Chapter

A Brief Summary of the Finite Element Method for Differential Equations

Mahboub Baccouch

Abstract

The finite element (FE) method is a numerical technique for computing approximate solutions to complex mathematical problems described by differential equations. The method was developed in the 1950s to solve complicated problems in engineering, notably in elasticity and structural mechanics modeling involving elliptic partial differential equations and complicated geometries. But nowadays the range of applications is quite extensive. In particular, the FE method has been successfully applied to many problems such as fluid–structure interaction, thermomechanical, thermochemical, thermo-chemo-mechanical problems, biomechanics, biomedical engineering, piezoelectric, ferroelectric, electromagnetics, and many others. This chapter contains a summary of the FE method. Since the remaining chapters of this textbook are based on the FE method, we present it in this chapter as a method for approximating solutions of ordinary differential equations (ODEs) and partial differential equations (PDEs).

Keywords: the finite element method, initial-value problems, boundary-value problems, Laplace equation, heat equation, wave equation

1. Introduction

1.1 An overview of the finite element method

Differential equations arise in many disciplines such as engineering, mathematics, sciences, economics, and many other fields. Unfortunately solutions to differential equations can rarely be expressed by closed formulas and numerical methods are needed to approximate their solutions. There are many numerical methods for approximating the solution to differential equations including the finite difference (FD), finite element (FE), finite volume (FV), spectral, and discontinuous Galerkin (DG) methods. These methods are used when the mathematical equations are too complicated to be solved analytically.

The FE method has become the standard numerical scheme for approximating the solution to many mathematical problems; see [1–9] and the references therein just to mention a few. In simple words, the FE method is a numerical method to solve differential equations by discretizing the domain into a finite mesh. Numerically speaking, a set of differential equations are converted into a set of algebraic equations to be solved for unknown at the nodes of the mesh. The FE method originated from the need to solve complex elasticity and structural analysis problems in civil and aeronautical engineering. The first development can be traced back

to the work by Hrennikoff in 1941 [10] and Courant in 1943 [11]. Although these pioneers used different perspectives in their FE approaches, they each identified the one common and essential characteristic: mesh discretization of a continuous domain into a set of discrete sub-domains, usually called elements. Another fundamental mathematical contribution to the FE method is represented by Gilbert Strang and George Fix [12]. Since then, the FE method has been generalized for the numerical modeling of physical systems in many engineering disciplines including electromagnetism, heat transfer, and fluid dynamics.

The advantages of this method can be summarized as follows:

1. Numerical efficiency: The discretization of the calculation domain with finite elements yields matrices that are in most cases sparse and symmetric. Therefore, the system matrix, which is obtained after spatial and time discretization, is sparse and symmetric too. Both the storage of the system matrix and the solution of the algebraic system of equations can be performed in a very efficient way.

2. Treatment of nonlinearities: The modeling of nonlinear material behavior is well established for the FE method (e.g., nonlinear curves, hysteresis).

3. Complex geometry: By the use of the FE method, any complex domain can be discretized by triangular elements in 2D and by tetrahedra elements in 3D.

4. Applicable to many field problems: The FE method is suited for structural analysis, heat transfer, electrical/magnetical analysis, fluid and acoustic analysis, multi-physics, etc.

COMSOL Multiphysics (known as FEMLAB before 2005) is a commercial FE software package designed to address a wide range of physical phenomena. It is widely used in science and industry for research and development. It excels at modeling almost any multi-physics problem by solving the governing set of PDEs via the FE method. This software package is able to solve one, two and three-dimensional problems. It comes with a modern graphical user interface to set up simulation models and can be scripted from Matlab or via its native Java API.

In this chapter, we introduce the FE method for several one-dimensional and two-dimensional model problems. Although the FE method has been extensively used in the field of structural mechanics, it has been successfully applied to solve several other types of engineering problems, such as heat conduction, fluid dynamics, seepage flow, and electric and magnetic fields. These applications prompted mathematicians to use this technique for the solution of complicated problems. For illustration, we will use simple one-dimensional and two-dimensional model problems to introduce the FE method.

2. The FE method for ODEs

2.1 The FE method for first-order linear IVPs

We first present the FE method as an approximation technique for solving the following first-order initial-value problem (IVP) using piecewise linear polynomials

$$u' = f(x), \quad x \in [a,b], \quad u(a) = u_0. \tag{1}$$

In order to apply the FE method to solve this problem, we carry out the following process.

1. Derive a weak form (variational formulation). This can be done by multiplying the ODE in (1) by a test function $v(x) \in V_0 = \left\{ v \in L^2[a,b] : \|v\|^2 + \|v'\|^2 < \infty,\ v(a) = 0 \right\}$, where $\|v\|^2 = \int_a^b v^2(x)dx$, integrating from a to b, using integration by parts, and applying $v(a) = 0$, to get $\int_a^b fvdx = \int_a^b u'vdx = -\int_a^b uv'dx + u(b)v(b) - u(a)v(a) = -\int_a^b uv'dx + u(b)v(b)$.

2. Generate a triangulation (also called a mesh) of the computational domain $[a,b]$. For a one-dimensional problem, a mesh is a set of points in the interval $[a,b]$, say, $a = x_0 \le x_1 \le \cdots \le x_N = b$. The point x_i is called a node or nodal point. The length of the interval (called an element) $I_i = [x_{i-1}, x_i]$ is $h_i = x_i - x_{i-1}$. Let $h = \max_{1 \le i \le N} h_i$ (called a mesh size that measures how fine the partition is). If the mesh is uniformly distributed, then $x_i = a + ih, i = 0, 1, \ldots, N$, where $h = \frac{b-a}{N}$.

3. Define a finite dimensional space over the triangulation: Let the solution u be in the space V. For the model problem (1), the solution space is $V = C^1[a,b]$. We wish to construct a finite dimensional space (subspace) $V_h \subset V$ based on the mesh. When the FE space is a subspace of the solution space, the method is called conforming. It is known that in this case, the FE solution converges to the true solution provided the FE space approximates the given space in some sense [3]. Different finite dimensional spaces will generate different FE solutions.

   Define the FE space as the set of all continuous piecewise linear polynomials $V_h = \{v : v|_{I_i} \in P^1(I_i),\ \ i = 1, 2, \ldots, N,\ \ v(a) = 0\}$, where $P^1(I_i)$ is the space of polynomials of degree ≤ 1 on I_i. Functions in V_h are linear on each I_i, and continuous on the whole interval $[a,b]$. An example of such a function is shown in **Figure 1**.

   We remark that any function $v \in V_h$ is uniquely determined by its nodal values $v(x_i)$.

4. Construct a set of basis functions based on the triangulation. Since V_h has finite dimension, we can find one set of basis functions. A basis for V_h is $\left\{\phi_j\right\}_{j=0}^{N}$, where $\phi_j \in V_h$ are linearly independent. Then

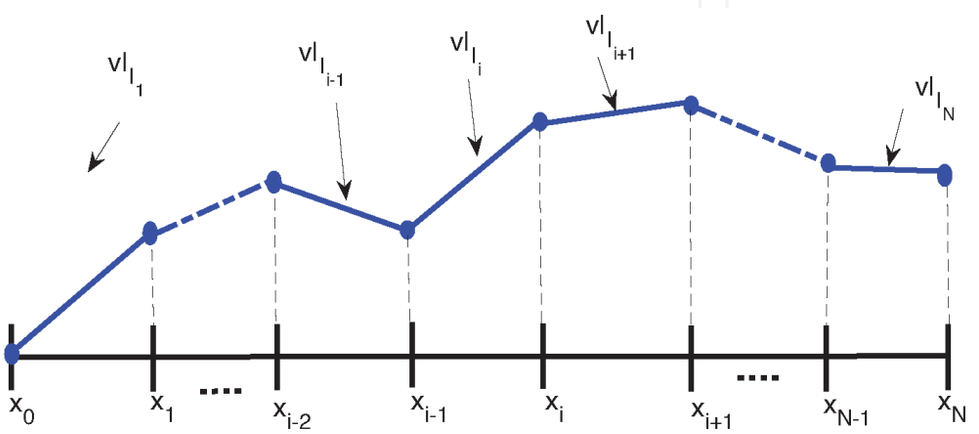

Figure 1.
A continuous piecewise linear function v.

$V_h = \left\{ v_h(x) \in V,\ \ v_h(x) = \sum_{j=0}^{N} c_j \phi_j(x) \right\}$ is the space spanned by the basis functions $\{\phi_i\}_{i=0}^{N}$. The simplest finite dimensional space is the piecewise continuous linear function space defined over the triangulation.

$$V_h = \{ v_h(x) \in V,\ \ v_h(x) \text{ is piecewise continuous linear over } [a, b] \text{ with } v_h(a) = 0 \}.$$

There are infinite number of sets of basis functions. We should choose a set of basis functions that are simple, have compact (minimum) support (that is, zero almost everywhere except for a small region), and meet the regularity requirement, that is, they have to be continuous, and differentiable except at nodal points. The simplest ones are the so-called hat functions satisfying $\phi_i(x_i) = 1$ and $\phi_i(x_j) = 0$ for $i \neq j$. The analytic form is (see **Figure 2**)

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{h}, & x \in I_1, \\ 0, & \text{else}, \end{cases}, \quad \phi_N(x) = \begin{cases} \dfrac{x - x_{N-1}}{h}, & x \in I_N, \\ 0, & \text{else}, \end{cases},$$

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h}, & x \in I_i, \\ \dfrac{x_{i+1} - x}{h}, & x \in I_{i+1}, \\ 0, & \text{else}. \end{cases}$$

5. Approximate the exact solution u by a continuous piecewise linear function $u_h(x)$. The FE method consists of finding $u_h \in V_h$ such that

$$-\int_a^b u_h v' dx + u_h(b) v(b) = \int_a^b f v dx, \quad \forall\ v \in V_h.$$

This type of FE method (with similar trial and test space) is sometimes called a Galerkin method, named after the famous Russian mathematician and engineer Galerkin.

Implementation: The FE solution is a linear combination of the basis functions. Writing $u_h(x) = \sum_{j=0}^{N} c_j \phi_j(x)$, where $c_0, c_1, \ldots, c_N$ are unknowns, and choosing $v = \phi_i,\ \ i = 1, 2, \ldots, N$ to get

$$-\sum_{j=0}^{N} c_j \int_a^b \phi_j \phi_i' dx + c_N \phi_i(b) = \int_a^b f \phi_i dx, \quad i = 1, 2, \ldots, N,$$

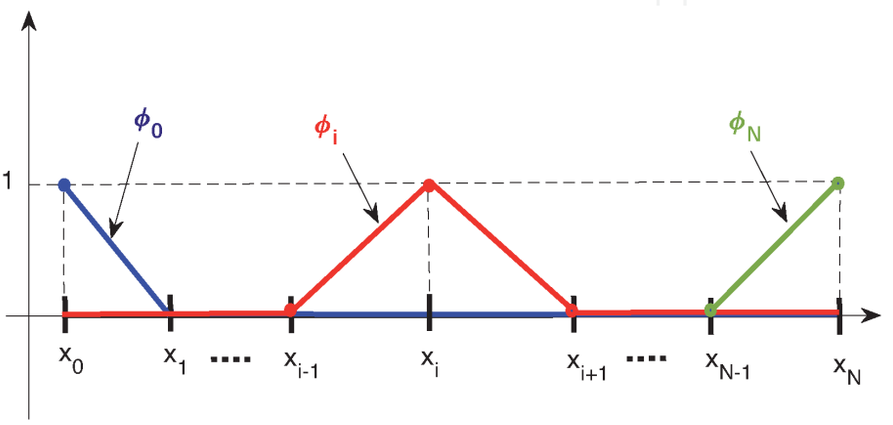

Figure 2.
A typical hat function ϕ_i on a mesh. Also shown is the half hat functions ϕ_0 and ϕ_N.

since $u_h(b) = c_N$. Note that using the hat functions, we have $u_h(x_0) = 0$ and $u_h(x_i) = \sum_{j=0}^{N} c_j \phi_j(x_i) = c_i \phi_i(x_i) = c_i$ for $i = 1, 2, \ldots, N$. Thus, we get the following linear system

$$-\sum_{j=1}^{N} c_j \int_a^b \phi_j \phi_i' dx + c_N \phi_i(b) = \int_a^b \phi_0 \phi_i' dx, \quad i = 1, 2, \ldots, N.$$

Finally, we solve the linear system for $c_1, \ldots, c_N$. We note that for $i = 1, 2, \ldots, N-1$, we have

$$\int_a^b \phi_i \phi_i' dx = \int_{x_{i-1}}^{x_{i+1}} \phi_i \phi_i' dx = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} \left(\frac{x - x_{i-1}}{h_i}\right) dx - \frac{1}{h_i} \int_{x_i}^{x_{i+1}} \left(\frac{x_{i+1} - x}{h_i}\right) dx = 0.$$

However, for $i = N$, we have

$$\begin{aligned} \int_a^b \phi_N \phi_N{}' dx &= \int_{x_{N-1}}^{x_N} \phi_N \phi_N{}' dx = \int_{x_{N-1}}^{x_N} \left(\frac{x - x_{N-1}}{h_N}\right) \left(\frac{x - x_{N-1}}{h_N}\right) dx \\ &= \frac{1}{h_N} \int_{x_{N-1}}^{x_N} \frac{x - x_{N-1}}{h_N} dx = \frac{1}{2}. \end{aligned}$$

Similarly, for $i = 1, 2, \ldots, N$, we have

$$\begin{aligned} \int_a^b \phi_{i-1} \phi_i' dx &= \int_{x_{i-1}}^{x_i} \phi_{i-1} \phi_i' dx = \int_{x_{i-1}}^{x_i} \left(\frac{x_i - x}{h_i}\right) \left(\frac{x - x_{i-1}}{h_i}\right) dx = \frac{1}{h_i} \int_{x_{i-1}}^{x_i} \frac{x_i - x}{h_i} dx = \frac{1}{2}, \\ \int_a^b \phi_{i+1} \phi_i' dx &= \int_{x_i}^{x_{i+1}} \phi_{i+1} \phi_i' dx = \int_{x_i}^{x_{i+1}} \left(\frac{x - x_i}{h_{i+1}}\right) \left(\frac{x_{i+1} - x}{h_{i+1}}\right) dx = -\frac{1}{h_{i+1}} \int_{x_i}^{x_{i+1}} \frac{x - x_i}{h_{i+1}} dx \\ &= -\frac{1}{2}. \end{aligned}$$

We next calculate $\int_a^b f \phi_i dx$. Since it depends on f, we cannot generally expect to calculate it exactly. However, we can approximate it using a quadrature rule. Using the Trapezoidal rule $\int_a^b f(x) dx \approx \frac{b-a}{2} (f(a) + f(b))$ and using $\phi_i(x_{i-1}) = \phi_i(x_{i+1}) = 0$ and $\phi_i(x_i) = 1$, we get

$$\int_a^b f \phi_i dx = \int_{x_{i-1}}^{x_i} f \phi_i dx + \int_{x_i}^{x_{i+1}} f \phi_i dx \approx \frac{h_i + h_{i+1}}{2} f(x_i), \quad i = 1, 2, \ldots, N-1,$$

$$\int_a^b f(x) \phi_N dx = \int_{x_{N-1}}^{x_N} f(x) \phi_N dx \approx \frac{h_N}{2} (f(x_{N-1}) \phi_N(x_{N-1}) + f(x_N) \phi_N(x_N)) = \frac{h_N}{2} f(x_N).$$

Thus, we obtain the following linear system of equations

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \cdots & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{N-1} \\ c_N \end{bmatrix} = \begin{bmatrix} \frac{h_1 + h_2}{2} f(x_1) \\ \frac{h_2 + h_3}{2} f(x_2) \\ \vdots \\ \frac{h_{N-1} + h_N}{2} f(x_{N-1}) \\ \frac{h_N}{2} f(x_N) \end{bmatrix}.$$

The determinant of the above matrix is $\frac{1}{2^N}$. Thus, the system has a unique solution $c_1, c_2, \ldots, c_N$.

Remark 2.1 *Suppose that $u(a) = u_0$, then we let $u_h(x) = \sum_{j=0}^{N} c_j \phi_j(x)$. Since $u_0 = u_h(x_0) = \sum_{j=1}^{N} c_j \phi_j(x_0) = c_0 \phi_0(x_0) = c_0$, we only need to find $c_1, c_2, \ldots, c_N$. Choosing $v = \phi_i$, $i = 1, 2, \ldots, N$, we get the following linear system*

$$-\sum_{j=1}^{N} c_j \int_a^b \phi_j \phi_i' dx + c_N \phi_i(b) = \int_a^b f \phi_i dx + u_0 \int_a^b \phi_0 \phi_i' dx, \quad i = 1, 2, \ldots, N.$$

Finally, we solve the linear system for $c_1, \ldots, c_N$. We note that $\int_a^b \phi_0 \phi_i' dx = 0$ for $i = 2, \ldots, N$ and

$$\int_a^b \phi_0 \phi_1{}' dx = \int_{x_0}^{x_1} \left(\frac{x_1 - x}{h_1}\right)\left(\frac{x - x_0}{h_1}\right)' dx = \frac{1}{h_1}\int_{x_0}^{x_1} \frac{x_1 - x}{h_1} dx = \frac{1}{2}.$$

Following the same steps used for the case $u(a) = 0$, we obtain the following linear system of equations

$$\begin{bmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \cdots & 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{N-1} \\ c_N \end{bmatrix} = \begin{bmatrix} \frac{h_1 + h_2}{2} f(x_1) + \frac{u_0}{2} \\ \frac{h_2 + h_3}{2} f(x_2) \\ \vdots \\ \frac{h_{N-1} + h_N}{2} f(x_{N-1}) \\ \frac{h_N}{2} f(x_N) \end{bmatrix}.$$

2.2 The FE method for first-order nonlinear IVPs

Here, we extend the FE method for the nonlinear IVP using piecewise linear polynomials

$$u' = f(x, u), \quad x \in [a, b], \quad u(a) = u_0. \tag{2}$$

The FE method consists of finding $u_h \in V_h = \{v : v|_{I_i} \in P^1(I_i),$ $i = 1, 2, \ldots, N,\ v(a) = 0\}$, such that

$$u_h(b)v(b) - \int_a^b u_h v' dx = \int_a^b f(x, u_h) v dx, \quad \forall\ v \in V_h.$$

Writing $u_h(x) = \sum_{j=0}^{N} c_j \phi_j(x)$ and choosing $v = \phi_i$, $i = 1, 2, \ldots, N$, we get

$$c_N\left(\phi_i - \int_a^b \phi_N \phi_i' dx\right) - \sum_{j=0}^{N-1} c_j \int_a^b \phi_j \phi_i' dx - \int_a^b f\left(x, \sum_{j=0}^{N} c_j \phi_j\right) \phi_i dx = 0,$$
$$i = 1, 2, \ldots, N,$$

where $u_h(x_0) = c_0 = u_0$. Finally, we solve the nonlinear system for $c_1, c_2, \ldots, c_N$ using *e.g.*, Newton's method for systems of nonlinear equations. The system can be written as $F_i(c_1, c_2, \ldots, c_N) = 0$, $i = 1, 2, \ldots, N$, where

$$F_i = c_N\left(\phi_i - \int_a^b \phi_N \phi_i' dx\right) - \sum_{j=0}^{N-1} c_j \int_a^b \phi_j \phi_i' dx - \int_a^b f\left(x, \sum_{j=0}^{N} c_j \phi_j\right)\phi_i dx,$$
$$i = 1, 2, \ldots, N.$$

Let $\alpha_i = \sum_{j=0}^{N} c_j \int_a^b \phi_j \phi_i' dx$ and $\beta_i = \int_a^b f\left(x, \sum_{j=0}^{N} c_j \phi_j\right)\phi_i dx$. Then, for $i = 1, 2, \ldots, N-1$,

$$\begin{aligned}
\alpha_i &= c_{i-1}\int_{x_{i-1}}^{x_i} \phi_{i-1}\phi_i' dx + c_i\left(\int_{x_{i-1}}^{x_i} \phi_i \phi_i' dx + \int_{x_i}^{x_{i+1}} \phi_i \phi_i' dx\right) + c_{i+1}\int_{x_i}^{x_{i+1}} \phi_{i+1}\phi_i' dx \\
&= c_{i-1}\int_{x_{i-1}}^{x_i} \frac{x_i - x}{h_i^2} dx + c_i\left(\int_{x_{i-1}}^{x_i} \frac{x - x_{i-1}}{h_i^2} dx - \int_{x_i}^{x_{i+1}} \frac{x_{i+1} - x}{h_{i+1}^2} dx\right) - c_{i+1}\int_{x_i}^{x_{i+1}} \frac{x - x_i}{h_{i+1}^2} dx \\
&= \frac{1}{2}c_{i-1} + c_i\left(\frac{1}{2} - \frac{1}{2}\right) - \frac{1}{2}c_{i+1} = \frac{1}{2}c_{i-1} - \frac{1}{2}c_{i+1},
\end{aligned}$$

$$\begin{aligned}
\alpha_N &= c_{N-1}\int_{x_{N-1}}^{x_N} \phi_{N-1}\phi_N{}' dx + c_N\int_{x_{N-1}}^{x_N} \phi_N \phi_N{}' dx \\
&= c_{N-1}\int_{x_{N-1}}^{x_N} \frac{x_N - x}{h_N^2} dx + c_N\int_{x_{N-1}}^{x_N} \frac{x - x_{N-1}}{h_N^2} dx = \frac{1}{2}c_{N-1} + \frac{1}{2}c_N.
\end{aligned}$$

Similarly,

$$\beta_i = \int_{x_{i-1}}^{x_{i+1}} f\left(x, \sum_{j=0}^{N} c_j \phi_j\right)\phi_i dx = \int_{x_{i-1}}^{x_i} f\left(x, \sum_{j=0}^{N} c_j \phi_j\right)\phi_i dx + \int_{x_i}^{x_{i+1}} f\left(x, \sum_{j=0}^{N} c_j \phi_j\right)\phi_i dx.$$

Using Simpson's Rule $\int_a^b f(x)dx \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$, and using $\phi_i(x_{i-1}) = \phi_i(x_{i+1}) = 0$, $\phi_i(x_i) = 1$, $\sum_{j=0}^{N} c_j \phi_j\left(x_{i-1} + \frac{h_i}{2}\right) = \frac{c_{i-1}+c_i}{2}$, $\phi_i\left(x_{i-1} + \frac{h_i}{2}\right) = \frac{1}{2}$, $\sum_{j=0}^{N} c_j \phi_j(x_i) = c_i$, we have, for $i = 1, 2, \ldots, N-1$,

$$\beta_i \approx \frac{h_i}{3} f\left(x_{i-1} + \frac{h_i}{2}, \frac{c_{i-1} + c_i}{2}\right) + \frac{h_i + h_{i+1}}{6} f(x_i, c_i) + \frac{h_{i+1}}{3} f\left(x_i + \frac{h_{i+1}}{2}, \frac{c_i + c_{i+1}}{2}\right).$$

However, for $i = N$, we have

$$\beta_N \approx \frac{h_N}{6}\left(2f\left(x_{N-1} + \frac{h}{2}, \frac{c_{N-1} + c_N}{2}\right) + f(x_N, c_N)\right).$$

Next, we compute the Jacobian matrix with entries

$$J_{i,j} = \frac{\partial F_i}{\partial c_j} = \int_a^b \phi_j \phi_i' dx - \int_a^b f_u\left(x, \sum_{j=0}^{N} c_j \phi_j\right)\phi_j \phi_i \, dx = a_{i,j} - b_{i,j}, \quad i = 1, 2, \ldots, N.$$

We already computed the entries $a_{i,j}$ as

$$a_{i,i-1} = \int_a^b \phi_{i-1}\phi_i' dx = \frac{1}{2}, \quad a_{i,i} = \int_a^b \phi_i \phi_i' dx = 0, \quad i = 1, 2, \ldots, N-1,$$
$$a_{N,N} = \int_a^b \phi_N \phi_N{}' dx = \frac{1}{2}, \quad a_{i,i+1} = \int_a^b \phi_{i+1}\phi_i' dx = -\frac{1}{2}.$$

Using Simpson's Rule, we get

$$b_{i,i-1} = \int_{x_{i-1}}^{x_i} \phi_{i-1}\phi_i f_u\left(x, \sum_{j=0}^{N} c_j\phi_j\right)dx \approx \frac{h_i}{6} f_u\left(x_{i-1} + \frac{h_i}{2}, \frac{c_{i-1}+c_i}{2}\right),$$

$$b_{i,i+1} = \int_{x_i}^{x_{i+1}} \phi_{i+1}\phi_i f_u\left(x, \sum_{j=0}^{N} c_j\phi_j\right)dx \approx \frac{h_{i+1}}{6} f_u\left(x_i + \frac{h_{i+1}}{2}, \frac{c_i+c_{i+1}}{2}\right),$$

$$b_{i,i} = \int_{x_{i-1}}^{x_i} \phi_i^2 f_u\left(x, \sum_{j=0}^{N} c_j\phi_j\right)dx + \int_{x_i}^{x_{i+1}} \phi_i^2 f_u\left(x, \sum_{j=0}^{N} c_j\phi_j\right)dx$$

$$\approx \frac{h_i}{6} f_u\left(x_{i-1} + \frac{h_i}{2}, \frac{c_{i-1}+c_i}{2}\right) + \frac{h_i+h_{i+1}}{6} f_u(x_i, c_i) + \frac{h_{i+1}}{6} f_u\left(x_i + \frac{h_{i+1}}{2}, \frac{c_i+c_{i+1}}{2}\right),$$

$$b_{N,N} = \int_{x_{N-1}}^{x_N} \phi_N^2 f_u\left(x, \sum_{j=0}^{N} c_j\phi_j\right)dx \approx \frac{h_N}{6}\left(f_u\left(x_{N-1} + \frac{h}{2}, \frac{c_{N-1}+c_N}{2}\right) + f_u(x_N, c_N)\right).$$

2.3 The FE method for two-point BVPs

Here, we shall study the derivation and implementation of the FE method for two-point boundary-value problems (BVPs). For easy presentation, we consider the following model problem: Find $u \in C^2[a,b]$ such that

$$-u'' + q(x)u = f(x), \quad x \in \Omega = (a,b), \quad u(a) = u(b) = 0, \tag{3}$$

where $u : \overline{\Omega} = [a,b] \to \mathbb{R}$ is the sought solution, $q(x) \geq 0$ is a continuous function on $[a,b]$, and $f \in L^2[a,b]$. Under these assumptions, (3) has a unique solution $u \in C^2[a,b]$. For general $q(x)$, it is impossible to find an explicit form of the solution. Therefore, our goal is to obtain a numerical solution via the FE method.

2.3.1 Different mathematical formulations for the 1D model

The model problem (3) can be reformulated into three different forms:

(D)-form: the original differential equation (3).

(V)-form: the variational form or weak form: $\int_a^b u'v'dx + \int_a^b quvdx = \int_a^b fvdx$, for any test function v in the Sobolev space $H_0^1[a,b] = \{v \in L^2[a,b] : \|v\|^2 + \|v'\|^2 < \infty,\ v(a) = v(b) = 0\}$, where $\|v\|^2 = \int_a^b v^2(x)dx$. The corresponding FE method is often called the *Galerkin method*. In other words, a Galerkin FE method is a FE method obtained from the variational form.

(M)-form: the minimization form: $\min_{v(x) \in H_0^1[a,b]} \int_a^b \left(\frac{1}{2}(v')^2 + \frac{1}{2}qv^2 - fv\right)dx$. The corresponding FE method is often called the *Ritz method*.

Under some assumptions, the three different forms are equivalent, that is, they have the same solution as will be explained in the following theorem.

Theorem 2.1 (Mathematical equivalences) *Suppose that u'' exists and continuous on $[a,b]$. Then we have the following mathematical equivalences.*

(D) is equivalent to (V), (V) is equivalent to (M), and (M) is equivalent to (D).

2.3.2 Galerkin method of the problem

To solve (3) using the FE method, we carry out the process described below. Usually, a FE method is always derived from the weak or variational formulation of the problem at hand.

Weak formulation of the problem: The Galerkin FE method starts by rewriting (3) in an equivalent variational formulation. To this end, let us define the vector space $H_0^1 = \left\{v \in L^2(a,b) : \|v\|^2 + \|v'\|^2 < \infty, \;\; v(a) = v(b) = 0\right\}$. Multiplying (3) by a test function $v \in H_0^1$, integrating from a to b, and using integration by parts, we get

$$\int_a^b fvdx = \int_a^b -u''vdx + \int_a^b quvdx = \int_a^b u'v'dx + \int_a^b quvdx,$$

since $v(a) = v(b) = 0$. Hence, the weak (or variational) form of (3) reads: Find $u \in H_0^1$, such that

$$\int_a^b u'v'dx + \int_a^b quvdx = \int_a^b fvdx, \quad \forall \;\; v \in H_0^1. \tag{4}$$

We want to find $u \in H_0^1$ that satisfies (4). We note that a solution u to (4) is less regular than the solution u (3). Indeed, (4) has only u' whereas (3) contains u''. Furthermore, we can easily verify the following:

1. If u is strong solution (*i.e.*, solves (3)) then u is also weak solution (*i.e.*, solves (4)).

2. Conversely, if u is a weak solution with $u \in C^2[a,b]$, it is also strong solution.

3. Existence and uniqueness of weak solutions is obtained by the Lax-Milgram Theorem.

4. We can consider solutions with lower regularity using the weak formulation.

5. FE method gives an approximation of the weak solution.

From now on, we use the notation $\|v\| = \|v\|_\Omega$, where $\Omega = [a,b]$.

The FE formulation: The FE method is based on the variational form (4). We note that the space H_0^1 contains many functions and it is therefore just as hard to find a function $u \in H_0^1$ which satisfies the variational Eq. (4) as it is to solve the original problem (3). Next, we study in details a special Galerkin method called the FE method. Let $a = x_0 < x_1 < \cdots < x_N = b$ be a regular partition of $[a,b]$. Suppose that the length of $I_i = [x_{i-1}, x_i]$ is $h_i = x_i - x_{i-1}$. We define $h = \max\limits_{i=1,2,\ldots,N} h_i$ to be the mesh size. We wish to construct a subspace $V_h \subset V = H_0^1$. Since V_h has finite dimension, we can find one set of basis functions $\left\{\phi_j\right\}_{j=1}^{N-1}$ for V_h, where $\phi_j \in V_h, \;\; j = 1,2,\ldots,N-1$ are linearly independent. We remark that V_h is the space spanned by the basis functions *i.e.*, $V_h = \left\{v_h(x), \;\; v_h(x) = \sum_{j=1}^{N-1} c_j \phi_j(x)\right\}$. The FE method consists of choosing a basis for the subspace V_h that satisfies the following properties

1. The matrix A must be sparse (*e.g.* traditional or banded matrix). In this case, iterative methods for solving linear systems can be adapted to obtain an efficient solution.

2. u_h must converge to the solution u of the original problem as $h \to 0$.

It is natural to obtain an approximation u_h to u as follows: Find $u_h \in V_h$ such that

$$\int_a^b u_h' v' dx + \int_a^b q u_h v dx = \int_a^b f v dx, \quad \forall\ v \in V_h. \tag{5}$$

We call u_h the FE approximation of u. We say that (5) is the Galerkin approximation of (4) and the method used to find $u_h \in V_h$ is called Galerkin method.

FE approximation using Lagrange $\mathbb{P}_1$ elements: The simplest finite dimensional space is the piecewise continuous linear function space defined over the triangulation

$$V_{h,0}^1 = \{v_h \in V,\ \ v_h \ \text{ is piecewise continuous linear over [a, b] with } \ v_h(a) = v_h(b) = 0\}.$$

It is easy to show that $V_{h,0}^1$ has a finite dimension even although there are infinite number of elements in $V_{h,0}^1$. The approximation of the FE method is therefore to look for an approximation u_h within a small (finite dimensional) subspace $V_{h,0}^1 = \left\{v \in V_h^1 |\ \ v(a) = v(b) = 0\right\}$ of H_0^1, consisting of piecewise linear polynomials, where $V_h^1 = \left\{v \in C^0[a,b] |\ \ v|_{I_i} \in P^1(I_i)\right\}$.

Let $V_{h,0}^1$ be the space of all continuous piecewise linear functions, which vanish at the end points a and b. There are many types of basis functions $\{\phi_i\}_{i=1}^{N-1}$. The simplest ones are the so-called hat functions satisfying $\phi_i(x_j) = \delta_{ij}$, where δ_{ij} is the Kronecker symbol. Note especially that there is no need to construct hat functions ϕ_0 and ϕ_N since any function of $V_{h,0}^1$ must vanish at the end points $x_0 = a$ and $x_N = b$.

The explicit expressions for the hat function $\phi_i(x)$ and its derivative $\phi_i'(x)$ are given by

$$\phi_i(x) = \begin{cases} 0, & a \le x \le x_{i-1}, \\ \dfrac{x - x_{i-1}}{h_i}, & x_{i-1} \le x \le x_i, \\ \dfrac{x_{i+1} - x}{h_{i+1}}, & x_i \le x \le x_{i+1}, \\ 0, & x_{i+1} \le x \le b, \end{cases}, \quad \phi_i'(x) = \begin{cases} 0, & a < x < x_{i-1}, \\ \dfrac{1}{h_i}, & x_{i-1} < x < x_i, \\ -\dfrac{}{h_{i+1}}, & x_i < x < x_{i+1}, \\ 0, & x_{i+1} < x < b, \end{cases},$$

for $i = 1, 2, \ldots, N-1$. The FE approximation of (4) thus reads: Find $u \in V_{h,0}^1$, such that

$$\int_a^b u_h' v' dx + \int_a^b q u_h v dx = \int_a^b f v dx, \quad \forall\ v \in V_{h,0}^1. \tag{6}$$

We call u_h the FE approximation of u. We say that (6) is the Galerkin approximation of (4) and the method used to find $u_h \in V_{h,0}^1$ is called Galerkin method.

It can be shown that (6) is equivalent to the $N-1$ equations

$$\int_a^b u_h' \phi_i' dx + \int_a^b q u_h \phi_i dx = \int_a^b f \phi_i dx, \quad i = 1, 2, \ldots, N-1. \tag{7}$$

Derivation of the discrete system: Since $u_h \in V_{h,0}^1$, we can express it as a linear combination of hat functions *i.e.*,

$$u_h = \sum_{j=1}^{N-1} c_j \phi_j(x), \tag{8}$$

where c_j are real numbers to be determined. We note that the coefficients $c_j,\ j = 1, 2, \ldots, N-1$ are the $N-1$ nodal values of u_h to be determined. Note that the index is only from 1 to $N-1$, because of the zero boundary conditions. We remark that $u_h(a) = u_h(b) = 0$ and $u_h(x_i) = c_i$. So c_i is an approximate solution to the exact solution at $x = x_i$.

We can use either the weak/variational form (V), or the minimization form (M), to derive a linear system of equations for the coefficients c_j.

Substituting (8) into (7) yields

$$\sum_{j=1}^{N-1} c_j \left(\int_a^b \phi_i' \phi_j' dx + \int_a^b q \phi_i \phi_j dx \right) = \int_a^b f \phi_i dx, \quad i = 1, 2, \ldots, N-1. \tag{9}$$

The problem (7) is now equivalent to the following: Find the real numbers $c_1, c_2, \ldots, c_{N-1}$ that satisfy the linear system (9).

We note that the linear system (9) is equivalent to the system in matrix–vector form

$$A\mathbf{c} = \mathbf{b}, \tag{10}$$

where $\mathbf{c} = [c_1, c_2, \ldots, c_{N-1}]^t \in \mathbb{R}^{N-1}$ is the unknown vector, A is an $(N-1) \times (N-1)$ matrix, the so-called stiffness matrix when $q = 0$, with entries

$$a_{ij} = \int_a^b \left(\phi_i' \phi_j' + q \phi_i \phi_j \right) dx, \quad i, j = 1, 2, \ldots, N-1, \tag{11}$$

and $\mathbf{b} \in \mathbb{R}^{N-1}$, the so-called load vector, has entries

$$b_i = \int_a^b f \phi_i dx, \quad i = 1, 2, \ldots, N-1. \tag{12}$$

To obtain the approximate solution we need to solve the linear system for the unknown vector $\mathbf{c}$. We note that $a_{ij} = a\left(\phi_i, \phi_j\right)$ and $b_i = (f, \phi_i)$, where $a(u, v) = \int_a^b (u'v' + quv) dx$ is a bi-linear and $(f, v) = \int_a^b fv dx$ is a linear form.

2.3.3 Ritz method of the problem

The Ritz method is one of the earliest FE methods. However, not every problem has a minimization form. The minimization form for the model problem (3) is

$$\min_{v(x) \in H_0^1[a,b]} F(v), \quad \text{where } F(v) = \int_a^b \left(\frac{1}{2} (v')^2 + \frac{1}{2} q v^2 - fv \right) dx.$$

As before, we look for an approximate solution of the form (8). If we plug this into the functional form above, we get

$$F(u_h) = \int_a^b \left(\frac{1}{2} \left(\sum_{j=1}^{N-1} c_j \phi_j'(x) \right)^2 + \frac{1}{2} q \left(\sum_{j=1}^{N-1} c_j \phi_j(x) \right)^2 - f \left(\sum_{j=1}^{N-1} c_j \phi_j(x) \right) \right) dx,$$

which is a multi-variable function of $c_1, c_2, \ldots, c_{N-1}$ and can be written as $F(u_h) = F(c_1, c_2, \ldots, c_{N-1})$. The necessary conditions for a global minimum are $\frac{\partial F}{\partial c_i} = 0, i = 1, 2, \ldots, N-1$. Taking the partial derivatives directly with respect to c_i, we get

$$\int_a^b \left(\phi_i'(x) \sum_{j=1}^{N-1} c_j \phi_j'(x) + q \phi_i(x) \sum_{j=1}^{N-1} c_j \phi_j(x) - f \phi_i(x) \right) dx = 0, \quad i = 1, 2, \ldots, N-1.$$

Exchange the order of integration and the summation, we get

$$\sum_{j=1}^{N-1} c_j \int_a^b \left(\phi_i'(x) \phi_j'(x) + q \phi_i(x) \phi_j(x) \right) dx = \int_a^b f \phi_i(x) dx = 0, \quad i = 1, 2, \ldots, N-1,$$

which is exactly the same linear system (9) obtained using the Galerkin method.

2.3.4 Computer implementation

It is straightforward to calculate the entries $\hat{a}_{i,j} = \int_a^b \phi_i' \phi_j' dx$. For $|i-j| > 1$, we have $\hat{a}_{i,j} = 0$, since ϕ_i and ϕ_j lack overlapping support. However, if $i = j$, then

$$\hat{a}_{i,i} = \int_a^b (\phi_i')^2 dx = \int_{x_{i-1}}^{x_i} \left(\frac{1}{h_i} \right)^2 dx + \int_{x_i}^{x_{i+1}} \left(-\frac{1}{h_{i+1}} \right)^2 dx = \frac{1}{h_i} + \frac{1}{h_{i+1}}, \quad i, j = 1, 2, \ldots, N-1.$$

Furthermore, if $j = i + 1$, then

$$\hat{a}_{i,i+1} = \int_a^b \phi_i' \phi_{i+1}' dx = \int_{x_i}^{x_{i+1}} \left(-\frac{1}{h_{i+1}} \right) \left(\frac{1}{h_{i+1}} \right) dx = -\frac{1}{h_{i+1}}, \quad i, j = 1, 2, \ldots, N-2. \tag{13}$$

By symmetry, we also have

$$\hat{a}_{i+1,i} = \int_a^b \phi_{i+1}' \phi_i' dx = -\frac{1}{h_{i+1}}, \quad i, j = 1, 2, \ldots, N-2.$$

To obtain $\tilde{a}_{i,j} = \int_a^b q \phi_i \phi_j dx$ and $b_i = \int_a^b f \phi_i dx$, we use the composite trapezoidal rule

$$\int_a^b f(x) dx = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} f(x) dx \approx \frac{1}{2} \left[h_1 f(x_0) + \sum_{i=1}^{N-1} (h_i + h_{i+1}) f(x_i) + h_N f(x_N) \right].$$

So, we can easily verify that

$$\tilde{a}_{i,j} = \int_a^b q \phi_i \phi_j dx \approx \begin{cases} \frac{q_i}{2}(h_i + h_{i+1}), & i = j \\ 0, & i \neq j \end{cases}, \quad b_i = \int_a^b f \phi_i dx \approx \frac{1}{2}(h_i + h_{i+1}) f_i,$$

where $q_i = q(x_i)$ and $f_i = f(x_i)$. Thus, the matrix $A = \left(\hat{a}_{i,j} + \tilde{a}_{i,j}\right)$ is tridiagonal and has the form

$$A = \begin{bmatrix} \frac{1}{h_1}+\frac{1}{h_2}+\frac{q_1}{2}(h_1+h_2) & -\frac{1}{h_2} & 0 & \cdots & 0 \\ -\frac{1}{h_2} & \frac{1}{h_2}+\frac{1}{h_3}+\frac{q_2}{2}(h_2+h_3) & -\frac{1}{h_3} & \ddots & 0 \\ 0 & -\frac{1}{h_3} & \ddots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{1}{h_{N-1}} \\ 0 & \cdots & 0 & -\frac{1}{h_{N-1}} & \frac{1}{h_{N-1}}+\frac{1}{h_N}+\frac{q_{N-1}}{2}(h_{N-1}+h_N) \end{bmatrix}.$$

Finally, we obtain the following system: $c_0 = c_N = 0$ and

$$-\frac{1}{h_i}c_{i-1} + \frac{1}{h_i + h_{i+1}}c_i - \frac{1}{h_{i+1}}c_{i+1} + \frac{q_i(h_i + h_{i+1})}{2}c_i = \frac{1}{2}(h_i + h_{i+1})f_i, \quad i = 1, 2, \ldots, N-1,$$

Remark 2.2 *Suppose that the partition is uniform i.e.,* $h_i = h = \frac{b-a}{N}$ *for all* $i = 1, 2, \ldots, N$. *Then the stiffness matrix A and the load vector* **b** *have the form:*

$$A = \begin{bmatrix} \frac{2}{h}+hq_1 & -\frac{1}{h} & 0 & \cdots & 0 \\ -\frac{1}{h} & \frac{2}{h}+hq_2 & -\frac{1}{h} & \ddots & 0 \\ 0 & -\frac{1}{h} & \ddots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{1}{h_{N-1}} \\ 0 & \cdots & 0 & -\frac{1}{h} & \frac{2}{h}+hq_{N-1} \end{bmatrix}, \quad \mathbf{b} = h\begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{N-1} \end{bmatrix}.$$

Finally, we obtain the following system: $c_0 = c_N = 0$ and

$$\frac{-c_{j-1} + 2c_j - c_{j+1}}{h} + hq_i c_j = hf_i \Rightarrow -\frac{c_{j-1} - 2c_j + c_{j+1}}{h^2} + q_i c_j = f_i, \quad i = 1, 2, \ldots, N-1,$$

which is the same system obtained using the finite difference method, where u'' is approximated using the second-order midpoint formula $u''(x_j) \approx \frac{u(x_{j-1}) - 2u(x_j) + u(x_{j+1})}{h^2}$. We conclude that the above FE method using the composite trapezoidal rule is equivalent to the finite difference method of order 2.

2.3.5 Existence, uniqueness, and basic a priori *error estimate*

Lemma 2.1 *The matrix A with entries* $a_{i,j} = \int_a^b \phi_i' \phi_j' dx$ *is symmetric positive definite i.e.,* $a_{i,j} = a_{j,i}$ *and*

$$\mathbf{x}^t A \mathbf{x} = \sum_{i,j=1}^{N-1} x_i a_{i,j} x_j > 0, \quad \text{for all nonzero } \mathbf{x} = [x_1, \ldots, x_{N-1}]^t \in \mathbb{R}^{N-1}.$$

Theorem 2.2 *The linear system (10) obtained using the FE method has a unique solution. Consequently, the FE method solution* u_h *is unique*.

Next, we state a general convergence result for the Galerkin method. We first define the following norm and semi-norm: For $v \in H_0^1$, we define

$$\|v\| = \left(\int_a^b v^2(x)dx\right)^{1/2}, \quad |v|_1 = \|v'\| = \left(\int_a^b (v'(x))^2 dx\right)^{1/2}.$$

Theorem 2.3 *Suppose that $q(x) \geq 0$, $\forall\ x \in [a,b]$. Let u be the solution to (4) and u_h be the solution to (6). Then there exists a constant C such that*

$$\|(u-u_h)'\| \leq C\|(u-v_h)'\|, \quad \forall\ v_h \in V_{h,0}^1, \tag{14}$$

where C is given by $C = 1 + \max_{x\in[a,b]}|q(x)|$, which is independent of the choice of $V_{h,0}^1$.

Remark 2.3 *From (14), taking the minimum over all $v_h \in V_{h,0}^1$, we get $\|(u-u_h)'\| \leq C \min_{v_h \in V_{h,0}} \|(u-v_h)'\|$. Thus, $|u-u_h|_1 \leq C \min_{v_h\in V_{h,0}}|u-v_h|_1$, where $C = 1 + \max_{x\in[a,b]}|q(x)|$.*

Next, we study the convergence of u_h to u. Let $u \in H_0^1$. Define the piecewise linear interpolant by

$$\pi u = \sum_{j=1}^{N} u(x_j)\phi_j(x) \in V_{h,0}^1, \quad x \in [a,b].$$

Since $\pi u \in V_{h,0}^1$, the estimate (14) gives

$$\|(u-u_h)'\| \leq C\|(u-\pi u)'\|.$$

This inequality suggest that the error between u and u_h is controlled by the interpolation error $u - \pi u$ in the $|\cdot|_1$-norm.

Theorem 2.4 (**A priori error estimate**) *Suppose that $q(x) \geq 0\ \forall\ x \in [a,b]$. Let u be the solution to (4) and u_h be the solution to (6). Then there exists a constant C such that*

$$\|(u-u_h)'\|^2 \leq C\sum_{i=1}^{N} h_i^2 \|u''\|_{I_i}^2,$$

where C is a constant independent of h. Consequently, if $h = \max_i h_i$, then

$$\|(u-u_h)'\|^2 \leq Ch^2\|u''\|^2.$$

Remark 2.4

1. If the partition is not uniform then we obtain the same error estimate with $h = \max_{i=1,2,\ldots,N}(x_i - x_{i-1})$.

2. The error is expressed in terms of the exact solution u. If it is expressed in terms of the computed solution u_h it is an *a posteriori* error estimate (this yields a computable error bound).

3. $u_h \to u$ in the $\|v'\|$-norm as $h = max_i(h_i) \to 0$. If $\|(u-u_h)'\| = 0$ then $u - u_h$ is constant, but since $u(0) = u_h(0)$ we also have $u - u_h = 0$ and therefore $u_h = u$.

4. u_h is the best approximation within the space $V_{h,0}^1$ with respect to the $\|v'\|$-norm.

5. The norm $\|v'\|$ is referred to as the energy norm and has often a physical meaning.

2.3.6 Boundary conditions

In problem (3) we considered a homogeneous Dirichlet boundary conditions. Here, we extend the FE method to boundary conditions of different types. There are three important types of boundary conditions (BCs):

1. Dirichlet BCs: $u(a) = \alpha$ and $u(b) = \beta$ for two real numbers α and β. This BC is also known as strong BC or essential BC.

2. Neumann BCs: $u'(a) = \alpha$ and $u'(b) = \beta$ for two real numbers α and β. This BC is also known as natural BCs.

3. Robin BCs: $u'(a) = \alpha u(a)$ and $u'(b) = \beta u(b)$ for two real numbers α and β.

Note that any combination is possible at the two boundary points.

Nonhomogeneous Dirichlet boundary conditions: Let us consider the following two-point BVP: find $u \in C^2(a,b)$ such that

$$-u'' = f(x), \quad x \in (a,b), \quad u(a) = \alpha, \quad u(b) = \beta, \tag{15}$$

where α and β are given constants and $f \in C(a,b)$ is a given function. In this case, the admissible function space $H_0^1 = \left\{ v : \|v\|^2 + \|v'\|^2 < \infty, \;\; v(a) = v(b) = 0 \right\}$ and the FE space $V_{h,0}^1$ defined earlier remain the same. Multiplying (15) by a test function $v \in H_0^1$ and integrating by parts gives

$$\int_a^b fv dx = \int_a^b -u''v dx = -u'(b)v(b) + u'(a)v(a) + \int_a^b u'v' dx = \int_a^b u'v' dx,$$

since $v(a) = v(b) = 0$. Hence, the weak or variational form of (15) reads: Given $u(a) = \alpha$, $u(b) = \beta$, find $u \in H^1 = \left\{ v : \|v\|^2 + \|v'\|^2 < \infty \right\}$, such that

$$\int_a^b u'v' dx = \int_a^b fv dx, \quad \forall \;\; v \in H_0^1. \tag{16}$$

Let V_h^1 and $V_{h,0}^1$, respectively, be the space of all continuous piecewise linear functions and the space of all continuous piecewise linear functions which vanish at the endpoints a and b. We also let $a = x_0 < x_1 < \cdots < x_N = b$ be a uniform partition of the interval $[a,b]$. Moreover let $\{\phi_i\}$ be the set of hat basis functions of V_h associated with the $N+1$ nodes x_j, $j = 0, 1, \ldots, N$, such that $\phi_i(x_j) = \delta_{ij}$. The FE approximation of (16) thus reads: Find $u_h \in V_h^1$ such that $u_h(a) = \alpha$, $u_h(b) = \beta$, and

$$\int_a^b u_h' v' dx = \int_a^b fv dx, \quad \forall \;\; v \in V_{h,0}^1. \tag{17}$$

It can be shown that (17) is equivalent to the $N-1$ equations

$$\int_a^b u_h' \phi_i' dx = \int_a^b f \phi_i dx, \quad i = 1, 2, \ldots, N-1. \tag{18}$$

Expanding u_h as a linear combination of hat functions

$$u_h = \sum_{j=0}^{N} c_j \phi_j(x) = \alpha \phi_0(x) + \sum_{j=1}^{N-1} c_j \phi_j(x) + \beta \phi_N(x), \tag{19}$$

where the coefficients c_j, $j = 1, 2, \ldots, N-1$ are the $N-1$ nodal values of u_h to be determined.

Substituting (19) into (18) yields

$$\sum_{j=1}^{N-1} c_j \left(\int_a^b \phi_i' \phi_j' dx \right) = \int_a^b \left(f \phi_i - \alpha \phi_0' \phi_i' - \beta \phi_N' \phi_i' \right) dx, \quad i = 1, 2, \ldots, N-1,$$

which is a $(N-1) \times (N-1)$ system of equations for c_j. In matrix form we write

$$A\mathbf{c} = \mathbf{b}, \tag{20}$$

where A is a $(N-1) \times (N-1)$ matrix, the so-called stiffness matrix, with entries

$$a_{i,j} = \int_a^b \phi_i' \phi_j' dx, \quad i, j = 1, 2, \ldots, N-1, \tag{21}$$

$\mathbf{c} = [c_1, c_2, \ldots, c_{N-1}]^t$ is a $(N-1)$ vector containing the unknown coefficients c_j, $j = 1, 2, \ldots, N-1$, and $\mathbf{b}$ is a $(N-1)$ vector, the so-called load vector, with entries

$$b_i = \int_a^b \left(f \phi_i - \alpha \phi_0' \phi_i' - \beta \phi_N' \phi_i' \right) dx, \quad i = 1, 2, \ldots, N-1. \tag{22}$$

Computer Implementation: The explicit expression for a hat function $\phi_i(x)$ is given by

$$\phi_i(x) = \begin{cases} 0, & a \le x \le x_{i-1}, \\ \dfrac{x - x_{i-1}}{h_i}, & x_{i-1} < x \le x_i, \\ \dfrac{x_{i+1} - x}{h_{i+1}}, & x_i < x \le x_{i+1}, \\ 0, & x_{i+1} < x \le b, \end{cases}, \quad i = 1, 2, \ldots, N-1,$$

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{h_1}, & x_0 < x \le x_1, \\ 0, & x_1 < x \le b, \end{cases}, \quad \phi_N(x) = \begin{cases} 0, & x_0 < x \le x_{N-1}, \\ \dfrac{x - x_{N-1}}{h_N}, & x_{N-1} < x \le b. \end{cases}$$

For simplicity we assume the partition is uniform so that $h_i = h$ for $i = 1, 2, \ldots, N$. Hence the derivative $\phi_i'(x)$ is either $-\frac{1}{h}$, $\frac{1}{h}$, or 0 depending on the interval.

It is straightforward to calculate the entries of the stiffness matrix. For $|i-j| > 1$, we have $a_{i,j} = 0$, since ϕ_i and ϕ_j lack overlapping support. However, if $i = j$, then

$$a_{i,j} = \int_a^b (\phi_i')^2 dx = \int_{x_{i-1}}^{x_i} \left(\frac{1}{h}\right)^2 dx + \int_{x_i}^{x_{i+1}} \left(-\frac{1}{h}\right)^2 dx = \frac{2}{h}, \quad i,j = 1, 2, \ldots, N-1,$$

where we have used that $x_i - x_{i-1} = x_{i+1} - x_i = h$. Furthermore, if $j = i+1$, then

$$a_{i,i+1} = \int_a^b \phi_i' \phi_{i+1}' dx = \int_{x_i}^{x_{i+1}} \left(-\frac{1}{h}\right)\left(\frac{1}{h}\right) dx = -\frac{1}{h}, \quad i,j = 1, 2, \ldots, N-2.$$

Changing i to $i-1$ we also have

$$a_{i-1,i} = \int_a^b \phi_{i-1}' \phi_i dx = \int_{x_{i-1}}^{x_i} \left(\frac{1}{h}\right)\left(-\frac{1}{h}\right) dx = -\frac{1}{h}, \quad i,j = 2, 3, \ldots, N-1.$$

Thus the stiffness matrix is

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

The entries b_i of the load vector must often be evaluated using quadrature, since they involve the function f which can be hard to integrate analytically. For example, using the trapezoidal rule one obtains the approximate load vector entries

$$b_1 = \int_a^b (f\phi_1 - \alpha\phi_0'\phi_1' - \beta\phi_N'\phi_1')dx = \int_{x_0}^{x_1} \left(f\phi_1 - \alpha\left(-\frac{1}{h}\right)\left(\frac{1}{h}\right)\right)dx + \int_{x_1}^{x_2} f\phi_1$$

$$= \frac{\alpha}{h} + \int_{x_0}^{x_2} f\phi_1 \approx \frac{\alpha}{h} + hf(x_1),$$

$$b_i = \int_a^b (f\phi_i - \alpha\phi_0'\phi_i' - \beta\phi_N'\phi_i')dx = \int_{x_{i-1}}^{x_{i+1}} f\phi_i dx \approx hf(x_i), \quad i = 2, \ldots, N-2,$$

$$b_{N-1} = \int_a^b (f\phi_{N-1}' - \alpha\phi_0'\phi_{N-1} - \beta\phi_N'\phi_{N-1}')dx$$

$$= \int_{x_{N-2}}^{x_{N-1}} f\phi_{N-1} dx + \int_{x_{N-1}}^{x_N} \left(f\phi_{N-1} - \beta\left(\frac{1}{h}\right)\left(-\frac{1}{h}\right)\right)dx = \int_{x_{N-2}}^{x_N} f\phi_{N-1} dx + \frac{\beta}{h} \approx hf(x_{N-1}) + \frac{\beta}{h}.$$

Assembly: We rewrite (20), (21), (22) as

$$\frac{1}{h}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & 0 \\ 0 & -1 & 2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_{N-1} \end{bmatrix} = \begin{bmatrix} hf(x_1) + \frac{\alpha}{h} \\ hf(x_2) \\ hf(x_3) \\ \vdots \\ hf(x_{N-1}) + \frac{\beta}{h} \end{bmatrix}$$

We note that $u_h(a) = \alpha = u(a)$ and $u_h(b) = \beta = u(b)$. Therefore, we see that the system matrix A remains the same, and only the first and last entries of the load vector **b** need to be modified because of the definition of the basis functions $\{\phi_0, \ldots, \phi_N\}$. An alternative approach is to use all the basis functions $\{\phi_0, \ldots, \phi_N\}$ to form a larger system of equation, *i.e.*, and $(N+1) \times (N+1)$ system. The procedure for inserting the boundary conditions into the system equation is: enter zeros in the first and $(N+1)$-th rows of the system matrix A except for unity in the main diagonal positions of these two rows, and enter α and β in the first and $(N+1)$-th rows of the vector **b**, respectively.

General boundary conditions: Let us consider the following two-point BVP: find $u \in C^2(a,b)$ such that

$$-u'' = f(x), \quad x \in [a,b], \quad u(a) = \alpha, \quad \gamma u(b) + u'(b) = \beta, \tag{23}$$

where α, β and γ are given numbers and $f \in C(a,b)$ is a given function. The boundary condition at $x = b$ is called a Robin boundary condition (combination and u and u' is prescribed at $x = b$). In this case, the admissible function space is modified to

$$H_0^1 = \left\{ v : \|v\|^2 + \|v'\|^2 < \infty, \;\; v(a) = 0 \right\}.$$

Multiplying (23) by a function $v \in H_0^1$ and integrating by parts gives

$$\begin{aligned} \int_a^b f v dx = \int_a^b -u'' v dx &= -u'(b)v(b) + u'(a)v(a) + \int_a^b u'v' dx \\ &= -(\beta - \gamma u(b))v(b) + u'(a)v(a) + \int_a^b u'v' dx. \end{aligned}$$

Since $v(a) = 0$, we are left with

$$\int_a^b u'v' dx + \gamma u(b)v(b) = \int_a^b f v dx + \beta v(b).$$

Hence, the weak or variational form of (23) reads: Given $u(a) = \alpha$, find the approximate solution $u \in H_0^1$, such that

$$\int_a^b u'v' dx + \gamma u(b)v(b) = \int_a^b f v dx + \beta v(b), \quad \forall \;\; v \in H_0^1. \tag{24}$$

The FE space V_h^1 is now the set of all continuous piecewise linear functions which vanish at the end point a. The FE approximation of (24) thus reads: Find the piecewise linear approximation u_h to the solution u satisfies

$$\int_a^b u_h'v' dx + \gamma u_h(b)v(b) = \int_a^b f v dx + \beta v(b), \quad \forall \;\; v \in V_h^1, \tag{25}$$

with $u_h(a) = \alpha$. As before, (25) can be formulated in matrix form.

2.4 Model problem with coefficient and general Robin BCs

Let us consider the following two-point BVP: find $u \in C^2(a,b)$ such that

$$\begin{aligned} &-(p(x)u')' = f(x), \quad x \in I = [a,b], \quad p(a)u'(a) = \kappa_0(u(a) - \alpha), \\ &p(b)u'(b) = \kappa_1(u(b) - \beta), \end{aligned} \tag{26}$$

where $p = p(x)$ with $p(x) \geq p_0 > 0, f \in L^2(I)$, κ_0, $\kappa_1 \geq 0$, and α, β are given numbers. Let

$$V = \left\{ v \in C^0(I) : \|v\|^2 + \|v'\|^2 < \infty \right\}.$$

Multiplying (26) by a function $v \in V$ and integrating by parts gives

$$\begin{aligned} \int_a^b fv dx &= \int_a^b -(pu')'v dx = \int_a^b pu'v' dx - p(b)u'(b)v(b) + p(a)u'(a)v(a) \\ &= \int_a^b pu'v' dx - \kappa_1(u(b) - \beta)v(b) + \kappa_0(u(a) - \alpha)v(a). \end{aligned}$$

We gather all u-independent terms on the left and obtain

$$\int_a^b pu'v' dx - \kappa_1 u(b)v(b) + \kappa_0 u(a)v(a) = \int_a^b fv dx - \kappa_1 \beta v(b) + \kappa_0 \alpha v(a), \quad \forall\ v \in V.$$

The FE method consists of finding $u_h \in V_h = \left\{ v \in C^0(a,b) |\ \ v|_{I_i} \in P^1(I_i) \right\}$ such that

$$\int_a^b pu_h'v' dx - \kappa_1 u_h(b)v(b) + \kappa_0 u_h(a)v(a) = \int_a^b fv dx - \kappa_1 \beta v(b) + \kappa_0 \alpha v(a), \quad \forall v \in V_h. \tag{27}$$

Implementation: We need to assemble a stiffness matrix A and a load vector b. Substituting $u_h = \sum_{i=0}^N c_i \phi_i$ into (27) and taking $v = \phi_j$ for $j = 0, 1, \ldots, N$ yields

$$\sum_{i=0}^N \int_a^b p\phi_i'\phi_j' dx - \kappa_1 \phi_i(b)\phi_j(b) + \kappa_0 \phi_i(a)\phi_j(a) = \int_a^b f\phi_j dx - \kappa_1 \beta \phi_j(b) + \kappa_0 \alpha \phi_j(a),$$
$$\forall j = 0, 1, \ldots, N.$$

which is a $(N+1) \times (N+1)$ system of equations for c_i. In matrix form we write $A\mathbf{c} = \mathbf{b}$, where $\mathbf{c} = [c_0, \ldots, c_N]^t$ is a $(N+1)$ vector containing the unknown coefficients c_i, $i = 0, 1, \ldots, N$, A is a $(N+1) \times (N+1)$ matrix with entries

$$a_{i,j} = \int_a^b p\phi_i'\phi_j' dx - \kappa_1 \phi_i(b)\phi_j(b) + \kappa_0 \phi_i(a)\phi_j(a), \quad i,j = 0, 1, \ldots, N,$$

and $\mathbf{b}$ is a $(N+1)$ vector with entries

$$b_j = \int_a^b f\phi_j dx - \kappa_1 \beta \phi_j(b) + \kappa_0 \alpha \phi_j(a), \quad j = 0, 1, \ldots, N.$$

Let for simplification $p = 1$. Then the matrix A and the vector $\mathbf{b}$ (when using the trapezoidal rule) are given by

$$A = \begin{bmatrix} \kappa_0 + \frac{1}{h_1} & -\frac{1}{h_1} & 0 & \cdots & 0 \\ -\frac{1}{h_1} & \frac{1}{h_1} + \frac{1}{h_2} & -\frac{1}{h_2} & \ddots & 0 \\ 0 & -\frac{1}{h_2} & \ddots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\frac{1}{h_N} \\ 0 & \cdots & 0 & -\frac{1}{h_N} & \frac{1}{h_N} - \kappa_1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} \frac{h_1}{2} f_0 + \kappa_0 \alpha \\ \frac{h_1 + h_2}{2} f_1 \\ \vdots \\ \frac{h_{N-1} + h_N}{2} f_{N-1} \\ \frac{h_N}{2} f_N - \kappa_1 \beta \end{bmatrix}.$$

2.5 The FE method using Lagrange $\mathbb{P}_2$ elements

Let $a = x_0 < x_1 < \cdots < x_N = b$ be a regular partition of the interval $[a, b]$. Suppose that the length of $I_i = [x_{i-1}, x_i]$ is $h_i = x_i - x_{i-1}$. Let $P^k = \left\{ p(x) = \sum_{j=0}^{k} c_j x^j, c_j \in \mathbb{R} \right\}$ denotes the vector space of polynomials in one variable and of degree less than or equal to k. The FE method for Lagrange P^2 elements involves the discrete space:

$$V_h^2 = \{ v(x) \in C^0[a, b], \quad v|_{I_i} \in P^2(I_i), \quad i = 1, \ldots, N \},$$

and its subspace $V_{0,h}^2 = \left\{ v \in V_h^2 | \ v(a) = v(b) = 0 \right\}$. These spaces are composed of continuous, piecewise parabolic functions (polynomials of degree less than or equal to 2). The P^2 FE method consists in applying the internal variational approximation approach to these spaces.

Lemma 2.2 *The space V_h^2 is a subspace of $H^1[a, b]$ of dimension $2N + 1$. Every function $v_h \in V_h^2$ is uniquely defined by its values at the mesh vertices x_j, $j = 0, 1, \ldots, N$ and at the midpoints $x_{j+\frac{1}{2}} = \frac{x_j + x_{j+1}}{2} = x_j + \frac{h_{j+1}}{2}$, $j = 0, 1, \ldots, N-1$, where $h_{j+1} = x_{j+1} - x_j$:*

$$v_h(x) = \sum_{j=0}^{N} v_h(x_j) \phi_j(x) + \sum_{j=0}^{N-1} v_h\left(x_{j+\frac{1}{2}}\right) \phi_{j+\frac{1}{2}}(x), \quad \forall \ x \in [a, b],$$

where $\left\{ \phi_j \right\}_{j=0}^{N}$ is the basis of the shape functions ϕ_j defined as:

$$\phi_j(x) = \phi\left(\frac{x - x_j}{h_{j+1}} \right), \quad j = 0, 1, \ldots, N, \quad \phi_{j+\frac{1}{2}}(x) = \psi\left(\frac{x - x_{j+\frac{1}{2}}}{h_{j+1}} \right), \quad j = 0, 1, \ldots, N-1,$$

with

$$\phi(\xi) = \begin{cases} (1 + \xi)(1 + 2\xi), & \xi \in [-1, 0], \\ (1 - \xi)(1 - 2\xi), & \xi \in [0, 1], \\ 0, & |\xi| > 1, \end{cases} \qquad \psi(\xi) = \begin{cases} 1 - 4\xi^2, & |\xi| \le \frac{1}{2}, \\ 0, & |\xi| > \frac{1}{2}, \end{cases} \tag{28}$$

Figure 3 shows the global shape functions for the space V_h^2 and the three quadratic Lagrange P^2 shape functions on the reference interval $[-1, 1]$.

DOI: http://dx.doi.org/10.5772/intechopen.95423

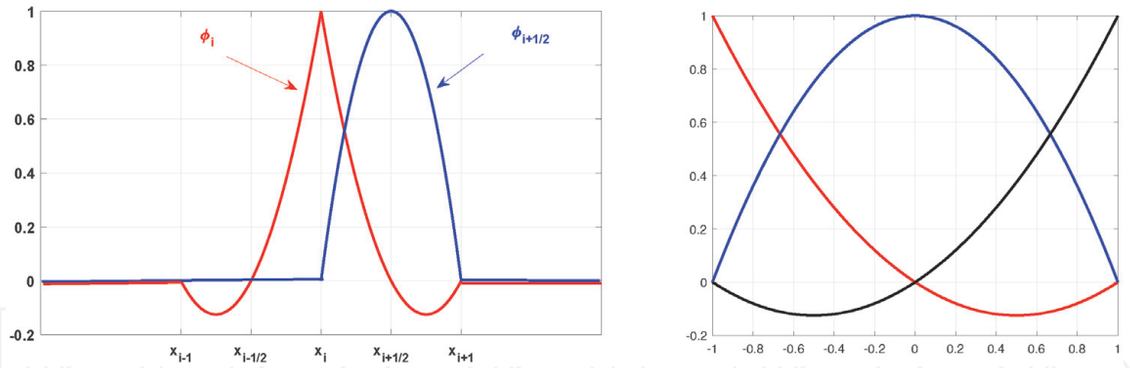

Figure 3.
(left) global shape functions for the space V_h^2. (right) the three quadratic Lagrange P^2 shape functions on the reference interval $[-1, 1]$.

Remark 2.5 *Notice that we have:*

$$\phi_j(x_j) = \delta_{ij}, \quad \phi_j\left(x_{j+\frac{1}{2}}\right) = 0, \quad \phi_{j+\frac{1}{2}}(x_j) = 0, \quad \phi_{j+\frac{1}{2}}\left(x_{j+\frac{1}{2}}\right) = \delta_{ij}.$$

Corollary 2.1 *The space $V_{0,h}^2$ is a subspace of $H_0^1[a,b]$ of dimension $2N-1$ and every function $v_h \in V_{0,h}^2$ is uniquely defined by its values at the mesh vertices x_j, $j = 1, 2, \ldots, N-1$ and at the midpoints $x_{j+\frac{1}{2}}$, $j = 0, 1, \ldots, N-1$:*

$$v_h(x) = \sum_{j=1}^{N-1} v_h(x_j)\phi_j(x) + \sum_{j=0}^{N-1} v_h\left(x_{j+\frac{1}{2}}\right)\phi_{j+\frac{1}{2}}(x), \quad \forall\ x \in [a,b],$$

where $\left\{\phi_j\right\}_{j=0}^{N}$ is the basis of the shape functions ϕ_j defined as:

$$\phi_j(x) = \phi\left(\frac{x - x_j}{h_{j+1}}\right), \quad j = 0, 1, \ldots, N, \quad \phi_{j+\frac{1}{2}}(x) = \psi\left(\frac{x - x_{j+\frac{1}{2}}}{h_{j+1}}\right), \quad j = 0, 1, \ldots, N-1,$$

with $\phi(\xi)$ and $\psi(\xi)$ are defined by (28).

2.5.1 Homogeneous boundary conditions

The variational formulation of the internal approximation of the Dirichlet BVP (3) consists now in finding $u_h \in V_{0,h}^2$, such that:

$$\int_a^b u_h' v' dx + \int_a^b q u_h v dx = \int_a^b f v dx, \quad \forall\ v \in V_{h,0}^2.$$

Here, it is convenient to introduce the notation $x_{\frac{j}{2}}$, $j = 1, \ldots, 2N-1$ for the mesh points and $\phi_{\frac{j}{2}}$, $j = 1, \ldots, 2N-1$ for the basis of $V_{0,h}^2$. Using these notations, we have:

$$u_h = \sum_{j=1}^{2N-1} c_{\frac{j}{2}} \phi_{\frac{j}{2}}(x),$$

where $c_{\frac{j}{2}} = u_h\left(x_{\frac{j}{2}}\right) \approx u\left(x_{\frac{j}{2}}\right)$ are the unknowns coefficients. This formulation leads to solve in $\mathbb{R}^{2N-1}$ a linear system:

$$A\mathbf{c} = \mathbf{b},$$

where $\mathbf{c} = \left[c_{\frac{1}{2}}, c_1, \ldots, c_{N-\frac{1}{2}}\right]^t \in \mathbb{R}^{2N-1}$ is the unknown vector containing the coefficients $c_{\frac{j}{2}}$, $j = 1, 2, \ldots, 2N-1$, A is an $(2N-1)\times(2N-1)$ matrix with entries

$$a_{ij} = \int_a^b \left(\phi_{\frac{i}{2}}{}'\phi_{\frac{j}{2}}{}' + q\phi_{\frac{i}{2}}\phi_{\frac{j}{2}}\right)dx, \quad i,j = 1, 2, \ldots, 2N-1,$$

and load vector $\mathbf{b} \in \mathbb{R}^{2N-1}$ has entries

$$b_{\frac{i}{2}} = \int_a^b f\phi_{\frac{i}{2}}dx, \quad i = 1, 2, \ldots, 2N-1.$$

Since the shape functions ϕ_i have a small support, the matrix A is mostly composed of zeros. However, the main difference with the Lagrange P^1 FE method, the matrix A is no longer a tridiagonal matrix.

Computer Implementation: The coefficients of the matrix A can be computed more easily by considering the following change of variables, for $\xi \in [-1, 1]$:

$$x = \frac{x_j + x_{j-1}}{2} + \frac{x_j - x_{j-1}}{2}\xi = x_{j-\frac{1}{2}} + \frac{x_j - x_{j-1}}{2}\xi, \quad \forall\ x \in \left[x_{j-1}, x_j\right],$$
$$j = 1, 2, \ldots, N.$$

Hence, the shape functions can be reduced to only three basic shape functions (**Figure 3**):

$$\hat{\phi}_{-1}(\xi) = \frac{\xi(\xi-1)}{2}, \quad \hat{\phi}_0(\xi) = (1-\xi)(1+\xi), \quad \hat{\phi}_1(\xi) = \frac{\xi(\xi+1)}{2}.$$

Their respective derivatives are

$$\frac{d\hat{\phi}_{-1}(\xi)}{d\xi} = \frac{2\xi-1}{2}, \quad \frac{d\hat{\phi}_0(\xi)}{d\xi} = -2\xi, \quad \frac{d\hat{\phi}_1(\xi)}{d\xi} = \frac{2\xi+1}{2}.$$

This approach consists in considering all computations on an interval $I_i = [x_{i-1}, x_i]$ on the reference interval $[-1, 1]$. Thus, we have:

$$\frac{d\phi_i(x)}{dx} = \frac{d\phi_i\left(x_{i-1/2} + \frac{x_i - x_{i-1}}{2}\xi\right)}{d\xi}\frac{d\xi}{dx} = \frac{2}{x_i - x_{i-1}}\frac{d\hat{\phi}_k(\xi)}{d\xi} = \frac{2}{h_i}\frac{d\hat{\phi}_k(\xi)}{d\xi}.$$

In this case, the elementary contributions of the element I_i to the stiffness matrix and to the mass matrix are given by the 3×3 matrices K^{I_i} and M^{I_i}:

$$K^{I_i} = \int_{I_i}\begin{bmatrix} \phi'_{i-1}\phi'_{i-1} & \phi'_{i-1}\phi'_{i-\frac{1}{2}} & \phi'_{i-1}\phi'_i \\ \phi'_{i-\frac{1}{2}}\phi'_{i-1} & \phi'_{i-\frac{1}{2}}\phi'_{i-\frac{1}{2}} & \phi'_{i-\frac{1}{2}}\phi'_i \\ \phi'_i\phi'_{i-1} & \phi'_i\phi'_{i-\frac{1}{2}} & \phi'_i\phi'_i \end{bmatrix} dx = \frac{2}{h_i}\int_{-1}^{1}\begin{bmatrix} \hat{\phi}'_{-1}\hat{\phi}'_{-1} & \hat{\phi}'_{-1}\hat{\phi}'_0 & \hat{\phi}'_{-1}\hat{\phi}'_1 \\ \hat{\phi}'_0\hat{\phi}'_{-1} & \hat{\phi}'_0\hat{\phi}'_0 & \hat{\phi}'_0\hat{\phi}'_1 \\ \hat{\phi}'_1\hat{\phi}'_{-1} & \hat{\phi}'_1\hat{\phi}'_0 & \hat{\phi}'_1\hat{\phi}'_1 \end{bmatrix} d\xi$$
$$= \frac{1}{3h_i}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix},$$

$$M^{I_i} = \int_{I_i} \begin{bmatrix} \phi_{i-1}\phi_{i-1} & \phi_{i-1}\phi_{i-\frac{1}{2}} & \phi_{i-1}\phi_i \\ \phi_{i-\frac{1}{2}}\phi_{i-1} & \phi_{i-\frac{1}{2}}\phi_{i-\frac{1}{2}} & \phi_{i-\frac{1}{2}}\phi_i \\ \phi_i\phi_{i-1} & \phi_i\phi_{i-\frac{1}{2}} & \phi_i\phi_i \end{bmatrix} dx = \frac{h_i}{2}\int_{-1}^{1} \begin{bmatrix} \hat{\phi}_{-1}\hat{\phi}_{-1} & \hat{\phi}_{-1}\hat{\phi}_0 & \hat{\phi}_{-1}\hat{\phi}_1 \\ \hat{\phi}_0\hat{\phi}_{-1} & \hat{\phi}_0\hat{\phi}_0 & \hat{\phi}_0\hat{\phi}_1 \\ \hat{\phi}_1\hat{\phi}_{-1} & \hat{\phi}_{-1}\hat{\phi}_0 & \hat{\phi}_1\hat{\phi}_1 \end{bmatrix} d\xi$$
$$= \frac{h_i}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}.$$

Coefficients of the right-hand side b: Usually, the function f is only known by its values at the mesh points $x_{\frac{i}{2}}$, $i = 0, 1, \ldots, 2N$ and thus, we use the decomposition of f in the basis of shape functions $\phi_{\frac{i}{2}}$, $i = 0, 1, \ldots, 2N$ as $f(x) = \sum_{j=0}^{2N} f\left(x_{\frac{j}{2}}\right)\phi_{\frac{j}{2}}$. Each component $b_{\frac{i}{2}}$ of the right-hand side vector is obtained as $b_{\frac{i}{2}} = \sum_{k=1}^{N}\int_{x_{k-1}}^{x_k} f\phi_{\frac{i}{2}}dx$. Using the previous decomposition of f, we obtain:

$$b_{\frac{i}{2}} = \sum_{k=1}^{N}\int_{x_{k-1}}^{x_k}\sum_{j=0}^{2N} f\left(x_{\frac{j}{2}}\right)\phi_{\frac{j}{2}}\phi_{\frac{i}{2}}dx = \sum_{j=0}^{2N} f\left(x_{\frac{j}{2}}\right)\left(\sum_{k=1}^{N}\int_{x_{k-1}}^{x_k}\phi_{\frac{i}{2}}\phi_{\frac{j}{2}}dx\right).$$

Thus, the problem is reduced to computing the integrals $\int_{x_{k-1}}^{x_k}\phi_{\frac{i}{2}}\phi_{\frac{j}{2}}dx$. It is easy to see that we obtain expressions very similar to that of the mass matrix. More precisely, the element $I_i = [x_{i-1}, x_i]$ will contribute to only three components of indices $i-1$, $i-\frac{1}{2}$ and i as:

$$\mathbf{b}^{I_i} = \frac{h_i}{30}\begin{bmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{bmatrix}\begin{bmatrix} f(x_{i-1}) \\ f\left(x_{i-\frac{1}{2}}\right) \\ f(x_i). \end{bmatrix}.$$

2.5.2 Nonhomogeneous boundary conditions

Consider the following two-point BVP: find $u \in C^2(a,b)$ such that

$$-u'' + q(x)u = f(x), \quad x \in [a,b], \quad u(a) = \alpha, \quad u(b) = \beta, \tag{29}$$

where α and β are given constants and $f \in C(a,b)$ is a given function.

Multiplying (29) by a function $v \in H_0^1 = \left\{v : \|v\|^2 + \|v'\|^2 < \infty,\ v(a) = v(b) = 0\right\}$ and integrating by parts gives

$$\int_a^b fvdx = \int_a^b (-u'' + qu)vdx = -u'(b)v(b) + u'(a)v(a) + \int_a^b (u'v' + quv)dx = \int_a^b u'v'dx.$$

Hence, the weak or variational form of (29) reads: Given $u(a) = \alpha$, $u(b) = \beta$, find $u \in H^1 = \left\{v : \|v\|^2 + \|v'\|^2 < \infty\right\}$, such that

$$\int_a^b (u'v' + quv)dx = \int_a^b fvdx, \quad \forall\ v \in H_0^1.$$

Let V_h^2 and $V_{h,0}^2$, respectively, be the space of all continuous piecewise quadratic functions and the space of all continuous piecewise quadratic functions which

vanish at the end points a and b, on a uniform partition $a = x_0 < x_1 < \cdots < x_N = b$ of the interval $[a, b]$.

The FE method scheme consists of finding $u_h \in V_h^2$, such that:

$$\int_a^b u_h' v' dx + \int_a^b q u_h v dx = \int_a^b f v dx, \quad \forall\ v \in V_{h,0}^2.$$

Introduce the notation $x_{\frac{j}{2}},\ j = 0, 1, \ldots, 2N-1, 2N$ for the mesh points and $\phi_{\frac{j}{2}},\ j = 0, 1, \ldots, 2N-1, 2N$ for the basis of V_h^2 and $\phi_{\frac{j}{2}},\ j = 1, \ldots, 2N-1$ for the basis of $V_{0,h}^2$. Using these notations, we have:

$$u_h = \sum_{j=0}^{2N} c_{\frac{j}{2}} \phi_{\frac{j}{2}}(x),$$

where $c_{\frac{j}{2}} = u_h\left(x_{\frac{j}{2}}\right) \approx u\left(x_{\frac{j}{2}}\right)$ are the unknowns coefficients. We note that $c_0 = u_h(x_0) = \alpha$ and $c_{2N} = u_h(x_N) = \beta$. This formulation leads to solve in $\mathbb{R}^{2N-1}$ a linear system:

$$A\mathbf{c} = \mathbf{b},$$

where $\mathbf{c} = \left[c_{\frac{1}{2}}, c_1, \ldots, c_{N-\frac{1}{2}}\right]^t \in \mathbb{R}^{2N-1}$ is the unknown vector containing the coefficients $c_{\frac{j}{2}},\ j = 1, 2, \ldots, 2N-1$, A is an $(2N-1) \times (2N-1)$ matrix with entries

$$a_{ij} = \int_a^b \left(\phi_{\frac{i}{2}}' \phi_{\frac{j}{2}}' + q \phi_{\frac{i}{2}} \phi_{\frac{j}{2}}\right) dx, \quad i, j = 1, 2, \ldots, 2N-1,$$

and the load vector $\mathbf{b} \in \mathbb{R}^{2N-1}$ has entries

$$b_{\frac{i}{2}} = \int_a^b f \phi_{\frac{i}{2}} dx - \alpha \int_a^b \left(\phi_{\frac{i}{2}}' \phi_0' + q \phi_{\frac{i}{2}} \phi_0\right) dx - \beta \int_a^b \left(\phi_{\frac{i}{2}}' \phi_N' + q \phi_{\frac{i}{2}} \phi_N\right) dx, \quad i = 1, 2, \ldots, 2N-1.$$

Clearly, the only extra terms are given in the vector with entries

$$\tilde{b}_{\frac{i}{2}} = -\alpha \int_a^b \left(\phi_{\frac{i}{2}}' \phi_0' + q \phi_{\frac{i}{2}} \phi_0\right) dx - \beta \int_a^b \left(\phi_{\frac{i}{2}}' \phi_N' + q \phi_{\frac{i}{2}} \phi_N\right) dx, \quad i = 1, 2, \ldots, 2N-1.$$

Suppose $q = 0$ then for $N \geq 2$, we have

$$\tilde{b}_{\frac{1}{2}} = -\alpha \int_a^b \phi_{\frac{1}{2}}' \phi_0' dx - \beta \int_a^b \phi_{\frac{1}{2}}' \phi_N' dx = -\alpha \int_{x_0}^{x_1} \phi_{\frac{1}{2}}' \phi_0' = \frac{8\alpha}{3h_1},$$

$$\tilde{b}_1 = -\alpha \int_a^b \phi_1' \phi_0' dx - \beta \int_a^b \phi_1' \phi_N' dx = -\alpha \int_{x_0}^{x_1} \phi_1' \phi_0' = -\frac{\alpha}{3h_1},$$

$$\tilde{b}_{\frac{i}{2}} = -\alpha \int_a^b \phi_{\frac{i}{2}}' \phi_0' dx - \beta \int_a^b \phi_{\frac{i}{2}}' \phi_N' dx = 0, \quad i = 3, \ldots, 2N-3,$$

$$\tilde{b}_{N-1} = -\alpha \int_a^b \phi_{N-1}' \phi_0' dx - \beta \int_a^b \phi_{N-1}' \phi_N' dx = -\beta \int_{x_{N-1}}^{x_N} \phi_{N-1}' \phi_N' dx = -\frac{\beta}{3h_1},$$

$$\tilde{b}_{N-\frac{1}{2}} = -\alpha \int_a^b \phi_{N-\frac{1}{2}}' \phi_0' dx - \beta \int_a^b \phi_{N-\frac{1}{2}}' \phi_N' dx = -\beta \int_{x_{N-1}}^{x_N} \phi_{N-\frac{1}{2}}' \phi_N' dx = \frac{8\beta}{3h_1}.$$

3. The FE for elliptic PDEs

Here, we apply the FE method for two-dimensional elliptic problem: Find u such that

$$-\nabla \cdot (a\nabla u) + bu = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad a\nabla u \cdot \mathbf{n} = \kappa(g - u), \quad \text{on } \partial\Omega, \tag{30}$$

where $a > 0$, $b \geq 0$, $\kappa \geq 0$, $f \in L^2(\Omega)$ and $g \in C^0(\partial\Omega)$.

3.1 Meshes

Let $\Omega \subset \mathbb{R}^2$ bounded with $\partial\Omega$ assumed to be polygonal. A triangulation $\mathcal{T}_h$ of Ω is a set of triangles T such that $\Omega = \bigcup_{T \in \mathcal{T}_h} T$, and two triangles intersect by either a common triangle edge, or a corner, or nothing. Corners will be referred to as nodes. We let $h_T = diam(T)$ the length or the largest edge.

Let $\mathcal{T}_h$ have N nodes and M triangles. The data is stored in two matrices. The matrix $P \in \mathbb{R}^{2\times N}$ describes the nodes $((x_1, y_1), \ldots, (x_N, y_N))$ and the matrix $K \in \mathbb{R}^{3\times M}$ describes the triangles, *i.e.*, it describes which nodes (numerated from 1 to N) form a triangle T and how it is orientated:

$$P = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ y_1 & y_2 & \cdots & y_N \end{bmatrix}, \quad K = \begin{bmatrix} n_1^\alpha & n_2^\alpha & \cdots & n_M^\alpha \\ n_1^\beta & n_2^\beta & \cdots & n_M^\beta \\ n_1^\gamma & n_2^\gamma & \cdots & n_M^\gamma \end{bmatrix}.$$

This means that triangle T_i is formed by the nodes n_i^α, n_i^β, and n_i^γ (enumeration in counter-clockwise direction).

The Delaunay algorithm determine a triangulation with the given points as triangle nodes. Delaunay triangulations are optimal in the sense that the angles of all triangles are maximal.

Matlab has a built in toolbox called PDE Toolbox and includes a mesh generation algorithm.

3.2 Piecewise polynomial spaces

Let T be a triangle with nodes $N_1 = (x_1, y_1)$, $N_2 = (x_2, y_2)$, and $N_3 = (x_3, y_3)$. We define

$$P^1(T) = \{v \in C^0(T) |\ v(x, y) = c_1 + c_2 x + c_3 y,\ \ c_1, c_2, c_3 \in \mathbb{R}\}.$$

Now let $v_i = v(N_i)$ for $i = 1, 2, 3$. Note that $v \in P^1(T)$ is determined by $\{v_i\}_{i=1}^3$. Given v_i we compute c_i by

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}.$$

This is solvable due to

$$\det\begin{bmatrix}1 & x_1 & y_1\\ 1 & x_2 & y_2\\ 1 & x_3 & y_3\end{bmatrix} = 2|T| \neq 0, \quad \begin{bmatrix}1 & x_1 & y_1\\ 1 & x_2 & y_2\\ 1 & x_3 & y_3\end{bmatrix}^{-1}$$
$$= \frac{1}{2|T|}\begin{bmatrix}x_2y_3 - x_3y_2 & x_3y_1 - x_1y_3 & x_1y_2 - x_2y_1\\ y_2 - y_3 & y_3 - y_1 & y_1 - y_2\\ x_3 - x_2 & x_1 - x_3 & x_3 - x_1\end{bmatrix},$$

where $|T| = \frac{1}{2}\left(x_2y_3 - x_3y_2 - x_1y_3 + x_3y_1 + x_1y_2 - x_2y_1\right)$, which is $\pm$ the area of the triangle T.

Let $\lambda_j \in P^1(T)$ be given by the nodal values $\lambda_j(N_i) = \delta_{ij}$, where δ_{ij} is the Kronecker symbol. This gives us $v(x,y) = \alpha_1\lambda_1(x,y) + \alpha_2\lambda_2(x,y) + \alpha_3\lambda_3(x,y)$, where $\alpha_i = v(N_i)$ for $i = 1,2,3$. We can compute $\lambda_i(x,y)$ as follows: Let $\lambda_i(x,y) = a_i + b_ix + c_iy$. Using $\lambda_j(N_i) = \delta_{ij}$, we get

$$\begin{bmatrix}1 & x_1 & y_1\\ 1 & x_2 & y_2\\ 1 & x_3 & y_3\end{bmatrix}\begin{bmatrix}a_1\\ b_1\\ c_1\end{bmatrix} = \begin{bmatrix}1\\ 0\\ 0\end{bmatrix}, \quad \begin{bmatrix}1 & x_1 & y_1\\ 1 & x_2 & y_2\\ 1 & x_3 & y_3\end{bmatrix}\begin{bmatrix}a_2\\ b_2\\ c_2\end{bmatrix} = \begin{bmatrix}0\\ 1\\ 0\end{bmatrix}, \quad \begin{bmatrix}1 & x_1 & y_1\\ 1 & x_2 & y_2\\ 1 & x_3 & y_3\end{bmatrix}\begin{bmatrix}a_3\\ b_3\\ c_3\end{bmatrix} = \begin{bmatrix}0\\ 0\\ 1\end{bmatrix}.$$

Solving the systems, we get

$$\lambda_1(x,y) = \frac{1}{2|T|}\left(x_2y_3 - x_3y_2 + (y_2 - y_3)x + (x_3 - x_2)y\right),$$
$$\lambda_2(x,y) = \frac{1}{2|T|}\left(x_3y_1 - x_1y_3 + (y_3 - y_1)x + (x_1 - x_3)y\right),$$
$$\lambda_3(x,y) = \frac{1}{2|T|}\left(x_1y_2 - x_2y_1 + (y_1 - y_2)x + (x_3 - x_1)y\right).$$

Let $\mathcal{T}_h$ be a triangulation of Ω, then we let

$$V_h = \left\{v \in C(\Omega) |\ v|_T \in P^1(T),\ \forall\ T \in \mathcal{T}_h\right\}.$$

Functions in V_h are piecewise linear and continuous. We know that $v \in V_h$ is uniquely determined by $\{v(N_i),\ i = 1,2,\ldots,N\}$. We let $\phi_j(N_i) = \delta_{ij}$ and let $\left\{\phi_j,\ j = 1,2,\ldots,N\right\} \subset V_h$ be a basis for V_h (hat functions), *i.e.*,

$$v(x,y) = \sum_{i=1}^{N} \alpha_i\phi_i(x,y), \quad \alpha_i = v(N_i), \quad i = 1,2,\ldots,N.$$

3.3 Interpolation

Given $u \in C(T)$ on a single triangle with nodes $N_i = (x_i, y_i)$, $i = 1,2,3$, we let

$$\pi u(x,y) = \sum_{i=1}^{3} u(N_i)\phi_i(x,y),$$

in particular $\pi u(N_i) = u(N_i)$, $i = 1,2,\ldots,N$. We want to estimate the interpolation error $u - \pi u$. Let

$$\|u\|^2_{L^2(\Omega)} = \int_\Omega |u(x)|^2 dxdy, \quad \|Du\|^2_{L^2(\Omega)} = \|u_x\|^2_{L^2(\Omega)} + \|u_y\|^2_{L^2(\Omega)},$$
$$\|D^2u\|^2_{L^2(\Omega)} = \|u_{xx}\|^2_{L^2(\Omega)} + 2\|u_{xy}\|^2_{L^2(\Omega)} + \|u_{yy}\|^2_{L^2(\Omega)}.$$

Theorem 3.1 *Suppose that $u \in C^2(T)$. Then the following hold*

$$\|u - \pi u\|_{L^2(T)} \le Ch_T^2 \|D^2u\|_{L^2(T)}, \quad \|D(u - \pi u)\|_{L^2(T)} \le Ch_T \|D^2u\|_{L^2(T)},$$

where C is a generic constant independent of h_T and u, but it depends on the ratio between smallest and largest interior angle of the triangle T.

Now, we consider the piecewise continuous interpolant $\pi u = \sum_{i=1}^N u(N_i)\phi_i$.

Theorem 3.2 *Suppose that $u \in C^2(T)$ for all $T \in \mathcal{T}_h$. Then the following hold*

$$\|u - \pi u\|^2_{L^2(\Omega)} \le C \sum_{T \in \mathcal{T}_h} h_T^4 \|D^2u\|^2_{L^2(T)}, \quad \|D(u - \pi u)\|^2_{L^2(\Omega)} \le C \sum_{T \in \mathcal{T}_h} h_T^2 \|D^2u\|^2_{L^2(T)},$$

where C is a generic constant independent of h and u, but it depends on the ratio between smallest and largest interior angle of the triangles of $\mathcal{T}_h$. Here $\|D(u - \pi u)\|^2_{L^2(\Omega)} = \sum_{T \in \mathcal{T}_h} \|D(u - \pi u)\|^2_{L^2(T)}$.

3.4 L^2-projection

Let $\Omega \subset \mathbb{R}^2$. We consider the space $L^2(\Omega) = \left\{v | \int_\Omega v^2(x,y)dxdy < \infty\right\}$. Let $u \in L^2(\Omega)$. We define the L^2-projection $P_h : \ L^2(\Omega) \to V_h = \left\{v \in C^0(\Omega) |\ v|_T \in P^1(T), \ \forall \ T \in \mathcal{T}_h\right\}$ by $P_h u \in V_h$ such that

$$\int_\Omega (u - P_h u) v_h dxdy = 0, \quad \forall v_h \in V_h.$$

The problem of finding $P_h u \in V_h$ is equivalent to solve the following linear system

$$\int_\Omega (u - P_h u)\phi_i dxdy = 0, \quad i = 1, 2 \ldots, N,$$

where $\{\phi_i\}_{i=1}^N$ is a basis of V_h.

Since $P_h u \in V_h$ we can express it as $P_h u = \sum_{i=1}^N c_i \phi_i(x,y)$, where $c_i \in \mathbb{R}$. Therefore, to find $P_h u \in V_h$ we need to find $c_1, c_2, \ldots, c_N \in \mathbb{R}$ such that

$$\sum_{i=1}^N c_i \int_\Omega \phi_i \phi_j dxdy = \int_\Omega u\phi_j dxdy, \quad j = 1, 2, \ldots, N.$$

The problem can be expressed as a linear system of equations $M\mathbf{c} = \mathbf{b}$, where $\mathbf{c} = [c_1, c_2, \ldots, c_N]^t$ and the entries of the matrix $M \in \mathbb{R}^{N \times N}$ and the vector $\mathbf{b} \in \mathbb{R}^N$ are given by

$$m_{ij} = \int_\Omega \phi_i \phi_j dxdy, \quad b_j = \int_\Omega u\phi_j dxdy.$$

In general, we use a quadrature rule to approximate integrals. The general form is

$$\int_T f(x,y)dxdy \approx \sum_{j=1}^{n} \omega_j f\left(\overline{N}_j\right),$$

where the ω_j 's denote the weights and the $\left(\overline{N}_j\right)$'s the quadrature points.

Lemma 3.1 *The mass matrix M with entries* $m_{ij} = \int_\Omega \phi_i \phi_j dxdy$ *is symmetric and positive definite.*

Theorem 3.3 *For any* $u \in L^2(\Omega)$ *the* L^2*-projection* $P_h u$ *exists and is unique.*

3.5 A priori error estimate

Theorem 3.4 *Let* $u \in L^2(\Omega)$ *and let* $P_h u$ *be the* L^2*-projection of* u*, then*

$$\|u - P_h u\|_{L^2(\Omega)} \leq \|u - v_h\|_{L^2(\Omega)}, \quad \forall\ v_h \in V_h.$$

Theorem 3.5 *Suppose that* $u \in C^2(\Omega)$ *with* $u \in C^2(T)$ *for all* $T \in \mathcal{T}_h$*. Then there exists a constant C such that*

$$\|u - P_h u\|^2_{L^2(\Omega)} \leq C \sum_{T \in \mathcal{T}_h} h_T^4 \|D^2 u\|^2_{L^2(T)}.$$

3.6 The FE method for general elliptic problem

The FE method was designed to approximate solutions to complicated equations of elasticity and structural mechanics, usually modeled by elliptic type equations, with complicated geometries. It has been developed for other applications as well.

Consider the following two-dimensional elliptic problem: Find u such that

$$-\nabla \cdot (a\nabla u) + bu = f, \quad \text{in } \Omega, \quad a\nabla u \cdot \mathbf{n} = \kappa(g - u), \quad \text{on } \partial\Omega, \tag{31}$$

where $a > 0$, $b \geq 0$, $\kappa \geq 0$, $f \in L^2(\Omega)$ and $g \in C^0(\partial\Omega)$. We seek a weak solution u in

$$V = H^1(\Omega) = \left\{ v \in L^2(\Omega) |\ v \text{ has a weak derivative and } \|v\|_{L^2(\Omega)} + \|\nabla v\|_{L^2(\Omega)} < \infty \right\}.$$

In order to derive the weak formulation, we multiply (31) with $v \in V$, integrate over Ω and use Green's formula to obtain

$$\begin{aligned}
\int_\Omega f v dxdy &= -\int_\Omega v \nabla \cdot (a\nabla u) dxdy + \int_\Omega buv dxdy \\
&= \int_\Omega a\nabla u \cdot \nabla v dxdy - \int_{\partial\Omega} v(a\nabla u) \cdot \mathbf{n} ds + \int_\Omega buv dxdy \\
&= \int_\Omega a\nabla u \cdot \nabla v dxdy + \int_\Omega buv dxdy + \int_{\partial\Omega} \kappa(u - g) v ds.
\end{aligned}$$

We obtain the weak form: Find $u \in V$ such that

$$\int_\Omega a\nabla u \cdot \nabla v dxdy + \int_\Omega buv dxdy + \int_{\partial\Omega} \kappa uv ds = \int_\Omega f v dxdy + \int_{\partial\Omega} \kappa g v ds, \quad v \in V. \tag{32}$$

We can formulate the method as in the 1D case by using the weak formulation (32). The FE method in 2D is defined as follows: Find $u_h \in V_h$ such that

$$\int_\Omega a\nabla u_h \cdot \nabla v_h dxdy + \int_\Omega b u_h v_h dxdy + \int_{\partial\Omega} \kappa u_h v_h ds = \int_\Omega f v_h dxdy + \int_{\partial\Omega} \kappa g v_h ds, \quad v_h \in V_h, \tag{33}$$

where $V_h = \{v \in V |\ v|_T \in P^1(T),\ \forall\ T \in \mathcal{T}_h\}$.

Implementation: Let $a = 1$ and $b = g = 0$. Substituting $u_h = \sum_{j=1}^N c_j \phi_j$ into (33) and picking $v_h = \phi_i$, we obtain

$$\sum_{j=1}^N c_j \left(\int_\Omega \nabla\phi_j \cdot \nabla\phi_i dxdy + \int_{\partial\Omega} \kappa\phi_j\phi_i ds \right) = \int_\Omega f\phi_i dxdy, \quad i = 1, 2, \ldots, N.$$

This gives us the system $(A + R)\mathbf{c} = \mathbf{b}$, where $\mathbf{c} = [c_1, c_2, \ldots, c_N]^t \in \mathbb{R}^N$ is the unknown vector and the entries of $A \in \mathbb{R}^{N\times N}$, $R \in \mathbb{R}^{N\times N}$, and $\mathbf{b} \in \mathbb{R}^N$ are given by

$$a_{ij} = \int_\Omega \nabla\phi_j \cdot \nabla\phi_i dxdy, \quad r_{ij} = \int_{\partial\Omega} \kappa\phi_j\phi_i ds, \quad b_i = \int_\Omega f\phi_i dxdy, \quad i, j = 1, 2, \ldots, N.$$

Assembly of the stiffness matrix A: We can again identify the local contributions that come form a particular triangle T

$$a_{ij}^T = \int_\Omega \nabla\phi_j \cdot \nabla\phi_i dxdy, \quad i, j = 1, 2, 3.$$

where T is an arbitrary triangle with vertices $N_i = (x_i, y_i)$ and ϕ_i are the hat functions *i.e.*, $\phi_j(N_i) = \delta_{ij}$. Let $\phi_i(x, y) = \alpha_i + \beta_i x + \gamma_i y$, for $i = 1, 2, 3$. Then, we compute $\alpha_i,\ \beta_i,\ \gamma_i$ by

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix} \begin{bmatrix} \alpha_3 \\ \beta_3 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

In general we have $B\alpha_i = \mathbf{e}_i$ for $i = 1, 2, 3$, where

$$B = \begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}, \quad \alpha_i = \begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{bmatrix}, \quad \mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Furthermore, we obviously have $\nabla\phi_i = [\beta_i, \gamma_i]^t$, which gives

$$a_{ij}^T = \int_\Omega \left(\beta_i\beta_j + \gamma_i\gamma_j\right) dx = \left(\beta_i\beta_j + \gamma_i\gamma_j\right)|T|, \quad i, j = 1, 2, 3.$$

Assembly of boundary matrix R: Let Γ_h^{out} denote the set of boundary edges of the triangulation, *i.e.* $\Gamma_h^{out} = \{E |\ E = T \cap \partial\Omega,\ for\ T \in \mathcal{T}_h\}$. Assume that κ is constant on E. For an edge $E \in \Gamma_h^{out}$, we define $R^E \in \mathbb{R}^{2\times 2}$ by the entries

$$r_{ij}^E = \int_E \kappa \phi_j \phi_i ds = \frac{\kappa}{6}(1+\delta_{ij})|E|, \quad i,j = 1,2,$$

where $|E|$ is the length of E and δ_{ij} is 1 for $i = j$ and 0 else.

Assembly of load vector: We use a corner quadrature rule for approximating the integral. We obtain for $T \in \mathcal{T}_h$

$$b_i^T = \int_T f\phi_i dxdy \approx \frac{|T|}{3} f(N_i), \quad i = 1,2,\ldots,N.$$

Given A, R and $\mathbf{b}$, we can solve $(A+R)\mathbf{c} = \mathbf{b}$ and write $u_h = \sum_{j=1}^N c_j \phi_j$.

3.7 The Dirichlet problem

Consider the following Dirichlet Problem: Find u such that

$$-\Delta u = f, \quad \text{in } \Omega, \quad u = g, \quad \text{on } \partial\Omega, \tag{34}$$

where $f \in L^2(\Omega)$ and $g \in C^0(\partial\Omega)$. We seek a weak solution u in $V_g = \{v \in V | \ v|_{\partial\Omega} = g\}$. Multiplying (34) by a test function $v \in V_0$ and integrating over Ω, we get

$$\int_\Omega fvdxdy = -\int_\Omega v\Delta u dxdy = \int_\Omega \nabla u \cdot \nabla v dxdy - \int_{\partial\Omega} v\nabla u \cdot \mathbf{n} ds = \int_\Omega \nabla u \cdot \nabla v dxdy.$$

So the weak problem reads: Find $u \in V_g$ such that

$$\int_\Omega \nabla u \cdot \nabla v dxdy = \int_\Omega fv dxdy, \quad v \in V_0.$$

Assume that g is piecewise linear on $\partial\Omega$ with respect to the triangulation. Then the FE method in 2D is defined as follows: Find $u_h \in V_{h,g} = \{v \in V_h | \ v|_{\partial\Omega} = g\}$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h dxdy = \int_\Omega f v_h dxdy, \quad v_h \in V_{h,0}.$$

Assume that we have N nodes and J boundary nodes, then the matrix form of the FE method problem reads:

$$\begin{bmatrix} A_{0,0} & A_{0,g} \\ A_{g,0} & A_{0,g} \end{bmatrix} \begin{bmatrix} \mathbf{c}_0 \\ \mathbf{c}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_0 \\ \mathbf{b}_1 \end{bmatrix},$$

where $A_{0,0} \in \mathbb{R}^{(N-J)\times(N-J)}$, $A_{g,g} \in \mathbb{R}^{J\times J}$, $A_{0,g} \in \mathbb{R}^{(N-J)\times J}$, $A_{g,0} \in \mathbb{R}^{J\times(N-J)}$. Note that $\mathbf{c}_1 \in \mathbb{R}^J$ is known (it contains the values of g in the boundary nodes). We can therefore solve the simplified problem reading: find $\mathbf{c}_0 \in \mathbb{R}^{N-J}$ with $A_{0,0}\mathbf{c}_0 = \mathbf{b}_0 - A_{0,g}\mathbf{c}_1$.

3.8 The Neumann problem

Consider the following Neumann Problem: Find u such that

$$-\Delta u = f, \quad \text{in } \Omega, \quad \nabla u \cdot \mathbf{n} = g, \quad \text{on } \partial\Omega, \tag{35}$$

where $f \in L^2(\Omega)$ and $g \in C^0(\partial\Omega)$. Let us try to seek a solution to this problem in the space $V = \left\{v| \ \|v\|_{L^2(\Omega)} + \|\nabla v\|_{L^2(\Omega)} < \infty\right\}$. Multiplying (35) by a test function $v \in V$, integrating over Ω, and using Green's formula, we get

$$\begin{aligned}\int_\Omega fvdxdy &= -\int_\Omega v\Delta udxdy = \int_\Omega \nabla u \cdot \nabla vdxdy - \int_{\partial\Omega} v\nabla u \cdot \mathbf{n}ds \\ &= \int_\Omega \nabla u \cdot \nabla vdxdy - \int_{\partial\Omega} vgds.\end{aligned}$$

Thus, the variational formulation reads: find $u \in V$ such that

$$\int_\Omega \nabla u \cdot \nabla vdxdy - \int_{\partial\Omega} vgds = \int_\Omega fvdxdy, \quad \forall \ v \in V.$$

In order to guarantee solvability, we note that if $v = 1$ then we have

$$0 = \int_\Omega \nabla u \cdot \nabla 1dxdy = \int_\Omega fdxdy + \int_{\partial\Omega} gds.$$

Therefore we need to assume the following compatibility condition

$$\int_\Omega fdxdy + \int_{\partial\Omega} gds = 0,$$

to ensure that a solution can exist. Note that if u exists, it is only determined up to a constant, since $u + c$ is a solution if u is a solution and $c \in \mathbb{R}$. To fix this constant and obtain a unique solution a common trick is to impose the additional constraint $\int_\Omega udxdy = 0$. We therefore define the weak solution space

$$\hat{V} = \left\{v \in V| \ \int_\Omega vdxdy = 0\right\},$$

which contains only functions with a zero mean value. This is a called a quotient space. This space guarantees a unique weak solution (with weak formulation as usual with test functions in V). So the weak problem reads: Find $u \in \hat{V}$ such that

$$\int_\Omega \nabla u \cdot \nabla vdxdy - \int_{\partial\Omega} vgds = \int_\Omega fvdxdy, \quad \forall \ v \in V.$$

Now, the FE method takes the form: find $u_h \in \hat{V}_h \subset \hat{V}$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h dxdy - \int_{\partial\Omega} v_h gds = \int_\Omega fv_h dxdy, \quad \forall \ v_h \in \hat{V}_h,$$

where $\hat{V}_h$ is the space of all continuous piecewise linear functions with a zero mean.

3.9 Finite elements for mixed Dirichlet-Neumann conditions

Here we describe briefly how Neumann conditions are handled in two-dimensional finite elements. Suppose Ω is a domain in either $\mathbb{R}^2$ or $\mathbb{R}^3$ and assume that $\partial\Omega$ has been partitioned into two disjoint sets: $\partial\Omega = \Gamma_1 \cup \Gamma_2$. We consider the following BVP:

$$-\nabla \cdot (\kappa(\mathbf{x})\nabla u) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad u = 0, \quad \mathbf{x} \in \Gamma_1, \quad \nabla u \cdot \mathbf{n} = 0, \quad \mathbf{x} \in \Gamma_2, \tag{36}$$

where $f \in L^2(\Omega)$. As for the 1-D case, Dirichlet conditions are termed essential boundary conditions because they must be explicitly imposed in the FE method, while Neumann conditions are called natural and need not be mentioned. We therefore define the space of test functions by

$$\hat{V} = \left\{ v \in C^2(\overline{\Omega}) : \ v(\mathbf{x}) = 0, \ \mathbf{x} \in \Gamma_1 \right\}.$$

Multiplying (36) by a test function $v \in \hat{V}$ and integrating over Ω, we get

$$\begin{aligned} \int_\Omega f v dx dy &= -\int_\Omega v \nabla \cdot (\kappa(\mathbf{x})\nabla u) dx dy = \int_\Omega \kappa(\mathbf{x}) \nabla u \cdot \nabla v dx dy - \int_{\partial\Omega} \kappa(\mathbf{x}) v \nabla u \cdot \mathbf{n} ds \\ &= \int_\Omega \kappa(\mathbf{x}) \nabla u \cdot \nabla v dx dy - \int_{\Gamma_1} \kappa(\mathbf{x}) v \nabla u \cdot \mathbf{n} ds - \int_{\Gamma_2} \kappa(\mathbf{x}) v \nabla u \cdot \mathbf{n} ds \\ &= \int_\Omega \kappa(\mathbf{x}) \nabla u \cdot \nabla v dx dy, \end{aligned}$$

since $v = 0$ on Γ_1 and $\nabla u \cdot \mathbf{n}$ on Γ_1. Thus the weak form of (36) is: Find $u \in \hat{V}$ such that

$$\int_\Omega \kappa(\mathbf{x}) \nabla u \cdot \nabla v dx dy = \int_\Omega f v dx dy, \quad v \in \hat{V}. \tag{37}$$

We now restrict our discussion once more to two-dimensional polygonal domains. To apply the FE method, we must choose an approximating subspace of $\hat{V}$. Since the boundary conditions are mixed, there are at least two points where the boundary conditions change from Dirichlet to Neumann. We will make the assumption that the mesh is chosen so that all such points are nodes (and that all such nodes belong to Γ_1, that is, that Γ_1 includes its "endpoints"). We can then choose the approximating subspace of $\hat{V}$ as follows:

$$V_h = \left\{ v \in C(\overline{\Omega}) : \ \text{v is linear on } \mathcal{T}_\text{h}, \ v(\mathbf{z}) = 0 \ \text{ for all nodes } \mathbf{z} \in \Gamma_1 \right\}.$$

A basis for V_h is formed by including all basis functions corresponding to interior boundary nodes that do not belong to Γ_1. If the BVP includes only Neumann conditions, then the stiffness matrix will be singular, reflecting the fact that BVP either does not have a solution or has infinitely many solutions. Special care must be taken to compute a meaningful solution to the resulting linear system.

3.10 The method of shifting the data

3.10.1 Inhomogeneous Dirichlet conditions on a rectangle

In a two-dimensional problem, inhomogeneous boundary conditions are handled just as in one dimension. Inhomogeneous Dirichlet conditions are addressed via the method of shifting the data (with a specially chosen piecewise linear function), while inhomogeneous Neumann conditions are taken into account directly when deriving the weak form. Both types of boundary conditions lead to a change in the load vector.

The method of shifting the data can be used to transform an inhomogeneous Dirichlet problem to a homogeneous Dirichlet problem. This technique works just as it did for a one-dimensional problem, although in two dimensions it is more difficult to find a function satisfying the boundary conditions. We consider the BVP

$$-\Delta u = f(\mathbf{x}), \quad \mathbf{x} \in \Omega = (0,a) \times (0,b), \quad u(\mathbf{x}) = g(\mathbf{x}) = \begin{cases} g_1(\mathbf{x}), & \mathbf{x} \in \Gamma_1, \\ g_2(\mathbf{x}), & \mathbf{x} \in \Gamma_2, \\ g_3(\mathbf{x}), & \mathbf{x} \in \Gamma_3, \\ g_4(\mathbf{x}), & \mathbf{x} \in \Gamma_4, \end{cases} \tag{38}$$

where Γ_1, Γ_2, Γ_3, and Γ_4 are, respectively, the bottom, right, top, and left boundary edges of the rectangular domain $\Omega = (0,a) \times (0,b)$. We will assume that the boundary data are continuous, so

$$g_1(0) = g_4(0), \quad g_1(a) = g_2(0), \quad g_2(b) = g_3(a), \quad g_3(0) = g_4(b).$$

Suppose we find a function w defined on $\overline{\Omega}$ and satisfying $w(\mathbf{x}) = g(\mathbf{x})$ for all $\mathbf{x} \in \partial\Omega$. We then define $v = u - w$ and note that

$$-\Delta v = -\Delta u + \Delta w = f(\mathbf{x}) + \Delta w = \hat{f}(\mathbf{x}),$$

and $v(\mathbf{x}) = u(\mathbf{x}) - w(\mathbf{x}) = 0$ for all $\mathbf{x} \in \partial\Omega$. We can then solve

$$-\Delta v = \hat{f}(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad v(\mathbf{x}) = 0, \quad \mathbf{x} \in \partial\Omega. \tag{39}$$

Finally, the solution u will be given by $u = v + w$.

We now describe a method for computing a function w that satisfies the given Dirichlet conditions. We first note that there is a polynomial of the form $q(x,y) = c_0 + c_1 x + c_2 y + c_3 xy$, which assumes the desired boundary values at the corners:

$$\begin{aligned} q(0,0) &= g_1(0) = g_4(0), \quad q(a,0) = g_1(a) = g_2(0), \quad q(a,b) = g_2(b) \\ &= g_3(a), \quad q(0,b) = g_3(0) = g_4(b). \end{aligned}$$

A direct calculation shows that

$$\begin{aligned} c_0 &= g_1(0), \quad c_1 = \frac{g_1(a) - g_1(0)}{a}, \quad c_2 = \frac{g_4(b) - g_4(0)}{b}, \\ c_3 &= \frac{g_2(b) + g_1(0) - g_1(a) - g_4(b)}{ab}. \end{aligned}$$

We then define

$$h(\mathbf{x}) = \begin{cases} h_1(x) = g_1(x) - \left(g_1(0) + \dfrac{g_1(a) - g_1(0)}{a} x\right), & \mathbf{x} \in \Gamma_1, \\ h_2(y) = g_2(y) - \left(g_2(0) + \dfrac{g_2(b) - g_2(0)}{b} y\right), & \mathbf{x} \in \Gamma_2, \\ h_3(x) = g_3(x) - \left(g_3(0) + \dfrac{g_3(a) - g_3(0)}{a} x\right), & \mathbf{x} \in \Gamma_3, \\ h_4(y) = g_4(y) - \left(g_4(0) + \dfrac{g_4(b) - g_4(0)}{b} y\right), & \mathbf{x} \in \Gamma_4. \end{cases}$$

We have thus replaced each g_i by a function h_i which differs from g_i by a linear function, and which has value zero at the two endpoints:

$$h_1(0) = h_1(a) = h_2(0) = h_2(b) = h_3(0) = h_3(a) = h_4(0) = h_4(b) = 0.$$

Finally, we define

$$w(x,y) = (c_0 + c_1x + c_2y + c_3xy) + \left(h_1(x) + \frac{h_3(x) - h_1(x)}{b}y\right) + \left(h_4(y) + \frac{h_2(y) - h_4(y)}{a}x\right).$$

The reader should notice how the second term interpolates between the boundary values on Γ_1 and Γ_3, while the third term interpolates between the boundary values on Γ_2 and Γ_4. In order for these two terms not to interfere with each other, it is necessary that the boundary data be zero at the corners. It was for this reason that we transformed the g_i's into the h_i's. The first term in the formula for w undoes this transformation. It is straightforward to verify that w satisfies the desired boundary conditions.

3.10.2 Inhomogeneous Neumann conditions on a rectangle

We can also apply the technique of shifting the data to transform a BVP with inhomogeneous Neumann conditions to a related BVP with homogeneous Neumann conditions. However, the details are somewhat more involved than in the Dirichlet case. Consider the following BVP with the Neumann conditions

$$-\Delta u = f(\mathbf{x}), \quad \mathbf{x} \in \Omega = (0,a) \times (0,b), \quad \mathbf{n} \cdot \nabla u(\mathbf{x}) = g(\mathbf{x}) = \begin{cases} g_1(\mathbf{x}), & \mathbf{x} \in \Gamma_1, \\ g_2(\mathbf{x}), & \mathbf{x} \in \Gamma_2, \\ g_3(\mathbf{x}), & \mathbf{x} \in \Gamma_3, \\ g_4(\mathbf{x}), & \mathbf{x} \in \Gamma_4, \end{cases} \tag{40}$$

where Γ_1, Γ_2, Γ_3, and Γ_4 are, respectively, the bottom, right, top, and left boundary edges of the rectangular domain $\Omega = (0,a) \times (0,b)$. We first note that this is equivalent to

$$-u_y(\mathbf{x}) = g_1(x), \ \mathbf{x} \in \Gamma_1, \quad u_x(\mathbf{x}) = g_2(y), \ \mathbf{x} \in \Gamma_2, \quad u_y(\mathbf{x}) = g_3(x), \ \mathbf{x} \in \Gamma_3, \quad -u_x(\mathbf{x}) = g_4(y), \ \mathbf{x} \in \Gamma_4.$$

We make the following observation: If there is a twice-continuously differentiable function u satisfying the given Neumann conditions, then, since $u_{xy} = u_{yx}$, we have

$$-u_{xy}(x,0) = g_1'(x), \quad -u_{yx}(0,y) = g_{4'}(y),$$

which together imply that $g_1'(0) = g_{4'}(0)$. By similar reasoning, we have all of the following conditions:

$$g_1'(0) = g_{4'}(0), \quad g_1'(0) = g_{4'}(0), \quad -g_1'(a) = g_2'(0), \quad g_2'(b) = g_3'(a). \tag{41}$$

We will assume that (41) holds.

We now explain how to compute a function that satisfies the desired Neumann conditions. The method is similar to that used to shift the data in a Dirichlet problem: we will "interpolate" between the Neumann conditions in each dimension and arrange things so that the two interpolations do not interfere with each other. We use the fact that

$$\psi(x) = -\alpha x + \frac{\alpha+\beta}{2a}x^2 \quad \text{satisfies} \quad \psi'(0) = -\alpha, \quad \psi'(a) = \beta. \tag{42}$$

The first step is to transform the boundary data $g_l(x)$ to a function $h_1(x)$ satisfying $h_{1'}(0) = h_{1'}(a) = 0$, and similarly for g_2, g_3, g_4 and h_2, h_3, h_4. Since these derivatives of the boundary data at the corners are (plus or minus) the mixed partial derivatives of the desired function at the corners, it suffices to find a function $q(x,y)$ satisfying the conditions

$$u_{xy}(0,0) = -g_1'(0), \quad u_{xy}(a,0) = -g_1'(a), \quad u_{xy}(0,b) = -g_3'(0), \quad u_{xy}(a,b) = -g_2'(b).$$

We can satisfy these conditions with a function of the form $q(x,y) = c_0xy + c_1x^2y + c_2xy^2 + c_3x^2y^2$. The reader can verify that the necessary coefficients are

$$c_0 = -g_1'(0), \quad c_1 = \frac{g_1'(0) - g_1'(a)}{2a}, \quad c_2 = \frac{g_3'(0) + g_1'(0)}{2b},$$
$$c_3 = \frac{g_2'(b) + g_1'(a) - g_3'(0) - g_1'(0)}{4ab}.$$

If w is to satisfy the desired Neumann conditions, then $w - q = h_i$ on Γ_i, $i = 1-4$, where

$$h_1(x) = g_1(x) + c_0x + c_1x^2, \quad h_2(y) = g_2(y) - (c_0 + 2ac_1)y - (c_2 + 2ac_3)y^2,$$
$$h_3(x) = g_3(x) - (c_0 + 2bc_2)x - (c_1 + 2bc_3)x^2, \quad h_4(y) = g_4(y) + c_0y + c_2y^2.$$

We can now define $w - q$ by the interpolation described by (42):

$$w(x,y) = q(x,y) - h_1(x)y + \frac{h_3(x) + h_1(x)}{2b}y^2 - h_4(y)x + \frac{h_2(y) + h_4(y)}{2a}yx^2.$$

Then w satisfies the original Neumann conditions, as the interested reader can verify directly.

3.11 Eigenvalue problem

Consider the following Eigenvalue Problem: Find $\lambda \in \mathbb{R}$ and u such that

$$-\Delta u = \lambda u, \quad \text{in } \Omega, \quad \nabla u \cdot \mathbf{n} = 0, \quad \text{on } \partial\Omega. \tag{43}$$

In order to derive the weak formulation, we multiply (43) with $v \in V$, integrate over Ω and use Green's formula to obtain

$$\lambda \int_\Omega uvdxdy = -\int_\Omega v\Delta udxdy = \int_\Omega \nabla u \cdot \nabla vdxdy - \int_{\partial\Omega} v\nabla u \cdot \mathbf{n}ds = \int_\Omega \nabla u \cdot \nabla vdxdy.$$

We obtain the weak form: Find $u \in V$ such that

$$\int_\Omega \nabla u \cdot \nabla v dxdy = \lambda \int_\Omega uv dxdy, \quad v \in V. \tag{44}$$

The FE method in 2D is defined as follows: Find $\lambda_h \in \mathbb{R}$ and $u_h \in V_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h dxdy = \lambda_h \int_\Omega u_h v_h dxdy, \quad v_h \in V_h, \tag{45}$$

where $V_h = \{v \in V | \ v|_T \in P^1(T), \ \forall \ T \in \mathcal{T}_h\}$.

Implementation: Substituting $u_h = \sum_{j=1}^N c_j \phi_j$ into (45) and picking $v_h = \phi_i$, we obtain

$$\sum_{j=1}^N c_j \left(\int_\Omega \nabla \phi_j \cdot \nabla \phi_i dxdy - \lambda_h \int_\Omega \phi_i \phi_j dxdy \right) = 0, \quad i = 1, 2, \ldots, N.$$

This leads to an algebraic system of the form $A\mathbf{c} = \lambda_h M\mathbf{c}$, *i.e.* an algebraic eigenvalue problem.

3.12 Error analysis

Consider the following model Problem: Find u such that

$$-\Delta u = f, \quad \text{in } \Omega, \qquad u = 0, \quad \text{on } \partial\Omega.$$

The weak form: Find $u \in V_0$ such that

$$\int_\Omega \nabla u \cdot \nabla v dxdy = \int_\Omega fv dxdy, \quad v \in V_0.$$

The FE approximation is defined as follows: Find $u_h \in V_{h,0}$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h dxdy = \int_\Omega f v_h dxdy, \quad v_h \in V_{h,0},$$

where $V_h = \{v \in V | \ v|_T \in P^1(T), \ \forall \ T \in \mathcal{T}_h\}$. Expressing $u_h = \sum_{j=1}^N c_j \phi_j$ and picking $v_h = \phi_i$, we obtain

$$\sum_{j=1}^N c_j \left(\int_\Omega \nabla \phi_j \cdot \nabla \phi_i dxdy \right) = \int_\Omega f \phi_i dxdy, \quad i = 1, 2, \ldots, N.$$

This leads to system of the form $A\mathbf{c} = \mathbf{b}$, where the entries of $A \in \mathbb{R}^{N \times N}$ and $\mathbf{b} \in \mathbb{R}^N$ are

$$a_{ij} = \int_\Omega \nabla \phi_j \cdot \nabla \phi_i dxdy, \quad b_i = \int_\Omega f \phi_i dxdy, \quad i, j = 1, 2, \ldots, N.$$

Theorem 3.6 *The stiffness matrix A is symmetric and positive definite.*

Theorem 3.7 (Galerkin orthogonality) *Let $u \in V_0$ denote the weak solution and $u_h \in V_{h,0}$ the corresponding FE method approximation. Then*

$$\int_\Omega \nabla (u - u_h) \cdot \nabla v_h dxdy = 0, \quad v_h \in V_{h,0}.$$

Now, let $|||v|||^2 = \int_\Omega \nabla v \cdot \nabla v dxdy = \int_\Omega |\nabla v|^2 dxdy$ be the energy norm on V_0.

There are two different kinds of error estimates, *a priori* estimates, where the error is bounded in terms of the exact solution, and *a posteriori* error estimates, where the error is bounded in terms of the computed solution.

Theorem 3.8 (A priori error bound) *Let $u \in V_0$ denote the weak solution and $u_h \in V_{h,0}$ the corresponding FE method approximation. Then*

$$|||u - u_h||| \leq |||u - v_h|||, \quad v_h \in V_{h,0}.$$

Theorem 3.9 *Let $u \in V_0$ denote the weak solution and $u_h \in V_{h,0}$ the corresponding FE method approximation. If $u \in C^2(\Omega)$, then there exists C independent of h_T and u such that*

$$|||u - u_h|||^2_{L^2(\Omega)} \leq C \sum_{T \in \mathcal{T}_h} h_T^2 \|D^2 u\|^2_{L^2(T)}.$$

3.13 The FE method for elliptic problems with a convection term

Consider the following convection-diffusion problem: Find u such that

$$-\nabla \cdot (a\nabla u) + \mathbf{b} \cdot \nabla u + cu = f, \quad \text{in } \Omega, \quad u = 0, \quad \text{on } \partial\Omega. \tag{46}$$

We seek a weak solution u in $V_0 = \{v \in V | \ v|_{\partial\Omega} = 0\}$. In order to derive the weak formulation, we multiply (46) with $v \in V_0$, integrate over Ω and use Green's formula to obtain

$$\begin{aligned}\int_\Omega f v dxdy &= -\int_\Omega v \nabla \cdot (a\nabla u) dxdy + \int_\Omega v\mathbf{b} \cdot \nabla u dxdy + \int_\Omega cuv dxdy \\ &= \int_\Omega a\nabla u \cdot \nabla v dxdy - \int_{\partial\Omega} v\nabla u \cdot \mathbf{n} ds + \int_\Omega v\mathbf{b} \cdot \nabla u dxdy + \int_\Omega cuv dxdy \\ &= \int_\Omega a\nabla u \cdot \nabla v dxdy + \int_\Omega v\mathbf{b} \cdot \nabla u dxdy + \int_\Omega cuv dxdy.\end{aligned}$$

Note that there is no need to apply Green's formula to $\int_\Omega v\mathbf{b} \cdot \nabla u dxdy$. We obtain the weak form: Find $u \in V_0$ such that

$$\int_\Omega a\nabla u \cdot \nabla v dxdy + \int_\Omega v\mathbf{b} \cdot \nabla u dxdy + \int_\Omega cuv dxdy = \int_\Omega f v dxdy, \quad v \in V_0.$$

The FE method in 2D is defined as follows: Find $u_h \in V_{h,0} = \{v \in V_h | \ v|_{\partial\Omega} = 0\}$ such that

$$\int_\Omega a\nabla u_h \cdot \nabla v_h dxdy + \int_\Omega v_h\mathbf{b} \cdot \nabla u_h dxdy + \int_\Omega cu_h v_h dxdy = \int_\Omega f v_h dxdy, \quad v_h \in V_{h,0}, \tag{47}$$

where $V_h = \{v \in V | \ v|_T \in P^1(T), \ \forall \ T \in \mathcal{T}_h\}$.

Implementation: Substituting $u_h = \sum_{j=1}^N c_j \phi_j$ into (47) and picking $v_h = \phi_i$, we obtain

$$\sum_{j=1}^N c_j \left(\int_\Omega a\nabla \phi_j \cdot \nabla \phi_i dxdy + \int_\Omega \phi_i \mathbf{b} \cdot \nabla \phi_j dxdy + \int_\Omega c\phi_i \phi_j dxdy \right) = \int_\Omega f\phi_i dxdy, \quad i = 1, 2, \ldots, N.$$

This gives us the system $(A+B+C)\mathbf{c}=\mathbf{d}$, where $\mathbf{c}=[c_1,\ldots,c_N]^t\in\mathbb{R}^N$ is the unknown vector and the entries of $A,B,C\in\mathbb{R}^{N\times N}$ and $\mathbf{d}\in\mathbb{R}^N$ are given by

$$a_{ij}=\int_\Omega a\nabla\phi_j\cdot\nabla\phi_i dxdy,\quad b_{ij}=\int_\Omega \phi_i\mathbf{b}\cdot\nabla\phi_j dxdy,\quad c_{ij}=\int_\Omega c\phi_i\phi_j dxdy,\quad d_i=\int_\Omega f\phi_i dxdy,$$

for $i,j=1,2,\ldots,N$. Note that B is not symmetric, *i.e.* $b_{ij}\neq b_{ji}$.

4. The FE method for the heat equation

Consider the following heat/diffusion problem: Find $u(\mathbf{x},t)$ such that

$$\dot{u}-\Delta u=f,\quad \text{in } \Omega\subset\mathbb{R}^2,\quad t\in[0,T], \tag{48}$$

$$u(\cdot,t)=0,\quad \text{on } \partial\Omega \text{ and } t\in[0,T], \tag{49}$$

$$u(\mathbf{x},0)=u_0(\mathbf{x}),\quad \text{for } \mathbf{x}\in\Omega \text{ and } t=0. \tag{50}$$

We seek a weak solution u in $V_0=\left\{v|\ \|v\|+\|\nabla v\|<\infty,\ v|_{\partial\Omega}=0\right\}$. In order to derive the weak formulation, we multiply (48) with $v\in V_0$, integrate over Ω and use Green's formula to obtain, for $t\in[0,T]$,

$$\int_\Omega fvd\mathbf{x}=\int_\Omega \dot{u}vd\mathbf{x}+\int_\Omega\nabla u\cdot\nabla vd\mathbf{x}-\int_{\partial\Omega}v\nabla u\cdot\mathbf{n}ds=\int_\Omega\dot{u}vd\mathbf{x}+\int_\Omega\nabla u\cdot\nabla vd\mathbf{x}.$$

The weak form therefore reads: Find $u(\cdot,t)\in V_0$ such that for $t>0$

$$\int_\Omega\dot{u}vd\mathbf{x}+\int_\Omega\nabla u\cdot\nabla vd\mathbf{x}=\int_\Omega fvd\mathbf{x},\quad v\in V_0. \tag{51}$$

The semi-discrete FE method in 2D is defined as follows: Find $u_h(\cdot,t)\in V_{h,0}=\left\{v\in V_h|\ v|_{\partial\Omega}=0\right\}$ such that

$$\int_\Omega\dot{u}_hv_hd\mathbf{x}+\int_\Omega\nabla u_h\cdot\nabla v_hd\mathbf{x}=\int_\Omega fv_hd\mathbf{x},\quad v_h\in V_{h,0}, \tag{52}$$

where $V_h=\left\{v\in V|\ v|_T\in P^1(T),\ \forall\ T\in\mathcal{T}_h\right\}$.

Implementation: Substituting $u_h(\mathbf{x},t)=\sum_{j=1}^N c_j(t)\phi_j(\mathbf{x})$ into (52) and choosing $v_h=\phi_i$, we obtain

$$\sum_{j=1}^N\dot{c}_j\int_\Omega\phi_j\phi_id\mathbf{x}+\sum_{j=1}^N c_j\int_\Omega\nabla\phi_j\cdot\nabla\phi_id\mathbf{x}=\int_\Omega f\phi_id\mathbf{x},\quad i=1,2,\ldots,N.$$

This gives us the system of ODEs

$$M\dot{\mathbf{c}}(t)+A(t)\mathbf{c}(t)=\mathbf{b}(t),\quad t\in(0,T],\quad \mathbf{c}(0)=\mathbf{c}_0,$$

where $\mathbf{c}=[c_1,c_2,\ldots,c_N]^t=[u_h(N_1,t),\ldots,u_h(N_N,t)]^t\in\mathbb{R}^N$ (here N_i denotes the node that belongs to the basis function ϕ_i) is the unknown vector and the entries of $M,\ A\in\mathbb{R}^{N\times N}$ and $\mathbf{b}\in\mathbb{R}^N$ are given by

$$m_{ij}=\int_\Omega\phi_i\phi_jd\mathbf{x},\quad a_{ij}=\int_\Omega\nabla\phi_j\cdot\nabla\phi_id\mathbf{x},\quad b_i=\int_\Omega f\phi_id\mathbf{x},\quad i,j=1,2,\ldots,N.$$

Finally, the system of ODEs can be solved with *e.g.*, the backward Euler method as follows: Let $0 = t_0 < t_1 < \cdots < t_M = T$ be a discretization, let $k_m = t_m - t_{m-1}$ for $m = 1, 2, \ldots, M$ be the time step size and let $\mathbf{c}^m \approx \mathbf{c}(t_m)$ for $m = 1, 2, \ldots, M$ denote corresponding approximations. Then, we can compute $\mathbf{c}^m$ using

$$(M + k_m A_m)\mathbf{c}^m = M\mathbf{c}^{m-1} + k_m \mathbf{b}_m, \quad m = 1, 2, \ldots, M,$$

where $\mathbf{c}^0$ is obtained from $u_0(\mathbf{x})$. We can either use $\mathbf{c}^0 = [c_1(0), \ldots, c_N(0)]^t = [u_0(N_1), \ldots, u_0(N_N)]^t$, or we can let $\mathbf{c}^0$ to be the L^2-projection of u_0. We set $u_h^0 = \sum_{j=1}^N c_j^0 \phi_j(\mathbf{x})$ and solve for c_j^0 using

$$\sum_{j=1}^{N} c_j^0 \int_\Omega \phi_j \phi_i d\mathbf{x} = \int_\Omega u_0 \phi_i d\mathbf{x}, \quad i = 1, 2, \ldots, N.$$

Theorem 4.1 (Stability) *There hold continuous and discrete stability estimates*

$$\|u(\cdot, t)\| \le \|u(\cdot, 0)\| + \int_0^t \|f(\cdot, s)\| ds, \quad \left\|u_h^m\right\| \le \left\|u_h^{m-1}\right\| + k_m \left\|f_m\right\| \le \left\|u_h^0\right\| + \sum_{i=1}^{m} k_i \left\|f_i\right\|.$$

5. The FE method for the wave equation

Many physical phenomena exhibit wave characteristics. For instance light which is an electromagnetic wave have the ability to disperse and create diffraction patterns, which is typical of waves.

Consider the following wave problem: Find $u(\mathbf{x}, t)$ such that

$$\ddot{u} - \nabla \cdot (\varepsilon \nabla u) = f, \quad \text{in } \Omega \subset \mathbb{R}^2, \quad t \in [0, T], \tag{53}$$

$$\mathbf{n} \cdot \nabla u(\cdot, t) = 0, \quad \text{on } \partial\Omega \text{ and } t \in [0, T], \tag{54}$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \dot{u}(\mathbf{x}, 0) = v_0(\mathbf{x}), \quad \text{for } \mathbf{x} \in \Omega \text{ and } t = 0, \tag{55}$$

where f is a given load, $\varepsilon = \varepsilon(\mathbf{x}, t)$ is a positive parameter, u_0 and v_0 are a prescribed initial conditions, and Ω is a bounded domain with boundary $\partial\Omega$ and unit outward normal $\mathbf{n}$.

We seek a weak solution u in $V = H^1(\Omega) = \{v \mid \|v\| + \|\nabla v\| < \infty\}$. Multiplying the wave Eq. (53) with $v \in V$, integrating over Ω, and using Green's formula, we obtain, for $t \in [0, T]$,

$$\begin{aligned}\int_\Omega f v d\mathbf{x} = \int_\Omega \ddot{u} v d\mathbf{x} - \int_\Omega v \nabla \cdot (\varepsilon \nabla u) d\mathbf{x} &= \int_\Omega \ddot{u} v d\mathbf{x} + \int_\Omega \varepsilon \nabla u \cdot \nabla v d\mathbf{x} - \int_{\partial\Omega} v \varepsilon \nabla u \cdot \mathbf{n} ds \\ &= \int_\Omega \ddot{u} v d\mathbf{x} + \int_\Omega \varepsilon \nabla u \cdot \nabla v d\mathbf{x}.\end{aligned}$$

The weak form (variational formulation) therefore reads: Find $u(\cdot, t) \in V = H^1(\Omega)$ such that for all $t > 0$

$$\int_\Omega \ddot{u} v d\mathbf{x} + \int_\Omega \varepsilon \nabla u \cdot \nabla v d\mathbf{x} = \int_\Omega f v d\mathbf{x}, \quad v \in V. \tag{56}$$

Let $V_h = \left\{v \in V \mid v|_T \in P^1(T), \ \forall \ T \in \mathcal{T}_h\right\} \subset V$ be the space of all continuous piecewise linear functions on a triangle mesh of Ω. The semi-discrete FE method in 2D is defined as follows: Find $u_h(\cdot, t) \in V_h$ such that

$$\int_\Omega \ddot{u}_h v_h d\mathbf{x} + \int_\Omega \varepsilon \nabla u_h \cdot \nabla v_h d\mathbf{x} = \int_\Omega f v_h d\mathbf{x}, \quad v_h \in V_h. \tag{57}$$

Implementation: Substituting $u_h(\mathbf{x},t) = \sum_{j=1}^N c_j(t)\phi_j(\mathbf{x})$ into (57) and choosing $v_h = \phi_i$, we obtain

$$\sum_{j=1}^N \ddot{c}_j \int_\Omega \phi_j \phi_i d\mathbf{x} + \sum_{j=1}^N c_j \int_\Omega \varepsilon \nabla \phi_j \cdot \nabla \phi_i d\mathbf{x} = \int_\Omega f \phi_i d\mathbf{x}, \quad i = 1, 2, \ldots, N.$$

This gives us the system

$$M\ddot{\mathbf{c}}(t) + A(t)\mathbf{c}(t) = \mathbf{b}(t), \quad t \in (0, T], \tag{58}$$

where $\mathbf{c} = [c_1, \ldots, c_N]^t = [u_h(N_1,t), \ldots, u_h(N_N,t)]^t \in \mathbb{R}^N$ (here N_i denotes the node that belongs to the basis function ϕ_i) is the unknown vector and the entries of the mass and stiffness matrices M, $A \in \mathbb{R}^{N\times N}$ and the load vector $\mathbf{b} \in \mathbb{R}^N$ are given by

$$m_{ij} = \int_\Omega \phi_i \phi_j d\mathbf{x}, \quad a_{ij} = \int_\Omega \varepsilon \nabla \phi_j \cdot \nabla \phi_i d\mathbf{x}, \quad b_i = \int_\Omega f \phi_i d\mathbf{x}, \quad i,j = 1, 2, \ldots, N.$$

Eq. (58) is a semi-discretization of the wave equation in the sense that it does not contain any unknowns with spatial derivatives.

Time discretization: We first transform the system of ODEs into a first-order system. Let $\mathbf{d}(t) = \dot{\mathbf{c}}(t)$, we get the new coupled system

$$M\dot{\mathbf{c}}(t) - M\mathbf{d}(t) = 0, \quad M\dot{\mathbf{d}}(t) + A(t)\mathbf{c}(t) = \mathbf{b}(t), \quad t \in (0, T].$$

Let $\mathbf{w} = [\mathbf{c}, \mathbf{d}]^t$ then the system is equivalent to $\hat{M}\dot{\mathbf{w}}(t) + \hat{A}(t)\mathbf{w}(t) = \hat{\mathbf{b}}(t)$, $t \in (0, T]$, where

$$\hat{M} = \begin{bmatrix} M & 0 \\ 0 & M \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} 0 & -M \\ A & 0 \end{bmatrix}, \quad \hat{\mathbf{b}} = \begin{bmatrix} 0 \\ \mathbf{b} \end{bmatrix}.$$

Finally, the system of ODEs can be solved with *e.g.*, the backward Euler method as follows: Let $0 = t_0 < t_1 < \cdots < t_M = T$ be a discretization, let $k_m = t_m - t_{m-1}$ for $m = 1, 2, \ldots, M$ be the time step size and let $\mathbf{w}^m \approx \mathbf{w}(t_m)$ for $m = 1, 2, \ldots, M$ denote corresponding approximations. Then, we can compute $\mathbf{w}^m$ using

$$\left(\hat{M} + k_m \hat{A}_m\right)\mathbf{w}^m = \hat{M}\mathbf{w}^{m-1} + k_m \hat{\mathbf{b}}_m, \quad m = 1, 2, \ldots, M,$$

where $\mathbf{w}^0$ is obtained from $u_0(\mathbf{x})$ and $v_0(\mathbf{x})$.

There are several possible choices of initial data. We can either use $\mathbf{w}^0 = [w_1(0), \ldots, c_{2N}(0)]^t = [u_0(N_1), \ldots, u_0(N_N), v_0(N_1), \ldots, v_0(N_N)]^t$, or we can let $\mathbf{w}^0 = \left[\mathbf{w}_1^0, \mathbf{w}_2^0\right]^t$, where $\mathbf{w}_1^0$ and $\mathbf{w}_2^0$ are the L^2-projection of u_0 and v_0, respectively. We set $w_{h,1}^0 = \sum_{j=1}^N w_{j,1}^0 \phi_j(\mathbf{x})$ and $w_{h,2}^0 = \sum_{j=1}^N w_{j,2}^0 \phi_j(\mathbf{x})$ and solve for $w_{j,1}^0$, $w_{j,2}^0$ using

$$\sum_{j=1}^N w_{j,1}^0 \int_\Omega \phi_j \phi_i d\mathbf{x} = \int_\Omega u_0 \phi_i d\mathbf{x}, \quad \sum_{j=1}^N w_{j,2}^0 \int_\Omega \phi_j \phi_i d\mathbf{x} = \int_\Omega v_0 \phi_i d\mathbf{x}, \quad i = 1, 2, \ldots, N.$$

We can also use Crank–Nicolson scheme

$$\left(\hat{M}+\frac{k_m}{2}\hat{A}_m\right)\mathbf{w}^m=\left(\hat{M}-\frac{k_m}{2}\hat{A}_{m-1}\right)\mathbf{w}^{m-1}+\frac{k_m}{2}\left(\hat{\mathbf{b}}_{m-1}+\hat{\mathbf{b}}_m\right)\equiv\mathbf{g}_m.$$

Theorem 5.1 (Conservation of energy) *If* $f=0$, *then*

$$\|\dot{u}_h(\cdot,t)\|^2_{L^2(\Omega)}+\varepsilon\|\nabla u_h(\cdot,t)\|^2_{L^2(\Omega)}=\|\dot{u}(\cdot,0)\|^2_{L^2(\Omega)}+\varepsilon\|\nabla u(\cdot,0)\|^2_{L^2(\Omega)}.$$

6. Conclusion

In this chapter, we introduced the finite element (FE) method for approximation the solutions to ODEs and PDEs. More specifically, the FE method is presented for first-order initial-value problems for OEDs, second-order boundary-value problems for ODEs, second-order elliptic PDEs, second-order heat and wave equations. The remaining chapters of this textbook are based on the FE method. The derivation of the FE method for other problems is straightforward. In the remaining chapters, the FE method will developed to solve complicated problems in engineering, notably in elasticity and structural mechanics modeling involving elliptic partial differential equations and complicated geometries. For more details, we refer the reader to [1–4, 6–9] and the references therein.

Author details

Mahboub Baccouch
Department of Mathematics, University of Nebraska at Omaha, Omaha, NE, USA

*Address all correspondence to: mbaccouch@unomaha.edu

References

[1] Ainsworth M. and Oden J. T. *A posteriori Error Estimation in Finite Element Analysis*. John Wiley, New York, 2000.

[2] Brenner S. C. and Scott L. R. *The Mathematical Theory of Finite Element Methods, second edition*. Springer-Verlag, New York, 2002.

[3] Ciarlet P. G. *The finite element method for elliptic problems*. North-Holland Pub. Co., Amsterdam-New York-Oxford, 1978.

[4] Johnson C. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, New York, 1987.

[5] Kaltenbacher M. *Numerical simulation of mechatronic sensors and actuators: finite elements for computational multiphysics*. Springer, Heidelberg, 2015.

[6] Larson M. *The finite element method: theory, implementation, and applications*. Springer, Berlin New York, 2013.

[7] Oden J. T. and Carey G. F. *Finite Elements, Mathematical Aspects*. Prentice Hall, Englewood Cliffs, 1983.

[8] Schwab C. *p– and hp– Finite Element Methods*. Oxford University Press, New York, 1998.

[9] Szabo B. and Babu I. s̆ ka. *Finite element analysis*. Wiley, New York, 1991.

[10] Hrennikoff A. Solution of problems of elasticity by the framework method. *Journal of Applied Mechanics*, 8(4):169–175, 1941.

[11] Courant R. Variational methods for the solution of problems of equilibrium and vibrations. bulletin of the american mathematical society. 49: 1–23. doi: 10.1090. Technical report, 1943.

[12] Strang G. and Fix G. J. An analysis of the finite element method. 1973.