orthogonal state of the Galerkin projection error has been preserved

$$a(e, v_h) = 0, \quad \forall v_h \in V_h.$$

4.5.1 A posteriori Error Estimators

In the formula (4.51), assuming that the finite element approximation of u_h is computed, [Deka and Ahmed (2011)]. The base problem in the posteriori error estimation is the question:
(How to estimate the estimation error e **?)**
To provide an answer, one can make use of

1. Galerkin is approximating u_h itself.
2. Data h, g.
3. Equation characterizing the true error.

$$a(e, v) = L(v) - a(u_h, v), \quad \forall v \in V. \tag{4.52}$$

4. Property of Galerkin orthogonality.

$$a(e, v_h) = 0, \quad \forall v_h \in V_h. \tag{4.53}$$

The following section explain how these may be used to derive, [Courant (1943)] and [Ern and Guermond (2004)].

4.5.1.1 A simple a posteriori error estimation

The first step is to break down the equation (4.52) from each element to the local contribution.

$$
\begin{aligned}
a(e,v) &= L(v) - a(u_h, v) && (4.54)\\
&= \sum_{k\in\mathcal{T}_h}\left\{\int\limits_k hv\,dx + \int\limits_{\partial k\bigcap\Gamma_N} gv\,dx - \int\limits_k \Big(\nabla u_h\cdot\nabla v + cu_h v\Big)\,dx\right\}, && (4.55)\\
&= \sum_{k\in\mathcal{T}_h}\left\{\int\limits_k \Big(h - cu_h\Big)v\,dx + \int\limits_{\partial k\bigcap\Gamma_N} gv\,dx - \int\limits_k \nabla u_h\cdot\nabla v\,dx\right\}, && (4.56)\\
&= \sum_{k\in\mathcal{T}_h}\left\{\int\limits_k \Big(h + \Delta u_h - cu_h\Big)v\,dx + \int\limits_{\partial k\bigcap\Gamma_N} \Big(g - \frac{\partial u_h}{\partial n_k}\Big)v\,ds - \int\limits_{\partial k\backslash\Gamma_N} \frac{\partial u_h}{\partial n_k}\,v\,dx\right\}. && (4.57)
\end{aligned}
$$

$\forall v \in V$, integrating by parts over each element

$$
a(e,v) = \sum_{k\in\mathcal{T}_h}\left\{\int\limits_k rv\,dx + \int\limits_{\partial k\bigcap\Gamma_N} Rv\,ds - \int\limits_{\partial k\backslash\Gamma_N} \frac{\partial u_h}{\partial n_k}\,v\,ds\right\}. \tag{4.58}
$$

where

1. The residual interior is r and his equals

$$
r = h + \Delta u_h - cu_h \quad \text{in} \quad k.
$$

2. The residual boundary is R and his equals

$$
R = g - \frac{\partial u_h}{\partial n_k} \quad \text{on} \quad \partial k\bigcap\Gamma_N. \tag{4.59}
$$

3. The normal outward unite vector for ∂k is n_k.

These quantities are well defined due to the smoothness of the data and the nearly regularity of the u_h approximation. The contribution from the final term in equation (4.58) can be rewritten by observing that the (trace of) the function v continuous along an edge shared by two the elements giving

$$
a(e,v) = \sum_{k\in\mathcal{T}_h}\left\{\int\limits_k rv\,dx + \int\limits_{\partial k\bigcap\Gamma_N} Rv\,ds\right\} - \underbrace{\sum_{\gamma\in\partial\mathcal{T}_h\backslash\partial\Omega}\int\limits_\gamma \Big[\frac{\partial u_h}{\partial n}\Big]v\,ds}_{(*)} \tag{4.60}
$$

where the sum of $(*)$ is over all the inter edge γ on the interior of the mesh, the quantity

$$\Big[\frac{\partial u_h}{\partial n}\Big] = n_k \cdot (\nabla u_h)_k + n_{k'} \cdot (\nabla u_h)_{k'},$$

it is defined on the node of γ, which separates the element k, k' represents a jump discontinuity in the approximation to the flux. Identity (4.60) can be written more compactly by extending the definition of the boundary residual to include the jump discontinuity in the flux, so that on interior edge the definition (4.59) is augmented by $R = -\frac{1}{2}\Big[\frac{\partial u_h}{\partial n}\Big]$ so that equation (4.60) becomes

$$a(e, v) = \sum_{k \in \mathcal{T}_h} \int_k rv \, dx + \underbrace{\sum_{\gamma \in \partial \mathcal{T}_h} \int_\gamma Rv \, ds}_{(**)}, \quad \forall v \in V. \tag{4.61}$$

where the sum $(**)$ is over all edge in the partition $\mathcal{T}_h$, the property (4.53) can be used as follows, for $v \in V$, let $\pi_h v$ be an interpolation to v from V_h, using (4.53) and the identity (4.61) there hold

$$0 = \sum_{k \in \mathcal{T}_h} \int_k r\pi_h v \, dx + \sum_{\gamma \in \partial \mathcal{T}_h} \int_\gamma R\pi_h v \, ds, \tag{4.62}$$

subtract equation (4.62) from equation (4.61) to get

$$a(u, v) = \sum_{k \in \mathcal{T}_h} \int_k r(v - \pi_h v) \, dx + \sum_{\gamma \in \partial \mathcal{T}_h} \int_\gamma R(v - \pi_h v) \, ds, \quad \forall v \in V. \tag{4.63}$$

The identity (4.63) plays an significant role, either indirectly or directly, in a posteriori error analysis, using the Cauchy-Schwartz inequality that we have

$$a(e, v) \leq \sum_{k \in \mathcal{T}_h} \|r\|_{L^2(k)} \, \|v - \pi_h v\|_{L^2(k)} + \sum_{\gamma \in \partial \mathcal{T}_h} \|R\|_{L^2(\gamma)} \, \|v - \pi_h v\|_{L^2(\gamma)}. \tag{4.64}$$

Now, using theorem (4.3.1) in equation (4.64)

$$\begin{aligned} a(e, v) &\leq \sum_{k \in \mathcal{T}_h} Ch_k \|r\|_{L^2(k)} \, |v|_{H^1(\tilde{k})} + \sum_{\gamma \in \partial \mathcal{T}_h} Ch_k^{\frac{1}{2}} \|R\|_{L^2(\gamma)} \, |v|_{H^1(\tilde{k})} \\ &\leq C \, |v|_{H^1(\tilde{k})} \Bigg\{ \sum_{k \in \mathcal{T}_h} h_k \|r\|_{L^2(k)} + \sum_{\gamma \in \partial \mathcal{T}_h} h_k^{\frac{1}{2}} \|R\|_{L^2(\gamma)} \Bigg\}, \end{aligned}$$

applying the Cauchy-Schwartz inequality

$$a(e, v) \leq C \, |v|_{H^1(\tilde{k})} \Bigg\{ \sum_{k \in \mathcal{T}_h} h_k^2 \|r\|^2_{L^2(k)} + \sum_{\gamma \in \partial \mathcal{T}_h} h_k \|R\|^2_{L^2(\gamma)} \Bigg\}^{\frac{1}{2}}.$$

using the coircivity of the bilinear form over the global space V it follows $\|v\|_{H^1(\tilde{k})} \leq C\|v\|$, then replaces v with e which becomes

$$a(e,e) \leq C\,\|e\| \left\{ \sum_{k\in\mathcal{T}_h} h_k^2\|r\|^2_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h_k\|R\|^2_{L^2(\gamma)} \right\}^{\frac{1}{2}} \tag{4.65}$$

$$\|e\|^2 \leq C\,\|e\| \left\{ \sum_{k\in\mathcal{T}_h} h_k^2\|r\|^2_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h_k\|R\|^2_{L^2(\gamma)} \right\}^{\frac{1}{2}}, \tag{4.66}$$

divided both sides by $\|e\|$ the squaring to get **A posterior error equation**

$$\|e\|^2 \leq C \left\{ \sum_{k\in\mathcal{T}_h} h_k^2\|r\|^2_{L^2(k)} + \sum_{\gamma\in\partial\mathcal{T}_h} h_k\|R\|^2_{L^2(\gamma)} \right\}.$$

Both the consistency of the R.H.S can be easily determined from the data and the approximation of the finite element, so we can write the terms as a single sum.

$$\|e\|^2 \leq C \sum_{k\in\mathcal{T}_h} \left\{ h_k^2\|r\|^2_{L^2(k)} + \frac{1}{2}h_k\|R\|^2_{L^2(\partial(k)} \right\}.$$

The perpose is doing so is that defining the local error indicator by η_k on element k by

$$\eta^2 = h_k^2\|r\|^2_{L^2(k)} + \frac{1}{2}h_k\|R\|^2_{L^2(\partial(k)}$$

Finally the last equation can be written as

$$\|e\|^2 \leq C \sum_{k\in\mathcal{T}_h} \eta_k^2.$$

It is assumed that each of these quantity is a measure of the local discretization error over each element. In this way one can use η_k as a basis for guiding local mesh refinements.

Efficiency of the estimator:

The estimator of the a posteriori implies by

$$\|e\|^2 \leq C \sum_{k\in\mathcal{T}_h} \eta_k^2 \tag{4.67}$$

Provided that the upper bound of the estimation error is reach the (generally unknown) constant C, if the estimator is to be used as a basis for adaptation of the purification algorithm and, in particular, the stopping criterion, it is desirable that the estimator is effective in the sense that there must be a constant C, Fixed, which does not depend on the mesh size.

$$\sum_{k\in\mathcal{T}_h} \eta_k^2 \leq C\ \|e\|^2.$$

This form of bound is of particle significance, as it confirms, in conjunction with the lower bound (4.67), that the rate of change of the estimator as the mesh size reduces the action of the actual error, [Ainsworth and Oden (1997)], [Bustinza et al. (2005)] and [Becker et al. (2003)].

4.5.2 Numerical solution

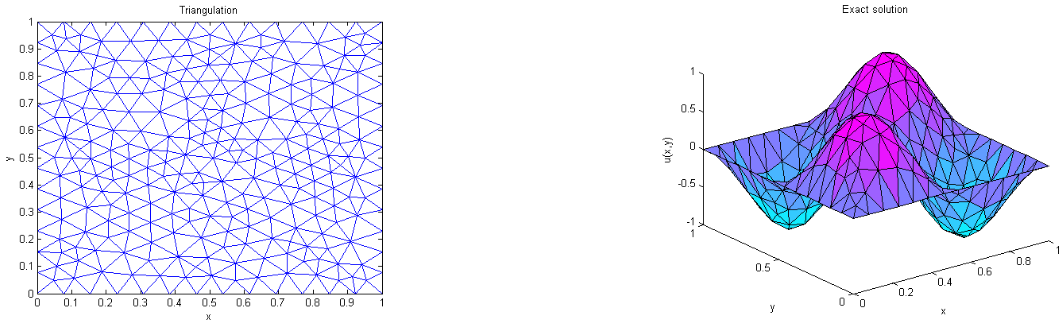
Example 1.

Consider the problem of the elliptical boundary value

$$\begin{cases} -\Delta u = 8\pi^2 \sin 2\pi x \sin 2\pi y, & x, y \in \Omega = [0,1]\times[0,1] \\ u = 0, & \text{on } \Gamma\ . \end{cases}$$

We denote to the maximum edge size by $h_{\max}$, and we consider the following two cases as computational examples.

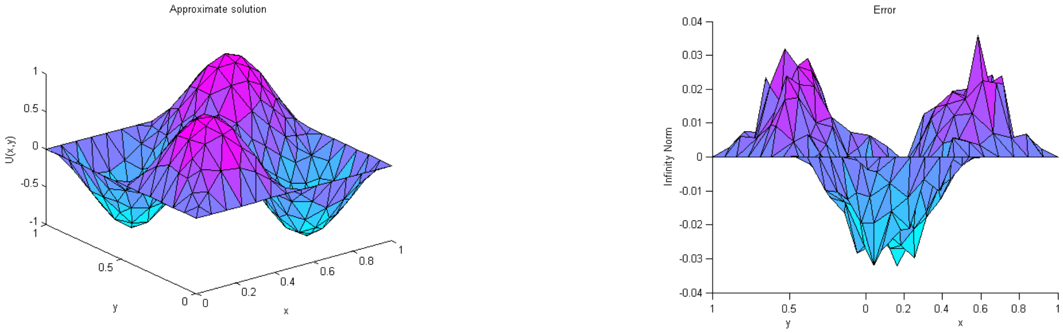
1. At $h_{\max} = 0.08$, the maximum norm error $= 0.035833$



(a) Triangulation

(b) Exact Solution

Figure 4.2 – Triangulation and the exact solution

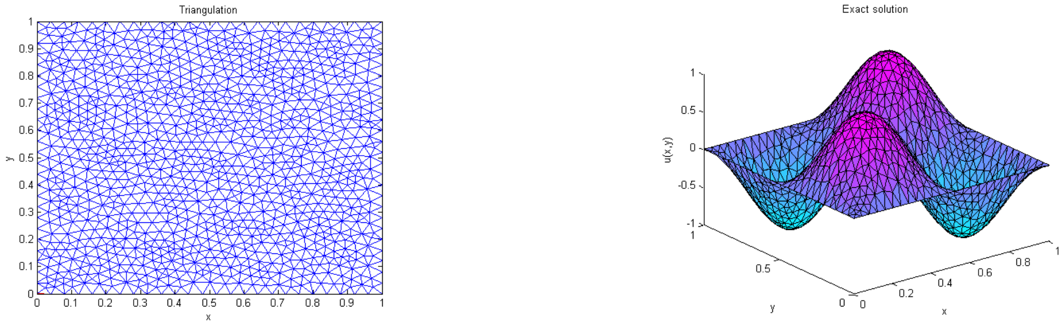


(a) Approximation

(b) Infinity norm error

Figure 4.3 – The approximation and the corresponding error

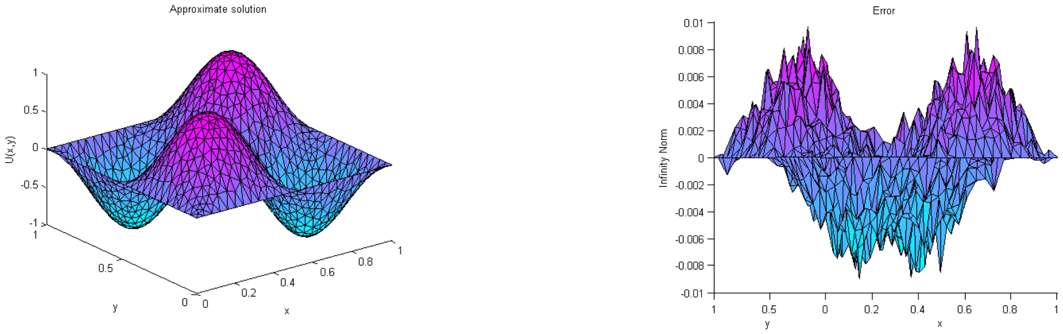
2. At $h_{\max} = 0.04,$ the maximum norm error $= 0.0096286$



(a) Triangulation

(b) Exact solution

Figure 4.4 – Refined triangulation and the exact solution



(a) Approximation

(b) Infinity norm error

Figure 4.5 – The approximation and the corresponding error

Figure 4.3a at $h_{\max} = 0.08$ with maximum error value $= 0.035833$, and figure 4.5a at $h_{\max} = 0.04$ with maximum error value $= 0.0096286$ show the finite element

approximate solution.
Clearly the approximate solution obtained with ($h_{\max} = 0.04$) nearly matches the exact solution and is closed to the exact solution than the solution obtained with ($h_{\max} = 0.08$). Therefore, the maximum norm error obtained by $h_{\max} = 0.04$ is less than the maximum norm error by $h_{\max} = 0.08$.

The following table shows that the maximum norm error values is decreasing as $h_{\max}$ becomes smaller. This means that refining the mesh provides better approximation.

$h_{\max}$	**maximum norm error**
0.01	0.00060786
0.02	0.0023331
0.03	0.0051082
0.04	0.0096286
0.05	0.015545
0.06	0.021533
0.07	0.025992
0.08	0.035833

Table 4.1 – comparing the maximum norm error values with different $h_{\max}$

CONCLUSION

⋆⋆⋆⋆⋆⋆⋆⋆⋆⋆⋆⋆

In this thesis we reviewed some basic and general theory of the finite element method. We also discussed the variational formulation and discretization of the method for one and two dimensional problems. After that, the error estimation in its both types, a posteriori and a priori, is explained.

The main goal of this thesis is to find a posteriori error estimations for reaction-diffusion problem and Poisson equation with homogeneous and mixed, Dirichlet/Neumann, boundary conditions. At the end, numerical examples are provided where the approximation is obtained using Matlab implementation,

APPENDIX

⋆⋆⋆*MATLAB CODE*⋆⋆⋆

Matlab code for Example (2)

```
% The program solves the heat eq.  k*w''+2w=0,
% w(0)=1 and w(1)=-1
% k=-1

 clear all
 clc
 format long
 k=-1;

 x_Int=[0 1];


 disp('"n" IS THE NUMBER OF SUBINTERVALS.');
 n=input('n=');
 a=x_Int(1);
 b=x_Int(2);
 h=(b-a)/n;
 xx=a:h:b;
 h=h*ones(size(xx));

% Computing M000
 M000=zeros(n-1,n-1);
 for i=1:n-1
       M000(i,i)=4;
 end
 for i=1:n-2
```

```
    M000(i,i+1)=1;
    M000(i+1,i)=1;
end
M000=(h(1)/6)*M000;

% computing M110
M110=zeros(n-1,n-1);
for i=1:n-1
    M110(i,i)=2;
end
for i=1:n-2
    M110(i,i+1)=-1;
    M110(i+1,i)=-1;
end
M110=(1/h(1))*M110;

% Computing the loud vector bb00
bb00=zeros(n-1,1);
bb00(1,1)=1*(1/h(1)-h(1)/3);
bb00(n-1,1)=-1*(1/h(1)-h(1)/3);


xi=(M110+2*M000)\bb00;
plot(xx,[1 xi' -1],'*r')
hold on
fplot('-0.321207702025859*exp(sqrt(2)*x)+1.321207702025859*exp(-sqrt(2)*x)',[0,1])
grid
legend('FEM approx', 'exact')
% Error
% exact=-0.321207702025859*exp(sqrt(2)*xx)+1.321207702025859*exp(-sqrt(2)*xx);
C2=(-(1+exp(sqrt(2)))/(exp(-sqrt(2))-exp(sqrt(2))));
C1=(1-(-(1+exp(sqrt(2)))/(exp(-sqrt(2))-exp(sqrt(2)))));
exact=C1*exp(sqrt(2)*xx)+C2*exp(-sqrt(2)*xx);
approx=[1 xi' -1];
disp('              Nodes                       Exact                   Approximation                Absolute Error')
disp('              =====                       ====                    ===========                  ===========')
disp([xx' exact' approx' abs(exact-approx)'])
```

Example (3) code + error

```
 function poi2D( )
% -div(grad u) = f, in [0,1]x[0,1],
% u = 0, on boundary
% f(x,y)=8*pi^2*sin(2*pi*x)*sin(2*pi*y) ; the source function
% u(x,y)=sin(2*pi*x)*sin(2*pi*y); the exact solution

clear all, clc

% triangulation
g = [2 0 1 0 0 1 0;2 1 1 0 1 1 0;2 1 0 1 1 1 0;2 0 0 1 0 1 0]' ;
[p,e,t] = initmesh(g,'hmax',0.04) ;% Try 0.07 , 0.06 , 0.05 , 0.04 , 0.03 , 0.02, 0.01 to refine the mesh

figure(1); clf
pdemesh(p,e,t)
title( 'Triangulation' )
xlabel('x'), ylabel('y')

% legend('FEM approx', 'exact')
% assemble
[A,b] = assemble(p,e,t,'f');
% solve
U = A\b; %U is the approximation
% visualize

figure(2); clf
pdesurf(p,t,U) % visualizing the approximation
shading faceted
title( 'Approximate solution' )
xlabel('x'), ylabel('y'), zlabel('U(x,y)')

% Exact Solution
u=zeros(size(U));
for i=1:size(p,2)
    x=p(:,i);
    u(i)=sin(2*pi*x(1)) * sin(2*pi*x(2));
end


```

```
figure(3),clf
pdesurf(p,t,u) % visualizing the exact solution
shading faceted
title( 'Exact solution' )
xlabel('x'), ylabel('y'), zlabel('u(x,y)')

% Compute the error
error = U - u;
enorm = max(abs(error))
disp(['Maximum norm error: ' num2str(enorm)])

figure(4); clf
pdesurf(p,t,error) % visualizing the error
shading faceted
title('Error')
xlabel('x'), ylabel('y'), zlabel('Infinity Norm')

% subroutines

 function z = f(x,y)
z = 8*pi^2*sin(2*pi*x).*sin(2*pi*y) ; % the source function
% z = 2*pi*sin(pi*x).*sin(pi*y) ;

function [A,b] = assemble( p,e ,t ,f)
Nt = size(t,2);
Np = size(p,2);
Ne = size(e,2);
A = sparse(Np,Np);
b = zeros(Np,1);
for i = 1:Nt
n = t(1:3,i);
x = p(1,n);
y = p(2,n);
dx = [y(2)-y(3); y(3)-y(1); y(1)-y(2)];
dy = [x(3)-x(2); x(1)-x(3); x(2)-x(1)];
area = 0.5*abs(x(2)*y(3)-y(2)*x(3)-x(1)*y(3)+y(1)*x(3)+x(1)*y(2)-y(1)*x(2));
A(n,n) = A(n,n) + (dx*dx'+dy*dy')/4/area ;
b(n) = b(n) + area/12*[2 1 1 ; 1 2 1 ; 1 1 2]*feval('f',x,y)' ;
end
```

```
% BC
for i = 1:Ne
n = e(1,i) ;
A(n,n) = 1e6 ;
b(n) = 0 ;
end
```

Bibliography

Ainsworth, M. and Oden, J. (1997). A posteriori error estimation in finite element analysis. 142(1):1–88.

Ainsworth, M. and Oden, J. T. (2000). A posterori error estimation in finite element analysis.

Bathe, K.-J. (2014). Finite Element Procedures.

Becker, R., Hansbo, P., and Larson, M. G. (2003). Energy norm a posteriori error estimation for discontinuous galerkin methods. 192(5):723–733.

Braess, D. (2007a). Finite elements. Theory, fast solvers and applications in elasticity theory. (Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie.) 4th revised and extended ed.

Braess, D. (2007b). Finite elements: Theory, fast solvers, and applications in solid mechanics.

Braess, D. (2013). Finite elemente: Theorie, schnelle löser und anwendungen in der elastizitätstheorie.

Braess, D. and Verfürth, R. (1996). A posteriori error estimators for the raviart–thomas element. 33(6):2431–2444.

Brenner, S. and Scott, R. (2008). The Mathematical Theory of Finite Element Methods.

Brezis, H. and Turner, R. E. L. (1977). On a class of superlinear elliptic problems. 2(6):601–614.

Burden, R. L., Faires, J. D., and Burden, A. M. (2015). Numerical Analysis.

Bustinza, R., Gatica, G. N., and Cockburn, B. (2005). An a posteriori error estimate for the local discontinuous galerkin method applied to linear and nonlinear diffusion problems. 22(1):147–185.

Cao, Y., Nie, S., and Wu, Z. (2019). Numerical simulation of parachute inflation: A methodological review. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 233(2):736–766.

Ciarlet, P. G. (2002). Finite element method for elliptic problems.

Courant, R. (1943). Variational methods for the solution of problems of equilibrium and vibrations. *Bulletin of the American Mathematical Society*, 49(1):1–23.

Deka, B. and Ahmed, T. (2011). Finite element methods for semilinear elliptic problems with smooth interfaces. *Indian Journal of Pure and Applied Mathematics*, 42(4):205.

Eriksson, K. (1996). Computational differential equations.

Ern, A. and Guermond, J.-L. (2004). Theory and practice of finite elements.

Evans, L. C. (2010). Partial Differential Equations: Second Edition.

Gaeta, G. and Rodríguez, M. A. (2017). Lectures on Hyperhamiltonian Dynamics and Physical Applications.

Gander, M. J. and Kwok, F. (2018). Numerical analysis of partial differential equations using maple and MATLAB. Google-Books-ID: UbFqDwAAQBAJ.

GIDAS, B. (1981). A priori bounds for positive solutions of nonlinear elliptic equations. 6:883–901.

Grätsch, T. and Bathe, K.-J. (2005). A posteriori error estimation techniques in practical finite element analysis. 83(4):235–265.

Gustafson, K. E. (2012). Introduction to partial differential equations and hilbert space methods. Google-Books-ID: TRi0AAAAQBAJ.

Hackbusch, W. (2017). Elliptic differential equations: Theory and numerical treatment.

Hicks, M. A., Brinkgreve, R. B. J., and Rohe, A. (2014). Numerical methods in geotechnical engineering. Google-Books-ID: 31DvBQAAQBAJ.

Houston, P. and Süli, E. (2001). Stabilized hp-finite element approximation of partial differential equations with nonnegative characteristic form. 66:99–119.

Izadi, M. (2007). Streamline diffusion method for treating coupling equations of hyperbolic scalar conservation laws. *Mathematical and Computer Modelling*, 45(1):201–214.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An introduction to statistical learning: with applications in r.

John, V. (2016). Finite element methods for incompressible flow problems.

Kawecki, E., Lakkis, O., and Pryer, T. (2018). A finite element method for the monge–amp\‘ere equation with transport boundary conditions.

Kelemen, S. and Quittner, P. (2010). Boundedness and a priori estimates of solutions to elliptic systems with dirichlet-neumann boundary conditions. 9(3):731.

Kirby, R. C. and Logg, A. (2012). The finite element method. In Logg, A., Mardal, K.-A., and Wells, G., editors, *Automated Solution of Differential Equations by the Finite Element Method: The FEniCS Book*, Lecture Notes in Computational Science and Engineering, pages 77–94. Springer.

Kunert, G. (2001). Robust a posteriori error estimation for a singularly perturbed reaction–diffusion equation on anisotropic tetrahedral meshes. 15(1):237–259.

Kuo, H.-J. and Trudinger, N. S. (1992). Discrete methods for fully nonlinear elliptic equations. 29(1):123–135.

Langtangen, H. P. and Mardal, K.-A. (2019). Introduction to Numerical Methods for Variational Problems.

Larson, M. G. and Bengzon, F. (2013). The Finite Element Method: Theory, Implementation, and Applications.

Larsson, S. and Thomee, V. (2003). Partial differential equations with numerical methods.

Le Dret, H. and Lucquin, B. (2016). Variational approximation methods for elliptic pdes. pages 145–166.

Li, Z. (2017). Numerical Solution of Differential Equations: Introduction to Finite Difference and Finite Element Methods.

Nussbaum, R. (1975). Positive solutions of nonlinear elliptic boundary value problems. 51(2):461–482.

Quarteroni, A. (2014). Numerical models for differential problems.

Quarteroni, A. M. and Valli, A. (2008). Numerical Approximation of Partial Differential Equations.

Quittner, P. and Souplet, P. (2004). A priori estimates and existence for elliptic systems via bootstrap in weighted lebesgue spaces. 174(1):49–81.

Rauch, J. (1997). Partial Differential Equations.

Renardy, M. and Rogers, R. C. (2004). An introduction to partial differential equations.

Saad, Y. (2003). Iterative Methods for Sparse Linear Systems.

Schopf, M. (2014). Error analysis of the galerkin FEM in l 2 -based norms for problems with layers.

Strauss, W. A. (2007). Partial differential equations: An introduction. Google-Books-ID: m2hvDwAAQBAJ.

Sun, J. and Zhou, A. (2016). Finite element methods for eigenvalue problems. Google-Books-ID: YC7FDAAAQBAJ.

Thomas, A., Ulrich, L., Arnd, M., and Olaf, S. (2019). Advanced finite element methods with applications: Selected papers from the 30th chemnitz finite element symposium 2017.

Thomas, J. W. (1998). Numerical partial differential equations: Finite difference methods.

Wait, A. R. M. (1631). Finite Element Method in Partial Differential Equations by A. Richard Mitchell.

Wu, J. (1996). Theory and applications of partial functional differential equations.

Yu, D. and Zhao, L. (2005). Boundary integral equations and a posteriori error estimates. 10(1):35–42.

Zeidler, E. (2007). Quantum Field Theory I: Basics in Mathematics and Physics: A Bridge between Mathematicians and Physicists. Google-Books-ID: XYt-nGl9enNgC.

Zhang, Z. and Yan, N. (2001). Recovery type a posteriori error estimates in finite element methods. 8(2):235.

Øksendal, B. (2003). Stochastic differential equations: An introduction with applications.