Contents

Chapter 6

Fundamentals of Numerical Analysis

In Chap. 5, covering several boundary value problems of elliptic partial differential equations, we saw that the existence of their unique solutions could be guaranteed using solutions of the weak form. These will be referred to as the exact solutions. Exact solutions can be found analytically if the shape of the domain is somewhat simple such as a rectangle or an ellipse. However, difficulties arise for domains whose shape may have arbitrarily moved as examined in this book. In order to solve a shape optimization problem, even if an exact solution is not possible, one can resort to a numerical analysis method to obtain an approximate solution.

This chapter looks at how to obtain an approximate solution to such boundary value problems. The Galerkin method is considered first in this chapter. In the Galerkin method, approximate functions with respect to the solution function and an arbitrarily selected variational function (test function) are constructed with linear combinations of given basis functions multiplied by undetermined multipliers which are determined by substituting the approximate functions into the weak form. The characteristics of this method are not only its clarity but also the fact that the unique existence of the approximate solution is guaranteed by the Lax–Milgram theorem shown in Chap. 5. However, constraints due to the shape of the domain may come into play depending on the choice of basis function.

In order to remove such constraints, the finite element method is introduced to devise an appropriate selection of the basis function. Furthermore, error analysis is possible for approximate solutions using the finite element method. These results show that the finite element method is flexible for domain shape approximation and provides a good method for reliably guaranteeing the convergence to an exact solution if the divisions into elements are increased.

6.1 Galerkin Method

Here, we examine how the Galerkin method can be applied to approximate a solution for a boundary problems of elliptic partial differential equation. This will be illustrated by solving a one-dimensional Poisson problem and a $d \in \{2, 3\}$-dimensional Poisson problem.

6.1.1 One-Dimensional Poisson Problem

Consider the following one-dimensional Poisson problem with a mixed boundary condition.

Problem 6.1.1 (One-Dimensional Poisson problem) Let the functions $b : (0,1) \to \mathbb{R}$, $p_{\mathrm{N}} \in \mathbb{R}$ and $u_{\mathrm{D}} : (0,1) \to \mathbb{R}$ be given. Find $u : (0,1) \to \mathbb{R}$ such that

$$-\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = b \quad \text{in } (0,1), \quad \frac{\mathrm{d}u}{\mathrm{d}x}(1) = p_{\mathrm{N}}, \quad u(0) = u_{\mathrm{D}}(0).$$

□

With respect to Problem 6.1.1, set

$$U = \left\{ v \in H^1((0,1);\mathbb{R}) \;\middle|\; v(0) = 0 \right\}. \tag{6.1.1}$$

Moreover, with respect to $u, v \in U$, let

$$a(u,v) = \int_0^1 \frac{\mathrm{d}u}{\mathrm{d}x}\frac{\mathrm{d}v}{\mathrm{d}x}\,\mathrm{d}x, \tag{6.1.2}$$

$$l(v) = \int_0^1 bv\,\mathrm{d}x + p_{\mathrm{N}} v(1). \tag{6.1.3}$$

The weak form of this problem is as follows.

Problem 6.1.2 (Weak form of 1D Poisson problem) Let $a(\,\cdot\,,\,\cdot\,)$ and $l(\,\cdot\,)$ be given by Eq. (6.1.2) and Eq. (6.1.3), respectively. For given functions $b \in L^2((0,1);\mathbb{R})$, $p_{\mathrm{N}} \in \mathbb{R}$ and $u_{\mathrm{D}} \in H^1((0,1);\mathbb{R})$, obtain $u - u_{\mathrm{D}} \in U$ satisfying

$$a(u,v) = l(v) \tag{6.1.4}$$

with respect to an arbitrary $v \in U$. □

In the Galerkin method, approximate functions are constructed in the following way with respect to Problem 6.1.2.

Definition 6.1.3 (Set of approximate functions) Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)^{\top} \in U^m$ be $m \in \mathbb{N}$ known linearly independent functions. Let the set of approximate functions for U be

$$U_h = \left\{ v_h(\boldsymbol{\alpha}) = \sum_{i \in \{1,\ldots,m\}} \alpha_i \phi_i = \boldsymbol{\alpha} \cdot \boldsymbol{\phi} \;\middle|\; \boldsymbol{\alpha} \in \mathbb{R}^m \right\}$$

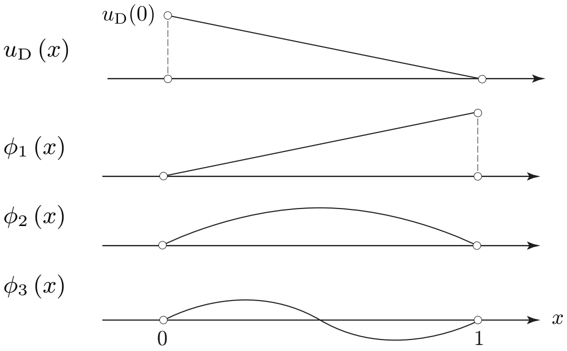


Fig. 6.1: An example of u_{D} and basis functions $\boldsymbol{\phi}$.

with $\boldsymbol{\alpha} = (\alpha_i)_i \in \mathbb{R}^m$ as undetermined multipliers. In this case, $\boldsymbol{\phi}$ is called a basis function. □

Figure 6.1 shows an example of u_{D} and $\boldsymbol{\phi}$. By using u_{D} and U_h defined in this way, the approximate functions of $u - u_{\mathrm{D}} \in U$ and $v \in U$ can be supposed to be $u_h - u_{\mathrm{D}} \in U_h$ and $v_h \in U_h$. If using Definition 4.2.6, this can be written as

$$
\begin{aligned}
u_h &= u_{\mathrm{D}} + \operatorname{span} \boldsymbol{\phi}, \\
v_h &= \operatorname{span} \boldsymbol{\phi}.
\end{aligned}
$$

In this case, U_h becomes a linear subspace (linear space) containing $\boldsymbol{\phi}$. Furthermore, if the inner product defined with respect to $H^1((0,1);\mathbb{R})$ is used, U_h becomes a Hilbert space. Here, the number of vectors which can be independently selected within U_h is m. In this case, U_h is a Hilbert space with m dimensions.

We have just defined the set of approximate functions U_h and u_{D}. Hence, using these functions we can now define the Galerkin method with respect to Problem 6.1.2 in the following way.

Definition 6.1.4 (Galerkin method) Let u_{D} and U_h be as in Definition 6.1.3. If $u_h(\boldsymbol{\alpha}) - u_{\mathrm{D}} \in U_h$ and $v_h(\boldsymbol{\beta}) \in U_h$ are substituted into $u - u_{\mathrm{D}} \in U$ and $v \in U$ of Eq. (6.1.4), simultaneous linear equations with $\boldsymbol{\alpha} \in \mathbb{R}^m$ as an unknown vector can be obtained. Using the solution $\boldsymbol{\alpha}$ to these equations, the method for obtaining the approximate solution of Problem 6.1.2 using $u_h(\boldsymbol{\alpha}) = u_{\mathrm{D}} + \boldsymbol{\phi} \cdot \boldsymbol{\alpha}$ is known as the Galerkin method. □

From the fact that U_h is a Hilbert space, if the Lax–Milgram theorem is applied, the solution by the Galerkin method $u_h(\boldsymbol{\alpha})$ can be said to exist uniquely. Actually, it is because $a(\,\cdot\,,\,\cdot\,)$ is bounded, $a(\,\cdot\,,\,\cdot\,)$ is coercive on $U_h \times U_h$ since U_h satisfies the homogeneous boundary condition, and $\hat{l}(\,\cdot\,)$ such as Exercise 5.2.5 is included in U_h'. Furthermore, $a(\,\cdot\,,\,\cdot\,)$ is also symmetric. The symmetry and coerciveness of $a\left(\,\cdot\,,\,\cdot\,\right)$ will appear later as the symmetry and positive definitiveness of coefficient matrix in simultaneous linear equations.

Let us look at how we can actually solve Problem 6.1.2 using the Galerkin method. Use u_{D} and U_h of Definition 6.1.3 to substitute $u_h - u_{\mathrm{D}} \in U_h$ and $v_h \in U_h$ into the weak form (Eq. (6.1.4)) and seek to find u_h such that

$$a\left(u_h, v_h\right) = l\left(v_h\right) \tag{6.1.5}$$

is satisfied with respect to an arbitrary $v_h \in U_h$. Both sides of Eq. (6.1.5) are respectively given by

$$\begin{aligned} a\left(u_h, v_h\right) &= \int_0^1 \frac{\mathrm{d}u_h}{\mathrm{d}x}\frac{\mathrm{d}v_h}{\mathrm{d}x}\,\mathrm{d}x, \\ l\left(v_h\right) &= \int_0^1 b v_h\,\mathrm{d}x + p_{\mathrm{N}} v_h\left(1\right). \end{aligned}$$

The terms in the integrand of $a\left(u_h, v_h\right)$ are given by

$$\begin{aligned} \frac{\mathrm{d}u_h}{\mathrm{d}x} &= \frac{\mathrm{d}u_{\mathrm{D}}}{\mathrm{d}x} + \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}x}\cdot\boldsymbol{\alpha} = \frac{\mathrm{d}u_{\mathrm{D}}}{\mathrm{d}x} + \begin{pmatrix} \frac{\mathrm{d}\phi_1}{\mathrm{d}x} & \cdots & \frac{\mathrm{d}\phi_m}{\mathrm{d}x} \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}, \\ \frac{\mathrm{d}v_h}{\mathrm{d}x} &= \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}x}\cdot\boldsymbol{\beta} = \begin{pmatrix} \frac{\mathrm{d}\phi_1}{\mathrm{d}x} & \cdots & \frac{\mathrm{d}\phi_m}{\mathrm{d}x} \end{pmatrix}\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}. \end{aligned}$$

Hence, $a\left(u_h, v_h\right)$ becomes

$$\begin{aligned} a\left(u_h, v_h\right) &= \int_0^1 \frac{\mathrm{d}u_h}{\mathrm{d}x}\frac{\mathrm{d}v_h}{\mathrm{d}x}\,\mathrm{d}x \\ &= \int_0^1 \begin{pmatrix} \beta_1 & \cdots & \beta_m \end{pmatrix}\begin{pmatrix} \frac{\mathrm{d}\phi_1}{\mathrm{d}x} \\ \vdots \\ \frac{\mathrm{d}\phi_m}{\mathrm{d}x} \end{pmatrix}\left(\frac{\mathrm{d}u_{\mathrm{D}}}{\mathrm{d}x} + \begin{pmatrix} \frac{\mathrm{d}\phi_1}{\mathrm{d}x} & \cdots & \frac{\mathrm{d}\phi_m}{\mathrm{d}x} \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix}\right)\mathrm{d}x \\ &= \int_0^1 \begin{pmatrix} \beta_1 & \cdots & \beta_m \end{pmatrix} \\ &\quad\times\left(\begin{pmatrix} \frac{\mathrm{d}\phi_1}{\mathrm{d}x}\frac{\mathrm{d}\phi_1}{\mathrm{d}x} & \cdots & \frac{\mathrm{d}\phi_1}{\mathrm{d}x}\frac{\mathrm{d}\phi_m}{\mathrm{d}x} \\ \vdots & \ddots & \vdots \\ \frac{\mathrm{d}\phi_m}{\mathrm{d}x}\frac{\mathrm{d}\phi_1}{\mathrm{d}x} & \cdots & \frac{\mathrm{d}\phi_m}{\mathrm{d}x}\frac{\mathrm{d}\phi_m}{\mathrm{d}x} \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} + \begin{pmatrix} \frac{\mathrm{d}u_{\mathrm{D}}}{\mathrm{d}x}\frac{\mathrm{d}\phi_1}{\mathrm{d}x} \\ \vdots \\ \frac{\mathrm{d}u_{\mathrm{D}}}{\mathrm{d}x}\frac{\mathrm{d}\phi_m}{\mathrm{d}x} \end{pmatrix}\right)\mathrm{d}x \\ &= \begin{pmatrix} \beta_1 & \cdots & \beta_m \end{pmatrix} \\ &\quad\times\left(\begin{pmatrix} a\left(\phi_1, \phi_1\right) & \cdots & a\left(\phi_1, \phi_m\right) \\ \vdots & \ddots & \vdots \\ a\left(\phi_m, \phi_1\right) & \cdots & a\left(\phi_m, \phi_m\right) \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} + \begin{pmatrix} a\left(u_{\mathrm{D}}, \phi_1\right) \\ \vdots \\ a\left(u_{\mathrm{D}}, \phi_m\right) \end{pmatrix}\right) \end{aligned}$$

$$= \boldsymbol{\beta} \cdot \left(\boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{a}_{\mathrm{D}} \right).$$

Here, we wrote $\boldsymbol{A} = (a_{ij})_{(i,j) \in \{1,\dots,m\}^2}$ and $\boldsymbol{a}_{\mathrm{D}} = (a_{\mathrm{D}i})_{i \in \{1,\dots,m\}}$, and let

$$a_{ij} = a\left(\phi_i, \phi_j\right) = \int_0^1 \frac{\mathrm{d}\phi_i}{\mathrm{d}x} \frac{\mathrm{d}\phi_j}{\mathrm{d}x} \, \mathrm{d}x, \tag{6.1.6}$$

$$a_{\mathrm{D}i} = a\left(u_{\mathrm{D}}, \phi_i\right) = \int_0^1 \frac{\mathrm{d}u_{\mathrm{D}}}{\mathrm{d}x} \frac{\mathrm{d}\phi_i}{\mathrm{d}x} \, \mathrm{d}x. \tag{6.1.7}$$

One the other hand, $l\left(v_h\right)$ becomes

$$\begin{aligned} l\left(v_h\right) &= \int_0^1 b v_h \, \mathrm{d}x + p_{\mathrm{N}} v_h\left(1\right) \\ &= \int_0^1 b \begin{pmatrix} \beta_1 & \cdots & \beta_m \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_m \end{pmatrix} \, \mathrm{d}x + p_{\mathrm{N}} \begin{pmatrix} \beta_1 & \cdots & \beta_m \end{pmatrix} \begin{pmatrix} \phi_1\left(1\right) \\ \vdots \\ \phi_m\left(1\right) \end{pmatrix} \\ &= \begin{pmatrix} \beta_1 & \cdots & \beta_m \end{pmatrix} \begin{pmatrix} l\left(\phi_1\right) \\ \vdots \\ l\left(\phi_m\right) \end{pmatrix} = \boldsymbol{\beta} \cdot \boldsymbol{l}. \end{aligned}$$

Here, we let $\boldsymbol{l} = (l_i)_{i \in \{1,\dots,m\}}$ and

$$l_i = l\left(\phi_i\right) = \int_0^1 b\phi_i \, \mathrm{d}x + p_{\mathrm{N}} \phi_i\left(1\right). \tag{6.1.8}$$

Therefore, Eq. (6.1.5) can be written as

$$\boldsymbol{\beta} \cdot \left(\boldsymbol{A}\boldsymbol{\alpha} + \boldsymbol{a}_{\mathrm{D}} \right) = \boldsymbol{\beta} \cdot \boldsymbol{l}.$$

Here, consider an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^m$ to get

$$\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{l} - \boldsymbol{a}_{\mathrm{D}} = \hat{\boldsymbol{l}}. \tag{6.1.9}$$

For confirmation, write the elements of vector and matrix of Eq. (6.1.9) to get

$$\begin{pmatrix} a\left(\phi_1, \phi_1\right) & \cdots & a\left(\phi_1, \phi_m\right) \\ \vdots & \ddots & \vdots \\ a\left(\phi_m, \phi_1\right) & \cdots & a\left(\phi_m, \phi_m\right) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{pmatrix} = \begin{pmatrix} l\left(\phi_1\right) - a\left(u_{\mathrm{D}}, \phi_1\right) \\ \vdots \\ l\left(\phi_m\right) - a\left(u_{\mathrm{D}}, \phi_m\right) \end{pmatrix}.$$

$\boldsymbol{A}$ is referred to as a coefficient matrix and $\hat{\boldsymbol{l}}$ is a known term vector. $\boldsymbol{A}$ is symmetric from $a\left(\phi_i, \phi_j\right) = a\left(\phi_j, \phi_i\right)$. Moreover, due to the coerciveness of $a\left(\,\cdot\,,\,\cdot\,\right)$ at U_h, there exists some $c_0 > 0$ and

$$\boldsymbol{\beta} \cdot \left(\boldsymbol{A}\boldsymbol{\beta} \right) = a\left(\boldsymbol{v}_h, \boldsymbol{v}_h\right) \geq c_0 \left\| \boldsymbol{\beta} \right\|_{\mathbb{R}^m}^2$$

holds with respect to an arbitrary $\boldsymbol{\beta} \in \mathbb{R}^m$. Therefore, $\boldsymbol{A}$ becomes a positive definite symmetric matrix (hence a regular matrix) and the inverse matrix of $\boldsymbol{A}$ can be taken to calculate $\boldsymbol{\alpha}$ by

$$\boldsymbol{\alpha} = \boldsymbol{A}^{-1}\hat{\boldsymbol{l}}. \tag{6.1.10}$$

The approximate solution u_h can be obtained via

$$u_h = u_{\mathrm{D}} + \boldsymbol{\phi} \cdot \boldsymbol{\alpha} \tag{6.1.11}$$

by using this $\boldsymbol{\alpha}$.

Based on the detail above, basis function u_{D} and $\boldsymbol{\phi}$ should be selected so that the calculation of $\int_0^1 (\mathrm{d}\phi_i/\mathrm{d}x)(\mathrm{d}\phi_j/\mathrm{d}x)\,\mathrm{d}x$ is easily possible. Furthermore, if selection can be made so that the derivatives of the basis function are mutually orthogonal, $\boldsymbol{A}$ becomes a diagonal matrix and the calculation of the inverse matrix becomes simple.

We solve the following problem using the Galerkin method.

Exercise 6.1.5 (Galerkin method for 1D Dirichlet problem) Let the basis functions be

$$\boldsymbol{\phi} = \begin{pmatrix}\phi_1 & \cdots & \phi_m\end{pmatrix}^\top = \begin{pmatrix}\sin(1\pi x) & \cdots & \sin(m\pi x)\end{pmatrix}^\top$$

and use the Galerkin method in order to obtain the approximate solution $u_h : (0,1) \to \mathbb{R}$ satisfying

$$-\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = 1 \quad \text{in } (0,1), \quad u(0) = 0, \quad u(1) = 0.$$

□

Answer Let $U = H_0^1((0,1);\mathbb{R})$ and write the weak form of this problem as

$$a(u,v) = l_1(v)$$

with respect to an arbitrary $v \in U$. Here, let $a(\,\cdot\,,\,\cdot\,)$ be Eq. (6.1.2) and $l_1(\,\cdot\,)$ be $l(\,\cdot\,)$ on Eq. (6.1.3) when $b = 1$ and $p = 0$. Let the approximate function with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta} \in \mathbb{R}^m$ be

$$u_h = \boldsymbol{\alpha} \cdot \boldsymbol{\phi}(x), \quad v_h = \boldsymbol{\beta} \cdot \boldsymbol{\phi}(x).$$

Substituting these into the weak form gives

$$\boldsymbol{\beta} \cdot (\boldsymbol{A}\boldsymbol{\alpha}) = \boldsymbol{\beta} \cdot \boldsymbol{l}_1.$$

Here, let $\boldsymbol{A} = (a_{ij})_{(i,j)\in\{1,\ldots,m\}^2}$ and $\boldsymbol{l}_1 = (l_{1i})_{i\in\{1,\ldots,m\}}$ to get

$$\begin{aligned} a_{ij} = a(\phi_i, \phi_j) &= \int_0^1 \frac{\mathrm{d}\phi_i}{\mathrm{d}x}\frac{\mathrm{d}\phi_j}{\mathrm{d}x}\ \mathrm{d}x = ij\pi^2 \int_0^1 \cos(i\pi x)\cos(j\pi x)\ \mathrm{d}x \\ &= \frac{1}{2}ij\pi^2 \int_0^1 [\cos\{(i+j)\pi x\} + \cos\{(i-j)\pi x\}]\,\mathrm{d}x = \frac{1}{2}ij\pi^2\delta_{ij}, \end{aligned}$$

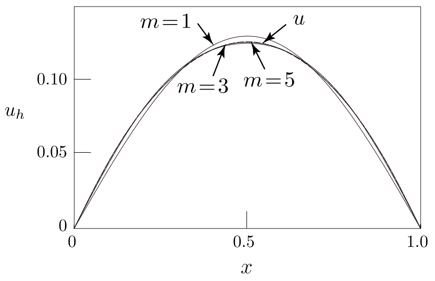


Fig. 6.2: Exact and approximate solutions of Exercise 6.1.5.

$$l_{1i} = \int_0^1 \sin(i\pi x)\,\mathrm{d}x = \frac{1}{i\pi}\left[-\cos(i\pi x)\right]_0^1 = \frac{(-1)^{i+1}+1}{i\pi},$$

where

$$\delta_{ij} = \begin{cases} 1 \ (i = j) \\ 0 \ (i \neq j) \end{cases}$$

represents the Kronecker delta. Hence, $\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{l}_1$ becomes

$$\frac{\pi^2}{2}\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 4 & 0 & \cdots & 0 \\ 0 & 0 & 9 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & m^2 \end{pmatrix}\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_m \end{pmatrix} = \frac{1}{\pi}\begin{pmatrix} 2 \\ 0 \\ 2/3 \\ \vdots \\ \dfrac{(-1)^{m+1}+1}{m} \end{pmatrix}.$$

Solving these simultaneous linear equations gives

$$\alpha_i = \frac{2\left\{(-1)^{i+1}+1\right\}}{i^3\pi^3}.$$

Therefore, an approximate solution becomes

$$u_h = \sum_{i\in\{1,\ldots,m\}} \frac{2\left\{(-1)^{i+1}+1\right\}}{i^3\pi^3}\sin(i\pi x).$$

On the other hand, the exact solution is

$$u = \frac{1}{2}x(x-1).$$

Figure 6.2 shows the comparison between the approximate solution and the exact solution. □

6.1.2 d-Dimensional Poisson Problem

Next, to think about solving the $d \in \{2,3\}$-dimensional Poisson problem using the Galerkin method, we revisit Problem 5.1.1.

Problem 6.1.6 (d-Dimensional Poisson problem) When $b : \Omega \to \mathbb{R}$, $p_{\mathrm{N}} : \Gamma_{\mathrm{N}} \to \mathbb{R}$ and $u_{\mathrm{D}} : \Omega \to \mathbb{R}$ are given, obtain $u : \Omega \to \mathbb{R}$ which satisfies

$$-\Delta u = b \quad \text{in } \Omega, \quad \frac{\partial u}{\partial \nu} = p_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}, \quad u = u_{\mathrm{D}} \quad \text{on } \Gamma_{\mathrm{D}}.$$

□

With respect to Problem 6.1.6, let

$$U = \left\{ u \in H^1\left(\Omega; \mathbb{R}\right) \;\middle|\; u = 0 \text{ on } \Gamma_{\mathrm{D}} \right\}. \tag{6.1.12}$$

Moreover, with respect to $u, v \in U$, let

$$a\left(u, v\right) = \int_\Omega \boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} v \,\mathrm{d}x, \tag{6.1.13}$$

$$l\left(v\right) = \int_\Omega b v \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} p_{\mathrm{N}} v \,\mathrm{d}\gamma. \tag{6.1.14}$$

The weak form of this problem becomes as follows.

Problem 6.1.7 (Weak form of d-Dimensional Poisson problem) Given $b \in L^2\left(\Omega; \mathbb{R}\right)$, $p_{\mathrm{N}} \in L^2\left(\Gamma_{\mathrm{N}}; \mathbb{R}\right)$ and $u_{\mathrm{D}} \in H^1\left(\Omega; \mathbb{R}\right)$, obtain $u - u_{\mathrm{D}} \in U$ such that

$$a\left(u, v\right) = l\left(v\right) \tag{6.1.15}$$

is satisfied with respect to an arbitrary $v \in U$. □

Construct approximate functions with respect to Problem 6.1.6 in the following way. Let m be a natural number.

Definition 6.1.8 (Set of approximate functions) Let $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_m\right)^\top \in U^m$ be m known functions of linear independence. Let

$$U_h = \left\{ v_h\left(\boldsymbol{\alpha}\right) = \sum_{i \in \{1,\ldots,m\}} \alpha_i \phi_i = \boldsymbol{\alpha} \cdot \boldsymbol{\phi} \;\middle|\; \boldsymbol{\alpha} \in \mathbb{R}^m \right\}$$

be the set of approximate functions with respect to U with $\boldsymbol{\alpha} = \left(\alpha_i\right)_i \in \mathbb{R}^m$ as unknown multipliers. Here, $\boldsymbol{\phi}$ is called basis functions. □

Comparing Definition 6.1.3 and Definition 6.1.8 shows that the same expression is being used other than the defined domain of the function being changed. Hence, the same set of equations used for the one-dimensional Poisson problem can be used for the d-dimensional Poisson problem. This is confirmed as follows. When $u_h - u_{\mathrm{D}} \in U_h$ and $v_h \in U_h$ with u_{D} and U_h in Definition 6.1.8 are substituted into the weak form (Eq. (6.1.15)),

$$a\left(u_h, v_h\right) = l\left(v_h\right) \tag{6.1.16}$$

becomes the same as Eq. (6.1.5). Here, the definitions of $a\left(\,\cdot\,,, \,\cdot\,\right)$ and $l\left(\,\cdot\,\right)$ are replaced by

$$\begin{aligned} a\left(u_h, v_h\right) &= \int_\Omega \boldsymbol{\nabla} u_h \cdot \boldsymbol{\nabla} v_h \;\mathrm{d}x, \\ l\left(v_h\right) &= \int_\Omega b v_h \;\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} p_{\mathrm{N}} v_h \;\mathrm{d}\gamma. \end{aligned}$$

After this, the same expansion of equations used for the one-dimensional Poisson problem can be performed to obtain

$$\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{l} - \boldsymbol{a}_{\mathrm{D}} = \hat{\boldsymbol{l}} \tag{6.1.17}$$

which coincides with Eq. (6.1.9), where $\boldsymbol{A} = \left(a_{ij}\right)_{(i,j)\in\{1,\ldots,m\}^2}$, $\boldsymbol{a}_{\mathrm{D}} = \left(a_{\mathrm{D}i}\right)_{i\in\{1,\ldots,m\}}$ and $\boldsymbol{l} = \left(l_i\right)_{i\in\{1,\ldots,m\}}$ are changed by

$$a_{ij} = a\left(\phi_i, \phi_j\right) = \int_\Omega \boldsymbol{\nabla}\phi_i \cdot \boldsymbol{\nabla}\phi_j \;\mathrm{d}x, \tag{6.1.18}$$

$$a_{\mathrm{D}i} = a\left(u_{\mathrm{D}}, \phi_j\right) = \int_\Omega \boldsymbol{\nabla} u_{\mathrm{D}} \cdot \boldsymbol{\nabla}\phi_j \;\mathrm{d}x, \tag{6.1.19}$$

$$l_i = l\left(\phi_i\right) = \int_\Omega b\phi_i \;\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} p_{\mathrm{N}}\phi_i \;\mathrm{d}\gamma, \tag{6.1.20}$$

respectively. Since $\boldsymbol{A}$ becomes a positive definite symmetric matrix, and the inverse matrix of $\boldsymbol{A}$ can be taken, Eq. (6.1.10) and Eq. (6.1.11) remain in effect.

We solve in the following exercise a two-dimensional Dirichlet problem using the Galerkin method with respect to a square domain. Looking at this it is apparent that the notations of Eq. (6.1.10) and Eq. (6.1.11) are established ([3, Exercise 3.2, p. 30]).

Exercise 6.1.9 (Galerkin method for 2D Dirichlet problem) Setting basis functions as

$$\boldsymbol{\phi} = \left(\phi_{ij}\left(\boldsymbol{x}\right)\right)_{(i,j)\in\{1,\ldots,m\}^2} = \left(\sin\left(i\pi x_1\right)\sin\left(j\pi x_2\right)\right)_{(i,j)\in\{1,\ldots,m\}^2},$$

use the Galerkin method in order to obtain the approximate solution $u_h : (0,1)^2 \to \mathbb{R}$ which satisfies

$$-\Delta u = 1 \quad \text{in } \Omega = (0,1)^2, \quad u = 0 \quad \text{on } \partial\Omega.$$

□

Answer Let $U = H_0^1\left((0,1)^2;\mathbb{R}\right)$ and denote the weak form of this problem as

$$a(u,v) = l_1(v)$$

with respect to an arbitrary $v \in U$, where we let $a(\,\cdot\,,\,\cdot\,)$ be as in Eq. (6.1.13) and $l_1(\,\cdot\,)$ be $l(\,\cdot\,)$ in Eq. (6.1.14) when $b = 1$ and $p = 0$. Let the approximate functions be

$$\begin{aligned} u_h &= \boldsymbol{\alpha}\cdot\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{(i,j)\in\{1,\dots,m\}^2} \alpha_{ij}\phi_{ij}(\boldsymbol{x}), \\ v_h &= \boldsymbol{\beta}\cdot\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{(i,j)\in\{1,\dots,m\}^2} \beta_{ij}\phi_{ij}(\boldsymbol{x}) \end{aligned}$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} \in \mathbb{R}^{m\times m}$. Substituting u_h and v_h into the weak form gives

$$\boldsymbol{\beta}\cdot(\boldsymbol{A}\boldsymbol{\alpha}) = \boldsymbol{\beta}\cdot\boldsymbol{l}_1.$$

Here, if we write $\boldsymbol{A} = (a(\phi_{ij},\phi_{kl}))_{(i,j,k,l)\in\{1,\dots,m\}^4}$ and $\boldsymbol{l}_1 = (l_1(\phi_{ij}))_{(i,j)\in\{1,\dots,m\}^2}$,

$$\begin{aligned} a(\phi_{ij},\phi_{kl}) &= \int_0^1\int_0^1 \left(\frac{\partial\phi_{ij}}{\partial x_1}\frac{\partial\phi_{kl}}{\partial x_1} + \frac{\partial\phi_{ij}}{\partial x_2}\frac{\partial\phi_{kl}}{\partial x_2}\right)\,\mathrm{d}x_1\mathrm{d}x_2 \\ &= \int_0^1\int_0^1 \left\{ki\pi^2\cos(k\pi x_1)\sin(l\pi x_2)\cos(i\pi x_1)\sin(j\pi x_2)\right. \\ &\qquad \left.+li\pi^2\sin(k\pi x_1)\cos(l\pi x_2)\sin(i\pi x_1)\cos(j\pi x_2)\right\}\,\mathrm{d}x_1\mathrm{d}x_2 \\ &= \frac{\pi^2}{4}(ki+lj)\,\delta_{ki}\delta_{lj}, \\ l_1(\phi_{ij}) &= \int_0^1\int_0^1 \sin(i\pi x_1)\sin(j\pi x_2)\,\mathrm{d}x_1\mathrm{d}x_2 \\ &= \frac{\left\{(-1)^{i+1}+1\right\}\left\{(-1)^{j+1}+1\right\}}{ij\pi^2} \end{aligned}$$

is obtained. Then $\boldsymbol{A}\boldsymbol{\alpha} = \boldsymbol{l}_1$ becomes

$$\sum_{(k,l)\in\{1,\dots,m\}^2} \frac{\pi^2}{4}(ki+lj)\,\delta_{ki}\delta_{lj}\alpha_{ij} = \frac{\left\{(-1)^{i+1}+1\right\}\left\{(-1)^{j+1}+1\right\}}{ij\pi^2}.$$

Solving these simultaneous first-order equations gives

$$\alpha_{ij} = \frac{4\left\{(-1)^{i+1}+1\right\}\left\{(-1)^{j+1}+1\right\}}{ij\left(i^2+j^2\right)\pi^4}$$

with respect to $i,j \in \{1,\dots,m\}$. Therefore, the approximate solution u_h becomes

$$u_h = \sum_{(i,j)\in\{1,\dots,m\}^2} \frac{4\left\{(-1)^{i+1}+1\right\}\left\{(-1)^{j+1}+1\right\}}{ij\left(i^2+j^2\right)\pi^4}\sin(i\pi x)\sin(j\pi x). \quad (6.1.21)$$

□

6.1.3 Ritz Method

In the Galerkin method, an approximate function was substituted into the weak form. However, the same approximate solution can also be obtained when an approximate function is substituted into a minimization problem. This method is called the Ritz method.

Take a look at the next problem rewriting Problem 6.1.7 as a minimization problem. $a(\,\cdot\,,\,\cdot\,)$, $l\,(\,\cdot\,)$, u_{D} and U are taken to be the same as those used in Problem 6.1.7.

Problem 6.1.10 (Minimization problem of dD Poisson problem) With respect to $b \in L^2\left(\Omega;\mathbb{R}\right)$, $p_{\mathrm{N}} \in L^2\left(\Gamma_{\mathrm{N}};\mathbb{R}\right)$ and $u_{\mathrm{D}} \in H^1\left(\Omega;\mathbb{R}\right)$, obtain a u which satisfies

$$\min_{u-u_{\mathrm{D}}\in U}\left\{f\left(u\right)=\frac{1}{2}a\left(u,u\right)-l\left(u\right)\right\}.$$

□

The Ritz method is defined as follows with respect to Problem 6.1.10.

Definition 6.1.11 (Ritz method) Let U_h be as in Definition 6.1.8. If $u_h\left(\boldsymbol{\alpha}\right) - u_{\mathrm{D}} \in U_h$ is substituted into $u - u_{\mathrm{D}} \in U$ in Problem 6.1.10, the stationary condition of $f\left(u\right)$ with respect to a variation of $\boldsymbol{\alpha} \in \mathbb{R}^m$ can be obtained as simultaneous linear equations with respect to $\boldsymbol{\alpha}$. The method for obtaining an approximate solution of Problem 6.1.10 using the solution $\boldsymbol{\alpha}$ via $u_h\left(\boldsymbol{\alpha}\right) = u_{\mathrm{D}} + \boldsymbol{\phi}\cdot\boldsymbol{\alpha}$ is called the Ritz method. □

We solve Problem 6.1.10 using the Ritz method. Substituting u_h into $f\left(u_h\right)$ gives

$$f\left(u_h\right)=\frac{1}{2}\left\{a\left(u_{\mathrm{D}},u_{\mathrm{D}}\right)+\boldsymbol{\alpha}\cdot\left(\boldsymbol{A}\boldsymbol{\alpha}\right)+2\boldsymbol{\alpha}\cdot\boldsymbol{a}_{\mathrm{D}}\right\}-\left(l\left(u_{\mathrm{D}}\right)+\boldsymbol{\alpha}\cdot\boldsymbol{l}\right),$$

where $\boldsymbol{A}$, $\boldsymbol{a}_{\mathrm{D}}$ and $\boldsymbol{l}$ are Eq. (6.1.18), Eq. (6.1.19) and Eq. (6.1.20) respectively.

From the minimum conditions of $f\left(u_h\right)$,

$$\frac{\partial f\left(u_h\right)}{\partial\boldsymbol{\alpha}}=\boldsymbol{A}\boldsymbol{\alpha}+\boldsymbol{a}_{\mathrm{D}}-\boldsymbol{l}=\boldsymbol{0}_{\mathbb{R}^m}$$

is obtained. This equation matches Eq. (6.1.17) of the Galerkin method.

The fact that the Ritz method just requires the construction of an approximate function with respect to u means it has an advantage in a sense that the theory can be compactly contained. However, compared with the Galerkin method, it should be noted that its use is limited to when the boundary value problem of elliptic partial differential equation can be rewritten as a minimization problem, namely $a(\,\cdot\,,\,\cdot\,)$ is symmetric, as seen in Section 5.2. Moreover, when $a(\,\cdot\,,\,\cdot\,)$ is symmetric, the two methods can be put together because the equation obtained from the Galerkin method is the same as that of the Ritz method, and referred to as the Ritz–Galerkin method.

6.1.4 Basic Error Estimation

The existence of a unique solution $u_h(\boldsymbol{\alpha})$ by the Galerkin method was guaranteed by the Lax–Milgram theorem from the fact that the set of approximate functions U_h is a Hilbert space. Furthermore, if using the norm of the Hilbert space containing the exact solution, clear results can be obtained regarding the stability and errors of the approximate solution.

Here, with respect to Problem 6.1.1,

$$\begin{aligned} V &= H^1((0,1);\mathbb{R}), \\ U &= \{v \in V \mid v(0) = 0\}, \\ U(u_{\mathrm{D}}) &= \{v \in V \mid v - u_{\mathrm{D}} \in U,\ u_{\mathrm{D}} \in V\} \end{aligned}$$

is set. With respect to Problem 6.1.6, let

$$\begin{aligned} V &= H^1(\Omega;\mathbb{R}), \\ U &= \{v \in V \mid v = 0 \text{ on } \Gamma_{\mathrm{D}}\}, \\ U(u_{\mathrm{D}}) &= \{v \in V \mid v - u_{\mathrm{D}} \in U,\ u_{\mathrm{D}} \in V\}. \end{aligned}$$

Let V' be the dual space of V. Moreover, the set of approximate functions U_h is as per Definition 6.1.3 with respect to Problem 6.1.1 and Definition 6.1.8 with respect to Problem 6.1.6. Furthermore, the approximate function of u_{D} has not been thought about but here, its approximate function is set to be $u_{\mathrm{D}h}$ and written as

$$U_h(u_{\mathrm{D}h}) = \{u_h \in V \mid h_h - u_{\mathrm{D}h} \in U_h\}.$$

One of the results is like the one below relating to the effect that the error in the given functions has on the approximate solution (cf. [1, Remark 1.2, p. 30], [2, Theorem 2.3, p. 34]).

Theorem 6.1.12 (Stability of approximate solution) Let $a : V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form. Let $u_{h1}, u_{h2} \in U_h(u_{\mathrm{D}h}) \subset U(u_{\mathrm{D}})$ be the approximate solutions by the Galerkin method of Problem 6.1.2 or Problem 6.1.7 with respect to arbitrary $l_1, l_2 \in V'$ respectively. In this case,

$$\|u_{h1} - u_{h2}\|_V \le \frac{1}{\alpha} \|l_1 - l_2\|_{V'}$$

is established. Here, $\alpha > 0$ is a constant representing the coerciveness of $a(\,\cdot\,,\,\cdot\,)$ (Definition 5.2.1). □

The other result shows that the approximate solution of the Galerkin method is the best (closest to the exact solution) element among the set of approximate solutions $U_h(u_{\mathrm{D}h})$. Here, the distance between the exact solution $u \in U(u_{\mathrm{D}})$ and $u_h \in U_h(u_{\mathrm{D}h})$ will be measured with $\sqrt{a(u - u_h, u - u_h)}$ or the norm $\|u - u_h\|_V$ in V. The result is called Cea's lemma (cf. [1, Theorem 13.1, p. 113], [5, Lemma 2.3, p. 54], [2, Theorem 2.4, p. 42]).

Theorem 6.1.13 (Basic error estimation) Let $u \in U(u_{\mathrm{D}})$ be the solution to Problem 6.1.2 or Problem 6.1.7 with respect to an arbitrary $l \in V'$, and $u_h \in U_h(u_{\mathrm{D}h})$ be the approximate solution from the Galerkin method. In this case,

$$a(u - u_h, u - u_h) \leq \inf_{v_h \in U_h(u_{\mathrm{D}h})} a(u - v_h, u - v_h),$$

$$\|u - u_h\|_V \leq \sqrt{\frac{\|a\|}{\alpha}} \inf_{v_h \in U_h(u_{\mathrm{D}h})} \|u - v_h\|_V + \left(1 + \sqrt{\frac{\|a\|}{\alpha}}\right) \|u_{\mathrm{D}} - u_{\mathrm{D}h}\|_V$$

is established. Here, $\|a\|$ is the norm of bilinear operator (Section 4.4.4) and $\alpha > 0$ is a constant providing the coerciveness of $a(\cdot, \cdot)$. □

Theorem 6.1.13 shows that the approximate solution from the Galerkin method is the best one out of U_h. Hence, it indicates that in order to reduce the error of approximate solution from the Galerkin method, it is effective to keep the approximate functions with the ability to be close to the exact solution within U_h. This result is also used when conducting error estimation of numerical analyses based on the Galerkin method. The error estimation with respect to finite element methods is examined in detail in Sect. 6.6.

6.2 One-Dimensional Finite Element Method

An approximate function is constructed as a linear combination of basis functions in the Galerkin method and the undetermined multipliers are found by substituting them into the weak form. Let us consider how to choose the basis function without changing this framework.

In the Galerkin method seen in Sect. 6.1, the basis functions were selected from the functions defined across the whole domain. In Exercise 6.1.9, as the basis functions, $(\sin(i\pi x_1)\sin(j\pi x_2))_{(i,j)\in\{1,\ldots,m\}^2}$ were chosen. These functions have the support on the domain $\Omega = (0,1)^2$ of the boundary value problem of partial differential equation. As long as the basis functions are chosen in this way, the shape of the domain on which the boundary value problem is defined will be limited to be a rectangle as shown in Fig. 6.3 (a), or an ellipse.

In contrast, think about constructing U_h using basis functions which have supports on simple triangular domains $\{\Omega_i\}_i$ which are formed by splitting a polygon domain Ω such as the one in Fig. 6.3 (b). This may enable an approximate solution of the boundary value problem defined over an arbitrary polygon shape to be obtained just by changing the way the basis functions are chosen without the need to change the framework of the Galerkin method. The Galerkin method obtained with these principles is the finite element method. The simple domain chosen in this case is called a finite element.

This section examines in detail the process for solving a one-dimensional Poisson problem using the finite element method. First, define the approximate functions used in the finite element method within the framework of the Galerkin

(a) Ω: Galerkin method in Sect. 6.1. (b) $\{\Omega_i\}_i$: Finite element method.

Fig. 6.3: Supports of approximate functions.

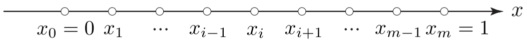


Fig. 6.4: Finite elements and nodes within a one-dimensional domain $\Omega = (0, 1)$.

method. Then suppose these approximate functions are defined for the split domains of each finite element. From this, the integration of the weak form can be replaced by the sum of integrations for each finite element domain.

6.2.1 Approximate Functions in Galerkin Method

Consider the finite element method with respect to a one-dimensional Poisson problem (Problem 6.1.1 and its weak form Problem 6.1.2). In the finite element method, the domain $\Omega = (0, 1)$ is split into (x_0, x_1), (x_1, x_2), …, (x_{m-1}, x_m) as in Fig. 6.4. Here, x_0, x_1, … , x_m are called nodes and (x_0, x_1), (x_1, x_2), … , (x_{m-1}, x_m) are called one-dimensional finite elements or the domains of finite elements. The finite elements are numbered and their set represented as $\mathcal{E} = \{1, \dots, m\}$. Moreover, nodes are also numbered and the set of numbers expressed as $\mathcal{N} = \{0, \dots, m\}$.

Select the basis functions in the one-dimensional finite element method with respect to Problem 6.1.1 as a pyramid-shaped function with unit height such as in Fig. 6.5 and defined as

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{x_1 - x_0} & \text{in } (0, x_1) \\ 0 & \text{in } (x_1, 1) \end{cases},$$

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}} & \text{in } (x_{i-1}, x_i) \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i} & \text{in } (x_i, x_{i+1}) \\ 0 & \text{in } (0, x_{i-1}) \cup (x_{i+1}, 1) \end{cases}, \quad \text{for } i \in \mathcal{N},$$

$$\phi_m(x) = \begin{cases} \dfrac{x - x_{m-1}}{x_m - x_{m-1}} & \text{in } (x_{m-1}, 1) \\ 0 & \text{in } (0, x_{m-1}) \end{cases}.$$

Approximate functions are constructed as

$$u_h(\bar{\boldsymbol{u}}) = u_0 \phi_0 + \sum_{i \in \{1, \dots, m\}} u_i \phi_i = \begin{pmatrix} \bar{u}_{\mathrm{D}} \\ \bar{\boldsymbol{u}}_{\mathrm{N}} \end{pmatrix} \cdot \begin{pmatrix} \phi_{\mathrm{D}} \\ \boldsymbol{\phi}_{\mathrm{N}} \end{pmatrix} = \bar{\boldsymbol{u}} \cdot \boldsymbol{\phi}, \tag{6.2.1}$$

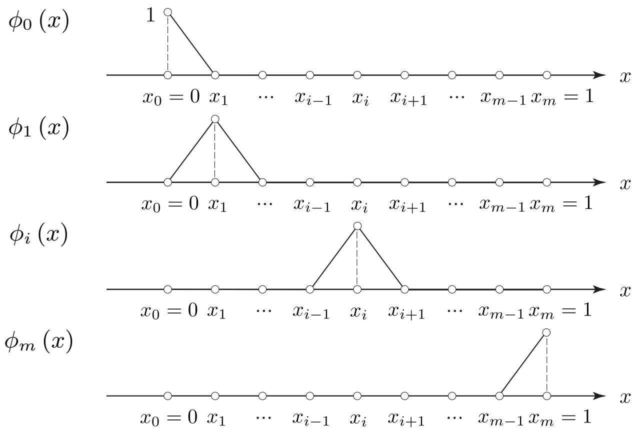

$\phi_0(x)$
1
x
$x_0 = 0$ x_1 $\cdots$ x_{i-1} x_i x_{i+1} $\cdots$ x_{m-1} $x_m = 1$
$\phi_1(x)$
$\phi_i(x)$
$\phi_m(x)$

Fig. 6.5: Basis functions $\phi_0, \ldots, \phi_m$ used in the one-dimensional finite element method.

$$v_h(\bar{\boldsymbol{v}}) = v_0\phi_0 + \sum_{i \in \{1,\ldots,m\}} v_i\phi_i = \begin{pmatrix} \bar{v}_{\mathrm{D}} \\ \bar{\boldsymbol{v}}_{\mathrm{N}} \end{pmatrix} \cdot \begin{pmatrix} \phi_{\mathrm{D}} \\ \boldsymbol{\phi}_{\mathrm{N}} \end{pmatrix} = \bar{\boldsymbol{v}} \cdot \boldsymbol{\phi}, \tag{6.2.2}$$

where $u_0 = \bar{u}_{\mathrm{D}} = u_{\mathrm{D}}$ and $v_0 = \bar{v}_{\mathrm{D}} = 0$. Here, $\bar{\boldsymbol{u}}_{\mathrm{N}} = (u_1, \ldots, u_m)^\top$ and $\bar{\boldsymbol{v}}_{\mathrm{N}} = (v_1, \ldots, v_m)^\top$ are undetermined multipliers. In this expression, $(\bar{\cdot})$ was included to indicate a vector, but $\bar{u}_{\mathrm{D}}$ and $\bar{v}_{\mathrm{D}}$ are one-dimensional vectors for Problem 6.1.1 in which the fundamental boundary condition was given at one node.

Take a look at the characteristics of the approximate functions defined above. First, the fact that the basis functions are continuous functions means that those are included in $H^1(\Omega; \mathbb{R})$ (Sobolev embedding theorem (Theorem 4.3.14)) and satisfy the requirements for substituting into the weak form. The fact that those are first-order polynomials in the finite elements represents that the evaluation of derivatives of ϕ_i and ϕ_j which appear in $a(\phi_i, \phi_j)$ becomes easier. Moreover, the basis functions defined for each node will have the supports only on the finite elements adjacent to the node, hence the domain of integration of $a(\phi_i, \phi_j)$ is limited to each of their finite elements. Furthermore, the unknown multiplier u_i with respect to basis function ϕ_i which is 1 at the node $i \in \mathcal{N}$ matches the node value of approximate function as in Fig. 6.6. From this, $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ are called nodal value vectors. In this book, the elements $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ are split into two types, $\bar{u}_{\mathrm{D}} = u_0$ and $\bar{v}_{\mathrm{D}} = v_0$ providing the fundamental boundary condition, called Dirichlet-type nodal value vectors (real numbers in this case) and $\bar{\boldsymbol{u}}_{\mathrm{N}} = (u_1, \ldots, u_m)^\top$ and $\bar{\boldsymbol{v}}_{\mathrm{N}} = (v_1, \ldots, v_m)^\top$, called Neumann-type nodal value vectors.

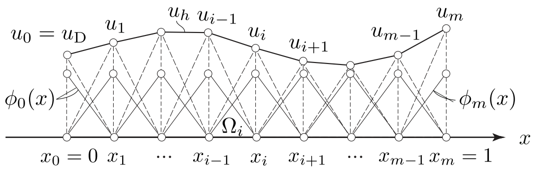


Fig. 6.6: Node value vector $\bar{\boldsymbol{u}}$ in the 1D finite element method.

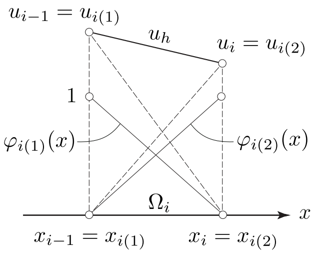


Fig. 6.7: Basis functions $\varphi_{i(1)}$, $\varphi_{i(2)}$ on a finite element in the 1D finite element method.

6.2.2 Approximate Functions in Finite Element Method

So far, based on the awareness that the finite element method is one form of the Galerkin method, the domain of an approximate function has been taken to be Ω. However, if we focus on the set of basis functions with support on the domain $\Omega_i = (x_{i-1}, x_i)$ for the finite element $i \in \mathcal{E}$, two basis functions $\phi_{i-1}(x)$ and $\phi_i(x)$ with respect to nodes $i-1 \in \mathcal{N}$ and $i \in \mathcal{N}$ such as those shown in Fig. 6.7 can be used by rewriting as

$$\varphi_{i(1)}(x) = \phi_{i-1}(x) = \frac{x_{i(2)} - x}{x_{i(2)} - x_{i(1)}}, \tag{6.2.3}$$

$$\varphi_{i(2)}(x) = \phi_i(x) = \frac{x - x_{i(1)}}{x_{i(2)} - x_{i(1)}}, \tag{6.2.4}$$

to define the approximate function as

$$u_h(\bar{\boldsymbol{u}}_i) = \begin{pmatrix} \varphi_{i(1)} & \varphi_{i(2)} \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix} = \boldsymbol{\varphi}_i \cdot \bar{\boldsymbol{u}}_i, \tag{6.2.5}$$

$$v_h(\bar{\boldsymbol{v}}_i) = \begin{pmatrix} \varphi_{i(1)} & \varphi_{i(2)} \end{pmatrix} \begin{pmatrix} v_{i(1)} \\ v_{i(2)} \end{pmatrix} = \boldsymbol{\varphi}_i \cdot \bar{\boldsymbol{v}}_i \tag{6.2.6}$$

on Ω_i with respect to all $i \in \mathcal{E}$. In this case, $\boldsymbol{\varphi}_i(x) = (\varphi_{i-1}(x), \varphi_i(x))^\top = (\varphi_{i(1)}(x), \varphi_{i(2)}(x))^\top$ is referred to as basis functions in a finite element. Moreover, in technical books on the finite element method, $\boldsymbol{\varphi}_i(x)$ is generally referred to as a shape function or an interpolation function. However, given that the main topic of this book is the shape optimization problem, the term shape function may bring some confusion. Hence, in this book $\boldsymbol{\varphi}_i(x)$ will be referred to as a basis function in the finite element or basis function when there is no danger of confusion. Moreover, in Eq. (6.2.5) and Eq. (6.2.6),

$$\bar{\boldsymbol{u}}_i = \begin{pmatrix} u_{i-1} \\ u_i \end{pmatrix} = \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix}, \quad \bar{\boldsymbol{v}}_i = \begin{pmatrix} v_{i-1} \\ v_i \end{pmatrix} = \begin{pmatrix} v_{i(1)} \\ v_{i(2)} \end{pmatrix}$$

is called the element nodal value vector of u and v with respect to finite element $i \in \mathcal{E}$. Hereafter, by using notation $(\,\cdot\,)_{i(\alpha)}$ as a function or a value corresponding to the local node number $\alpha \in \{1,2\}$ at the finite element $i \in \mathcal{E}$, $u_{i(\alpha)}$ and $v_{i(\alpha)}$ will be referred to as a local node number expression of the finite element $i \in \mathcal{E}$. In addition, since $u_h(\bar{\boldsymbol{u}}_i)$ and $v_h(\bar{\boldsymbol{v}}_i)$ of Eq. (6.2.5) and Eq. (6.2.6) are functions $\Omega_i \to \mathbb{R}$, then note that they can be written as $u_h(\bar{\boldsymbol{u}}_i)(x)$ and $v_h(\bar{\boldsymbol{v}}_i)(x)$ with respect to $x \in \Omega_i$. However, the reason that the notations $u_h(\bar{\boldsymbol{u}}_i)$ and $v_h(\bar{\boldsymbol{v}}_i)$ are used here is because undetermined multipliers $\bar{\boldsymbol{u}}_i$ and $\bar{\boldsymbol{v}}_i$ are variables in the finite element formulation. On the other hand, the basis functions $\boldsymbol{\varphi}_i(x) = \left(\varphi_{i(1)}(x), \varphi_{i(2)}(x)\right)^\top$ in the finite element are functions of $x \in \Omega_i$ and are constructed using

$$\bar{\boldsymbol{x}}_i = \begin{pmatrix} x_{i-1} \\ x_i \end{pmatrix} = \begin{pmatrix} x_{i(1)} \\ x_{i(2)} \end{pmatrix}.$$

$\bar{\boldsymbol{x}}_i$ in this case is called the element node vector with respect to the finite element $i \in \mathcal{E}$.

The basis functions on the finite element defined in this way satisfy

$$\varphi_{i(\alpha)}\left(x_{i(\beta)}\right) = \delta_{\alpha\beta} \tag{6.2.7}$$

with respect to $\alpha, \beta \in \{1,2\}$. Moreover, the equation

$$\sum_{\alpha \in \{1,2\}} \varphi_{i(\alpha)}(x) = 1 \tag{6.2.8}$$

holds for all $x \in \Omega_i$ with respect to all $i \in \mathcal{E}$. Equation (6.2.7) is a condition for the undetermined multipliers $\bar{\boldsymbol{u}}_i$ and $\bar{\boldsymbol{v}}_i$ to represent the node values of the approximate functions. Moreover, Eq. (6.2.8) is a condition for expressing $u = 1$ exactly over Ω_i using $\bar{\boldsymbol{u}}_i = (1,1)^\top$.

In order to relate functions $u_h(\bar{\boldsymbol{u}}_i)$ and $v_h(\bar{\boldsymbol{v}}_i)$ on $\Omega_i = (x_{i-1}, x_i)$ defined on Eq. (6.2.5) and Eq. (6.2.6) with the total node value vectors $\bar{\boldsymbol{u}} = (u_0, \ldots, u_m)^\top$ and $\bar{\boldsymbol{v}} = (v_0, \ldots, v_m)^\top$, a matrix $\boldsymbol{Z}_i \in \mathbb{R}^{3\times(m+1)}$ used as

$$u_h(\bar{\boldsymbol{u}}_i) = \begin{pmatrix} \varphi_{i(1)} & \varphi_{i(2)} \end{pmatrix} \begin{pmatrix} 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \end{pmatrix} \begin{pmatrix} u_0 \\ \vdots \\ u_{i-1} \\ u_i \\ \vdots \\ u_m \end{pmatrix}$$

$$\begin{align}
&= \boldsymbol{\varphi}_i \cdot \left(\boldsymbol{Z}_i \bar{\boldsymbol{u}}\right), \tag{6.2.9}\\
v_h\left(\bar{\boldsymbol{v}}_i\right) &= \boldsymbol{\varphi}_i \cdot \left(\boldsymbol{Z}_i \bar{\boldsymbol{v}}\right) \tag{6.2.10}
\end{align}$$

is introduced. Such a $\boldsymbol{Z}_i$ is called Boolean matrix.

6.2.3 Discretized Equations

The approximate functions for each finite element were constructed as Eq. (6.2.9) and Eq. (6.2.10). Therefore, it is possible to substitute these into the weak form (Problem 6.1.2) of the one-dimensional Poisson problem and to obtain the discretized equation with $\bar{\boldsymbol{u}}_{\mathrm{N}}$ as an unknown.

Substituting $u_h\left(\bar{\boldsymbol{u}}\right)$ and $v_h\left(\bar{\boldsymbol{v}}\right)$ of Eq. (6.2.1) and Eq. (6.2.2) into the weak form, we get

$$a\left(u_h\left(\bar{\boldsymbol{u}}\right), v_h\left(\bar{\boldsymbol{v}}\right)\right) = l\left(v_h\left(\bar{\boldsymbol{v}}\right)\right). \tag{6.2.11}$$

Here, the left-hand side of Eq. (6.2.11) can split the domain of integration for each element and can be written as

$$\begin{align}
a\left(u_h\left(\bar{\boldsymbol{u}}\right), v_h\left(\bar{\boldsymbol{v}}\right)\right) &= \sum_{i \in \{1,\ldots,m\}} \int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}u_h}{\mathrm{d}x}\left(\bar{\boldsymbol{u}}_i\right) \frac{\mathrm{d}v_h}{\mathrm{d}x}\left(\bar{\boldsymbol{v}}_i\right)\ \mathrm{d}x \nonumber\\
&= \sum_{i \in \{1,\ldots,m\}} a_i\left(u_h\left(\bar{\boldsymbol{u}}_i\right), v_h\left(\bar{\boldsymbol{v}}_i\right)\right). \tag{6.2.12}
\end{align}$$

Each term on the right-hand side of Eq. (6.2.12) can be summarized using $u_h\left(\bar{\boldsymbol{u}}_i\right)$ and $v_h\left(\bar{\boldsymbol{v}}_i\right)$ of Eq. (6.2.9) and Eq. (6.2.10) as

$$\begin{align}
&a_i\left(u_h\left(\bar{\boldsymbol{u}}_i\right), v_h\left(\bar{\boldsymbol{v}}_i\right)\right) \nonumber\\
&\quad= \begin{pmatrix} v_{i(1)} & v_{i(2)} \end{pmatrix} \nonumber\\
&\qquad \times \begin{pmatrix} \displaystyle\int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}\varphi_{i(1)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(1)}}{\mathrm{d}x}\ \mathrm{d}x & \displaystyle\int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}\varphi_{i(1)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(2)}}{\mathrm{d}x}\ \mathrm{d}x \\ \displaystyle\int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}\varphi_{i(2)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(1)}}{\mathrm{d}x}\ \mathrm{d}x & \displaystyle\int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}\varphi_{i(2)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(2)}}{\mathrm{d}x}\ \mathrm{d}x \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix} \nonumber\\
&\quad= \begin{pmatrix} v_{i(1)} & v_{i(2)} \end{pmatrix} \begin{pmatrix} a_i\left(\varphi_{i(1)}, \varphi_{i(1)}\right) & a_i\left(\varphi_{i(1)}, \varphi_{i(2)}\right) \\ a_i\left(\varphi_{i(2)}, \varphi_{i(1)}\right) & a_i\left(\varphi_{i(2)}, \varphi_{i(2)}\right) \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \end{pmatrix} \nonumber\\
&\quad= \bar{\boldsymbol{v}}_i \cdot \left(\bar{\boldsymbol{A}}_i \bar{\boldsymbol{u}}_i\right) = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^{\top} \bar{\boldsymbol{A}}_i \boldsymbol{Z}_i \bar{\boldsymbol{u}}\right) = \bar{\boldsymbol{v}} \cdot \left(\tilde{\boldsymbol{A}}_i \bar{\boldsymbol{u}}\right). \tag{6.2.13}
\end{align}$$

Here, $\bar{\boldsymbol{A}}_i = \left(\bar{a}_{i(\alpha\beta)}\right)_{\alpha\beta} \in \mathbb{R}^{2\times 2}$ is called the coefficient matrix of the finite element $i \in \mathcal{E}$. Let $\tilde{\boldsymbol{A}}_i \in \mathbb{R}^{(m+1)\times(m+1)}$ be a matrix which has been expanded with zero added to go with the total nodal value vectors. It should be noted that in contrast to the element nodal value vectors $\bar{\boldsymbol{u}}_i$ and $\bar{\boldsymbol{v}}_i$ of the finite element $i \in \mathcal{E}$ being elements of $\mathbb{R}^2$, the total nodal value vectors $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ are elements of $\mathbb{R}^{m+1}$.

If Eq. (6.2.3) and Eq. (6.2.4) are used to calculate $\bar{\boldsymbol{A}}_i$, one obtains

$$\begin{aligned}
\bar{a}_{i(11)} &= \int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}\varphi_{i(1)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(1)}}{\mathrm{d}x} \ \mathrm{d}x = \frac{1}{\left(x_{i(2)} - x_{i(1)}\right)^2} \int_{x_{i(1)}}^{x_{i(2)}} (-1)^2 \ \mathrm{d}x \\
&= \frac{1}{x_{i(2)} - x_{i(1)}}, \\
\bar{a}_{i(12)} &= \int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}\varphi_{i(1)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(2)}}{\mathrm{d}x} \ \mathrm{d}x = \frac{1}{\left(x_{i(2)} - x_{i(1)}\right)^2} \int_{x_{i(1)}}^{x_{i(2)}} 1 \cdot (-1) \ \mathrm{d}x \\
&= \frac{-1}{x_{i(2)} - x_{i(1)}}, \\
\bar{a}_{i(21)} &= \bar{a}_{i(12)}, \\
\bar{a}_{i(22)} &= \int_{x_{i(1)}}^{x_{i(2)}} \frac{\mathrm{d}\varphi_{i(2)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(2)}}{\mathrm{d}x} \ \mathrm{d}x = \frac{1}{\left(x_{i(2)} - x_{i(1)}\right)^2} \int_{x_{i(1)}}^{x_{i(2)}} 1^2 \ \mathrm{d}x \\
&= \frac{1}{x_{i(2)} - x_{i(1)}}
\end{aligned}$$

and

$$\bar{\boldsymbol{A}}_i = \begin{pmatrix} \bar{a}_{i(11)} & \bar{a}_{i(12)} \\ \bar{a}_{i(21)} & \bar{a}_{i(22)} \end{pmatrix} = \frac{1}{x_{i(2)} - x_{i(1)}} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{6.2.14}$$

On the other hand, the right-hand side of Eq. (6.2.11) can also be split into each element as

$$\begin{aligned}
l\left(v_h\left(\bar{\boldsymbol{v}}\right)\right) &= \sum_{i \in \{1,\ldots,m\}} \int_{x_{i(1)}}^{x_{i(2)}} b v_h\left(\bar{\boldsymbol{v}}_i\right) \ \mathrm{d}x + p_{\mathrm{N}} v_h\left(\bar{\boldsymbol{v}}_m\right) \\
&= \sum_{i \in \{1,\ldots,m\}} l_i\left(v_h\left(\bar{\boldsymbol{v}}_i\right)\right).
\end{aligned} \tag{6.2.15}$$

Here, for the finite element $i \in \{1, \ldots, m-1\}$ and m, let

$$\begin{aligned}
l_i\left(v_h\left(\bar{\boldsymbol{v}}_i\right)\right) &= \begin{pmatrix} v_{i(1)} & v_{i(2)} \end{pmatrix} \begin{pmatrix} \int_{x_{i(1)}}^{x_{i(2)}} b\varphi_{i(1)} \ \mathrm{d}x \\ \int_{x_{i(1)}}^{x_{i(2)}} b\varphi_{i(2)} \ \mathrm{d}x \end{pmatrix} = \begin{pmatrix} v_{i(1)} & v_{i(2)} \end{pmatrix} \begin{pmatrix} \bar{b}_{i(1)} \\ \bar{b}_{i(2)} \end{pmatrix} \\
&= \bar{\boldsymbol{v}}_i \cdot \bar{\boldsymbol{b}}_i = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^{\top} \bar{\boldsymbol{b}}_i\right) = \bar{\boldsymbol{v}} \cdot \tilde{\boldsymbol{b}}_i \\
&= \bar{\boldsymbol{v}}_i \cdot \bar{\boldsymbol{l}}_i = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^{\top} \bar{\boldsymbol{l}}_i\right) = \bar{\boldsymbol{v}} \cdot \tilde{\boldsymbol{l}}_i, \qquad (6.2.16) \\
l_m\left(v_h\left(\bar{\boldsymbol{v}}_m\right)\right) &= \begin{pmatrix} v_{m(1)} & v_{m(2)} \end{pmatrix} \left(\begin{pmatrix} \int_{x_{m(1)}}^{x_{m(2)}} b\varphi_{i(1)} \ \mathrm{d}x \\ \int_{x_{m(1)}}^{x_{m(2)}} b\varphi_{i(2)} \ \mathrm{d}x \end{pmatrix} + \begin{pmatrix} 0 \\ p_{\mathrm{N}} \end{pmatrix} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \begin{pmatrix} v_{m(1)} & v_{m(2)} \end{pmatrix} \left(\begin{pmatrix} \bar{b}_{m(1)} \\ \bar{b}_{m(2)} \end{pmatrix} + \begin{pmatrix} \bar{p}_{m(1)} \\ \bar{p}_{m(2)} \end{pmatrix} \right) \\
&= \bar{\boldsymbol{v}}_m \cdot \left(\bar{\boldsymbol{b}}_m + \bar{\boldsymbol{p}}_m \right) = \bar{\boldsymbol{v}} \cdot \left\{ \boldsymbol{Z}_m^\top \left(\bar{\boldsymbol{b}}_m + \bar{\boldsymbol{p}}_m \right) \right\} = \bar{\boldsymbol{v}} \cdot \left(\tilde{\boldsymbol{b}}_m + \tilde{\boldsymbol{p}}_m \right) \\
&= \bar{\boldsymbol{v}}_m \cdot \bar{\boldsymbol{l}}_m = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_m^\top \bar{\boldsymbol{l}}_m \right) = \bar{\boldsymbol{v}} \cdot \tilde{\boldsymbol{l}}_m, \qquad (6.2.17)
\end{aligned}
$$

respectively. Here, with respect to $i \in \mathcal{E} = \{1, \ldots, m\}$, $\bar{\boldsymbol{l}}_i$ is called the known term vector of the finite element i. $\bar{\boldsymbol{b}}_i$ and $\bar{\boldsymbol{p}}_i$ represent the components of the known term vector constructed from b and p_{N} respectively. $\tilde{\boldsymbol{l}}_i$, $\tilde{\boldsymbol{b}}_i$ and $\tilde{\boldsymbol{p}}_i$ are $\bar{\boldsymbol{l}}_i$, $\bar{\boldsymbol{b}}_i$ and $\bar{\boldsymbol{p}}_i$ respectively expanded by adding in 0 to match the total nodal value vectors.

When b is a constant function, $\bar{\boldsymbol{b}}_i = \left(\bar{b}_{i(1)}, \bar{b}_{i(2)} \right)^\top$ is calculated as

$$
\begin{aligned}
\bar{b}_{i(1)} &= b \int_{x_{i(1)}}^{x_{i(2)}} \varphi_{i(1)} \, \mathrm{d}x = b \int_{x_{i(1)}}^{x_{i(2)}} \frac{x_{i(2)} - x}{x_{i(2)} - x_{i(1)}} \, \mathrm{d}x = b \frac{x_{i(2)} - x_{i(1)}}{2}, \\
\bar{b}_{i(2)} &= b \int_{x_{i(1)}}^{x_{i(2)}} \varphi_{i(2)} \, \mathrm{d}x = b \int_{x_{i(1)}}^{x_{i(2)}} \frac{x - x_{i(1)}}{x_{i(2)} - x_{i(1)}} \, \mathrm{d}x = b \frac{x_{i(2)} - x_{i(1)}}{2}
\end{aligned}
$$

and is obtained as

$$
\bar{\boldsymbol{b}}_i = b \frac{x_{i(2)} - x_{i(1)}}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \qquad (6.2.18)
$$

Here, if Eq. (6.2.12) with Eq. (6.2.13) substituted in, and Eq. (6.2.15) with Eq. (6.2.16) and Eq. (6.2.17) substituted in are substituted into the weak form (Eq. (6.2.11)), we get

$$
\bar{\boldsymbol{v}} \cdot \sum_{i \in \{1, \ldots, m\}} \left(\tilde{\boldsymbol{A}}_i \bar{\boldsymbol{u}} \right) = \bar{\boldsymbol{v}} \cdot \sum_{i \in \{1, \ldots, m\}} \tilde{\boldsymbol{l}}_i,
$$

which can be rewritten as

$$
\bar{\boldsymbol{v}} \cdot \left(\bar{\boldsymbol{A}} \bar{\boldsymbol{u}} \right) = \bar{\boldsymbol{v}} \cdot \bar{\boldsymbol{l}}. \qquad (6.2.19)
$$

In other words we set

$$
\begin{aligned}
\bar{\boldsymbol{A}} &= \sum_{i \in \{1, \ldots, m\}} \tilde{\boldsymbol{A}}_i \in \mathbb{R}^{(m+1) \times (m+1)}, \\
\bar{\boldsymbol{l}} &= \sum_{i \in \{1, \ldots, m\}} \tilde{\boldsymbol{l}}_i = \sum_{i \in \{1, \ldots, m\}} \tilde{\boldsymbol{b}}_i + \tilde{\boldsymbol{p}}_m = \bar{\boldsymbol{b}} + \bar{\boldsymbol{p}} \in \mathbb{R}^{m+1}.
\end{aligned}
$$

$\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{l}}$ are called the total coefficient matrix and total known term vector, respectively. Moreover, $\bar{\boldsymbol{b}}$ and $\bar{\boldsymbol{p}}$ are called total nodal value vectors of b and p_{N}, respectively.

Equation (6.2.19) gives the weak form but there were no fundamental boundary conditions assumed with respect to u_h and v_h. Hence, substitute

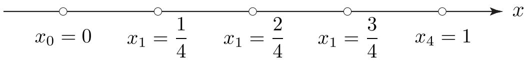


Fig. 6.8: Finite element mesh when $m = 4$.

in the fundamental boundary conditions into Eq. (6.2.19). $u_0 = \bar{u}_{\mathrm{D}} = u_{\mathrm{D}}$ ($\bar{u}_{\mathrm{D}}$ is the node value of the fundamental boundary condition, u_{D} is the known value of boundary value problem) and $v_0 = \bar{v}_{\mathrm{D}} = 0$. Substituting these into Eq. (6.2.19) gives

$$
\begin{aligned}
&\left(0 \,\middle|\, v_1 \quad \cdots \quad v_m\right) \left(\left(\begin{array}{c|ccc} \bar{a}_{00} & \bar{a}_{01} & \cdots & \bar{a}_{0m} \\ \hline \bar{a}_{10} & \bar{a}_{11} & \cdots & \bar{a}_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{m0} & \bar{a}_{m1} & \cdots & \bar{a}_{mm} \end{array} \right) \left(\begin{array}{c} u_{\mathrm{D}} \\ \hline u_1 \\ \vdots \\ u_m \end{array} \right) - \left(\begin{array}{c} l_0 \\ \hline l_1 \\ \vdots \\ l_m \end{array} \right) \right) \\
&= \begin{pmatrix} 0 & \bar{\boldsymbol{v}}_{\mathrm{N}}^{\top} \end{pmatrix} \left(\begin{pmatrix} \bar{A}_{\mathrm{DD}} & \bar{\boldsymbol{A}}_{\mathrm{DN}} \\ \bar{\boldsymbol{A}}_{\mathrm{ND}} & \bar{\boldsymbol{A}}_{\mathrm{NN}} \end{pmatrix} \begin{pmatrix} \bar{u}_{\mathrm{D}} \\ \bar{\boldsymbol{u}}_{\mathrm{N}} \end{pmatrix} - \begin{pmatrix} \bar{l}_{\mathrm{D}} \\ \bar{\boldsymbol{l}}_{\mathrm{N}} \end{pmatrix} \right) = 0. \qquad (6.2.20)
\end{aligned}
$$

Rearranging Eq. (6.2.20), we get

$$
\begin{aligned}
&\begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \cdot \left(\begin{pmatrix} \bar{a}_{11} & \cdots & \bar{a}_{1m} \\ \vdots & \ddots & \vdots \\ \bar{a}_{m1} & \cdots & \bar{a}_{mm} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix} - \begin{pmatrix} l_1 \\ \vdots \\ l_m \end{pmatrix} + \begin{pmatrix} u_{\mathrm{D}}\bar{a}_{10} \\ \vdots \\ u_{\mathrm{D}}\bar{a}_{m0} \end{pmatrix} \right) \\
&= \bar{\boldsymbol{v}}_{\mathrm{N}}^{\top} \left(\bar{\boldsymbol{A}}_{\mathrm{NN}} \bar{\boldsymbol{u}}_{\mathrm{N}} - \bar{\boldsymbol{l}}_{\mathrm{N}} + \bar{\boldsymbol{u}}_{\mathrm{D}} \bar{\boldsymbol{A}}_{\mathrm{ND}} \right) = 0.
\end{aligned}
$$

$\bar{\boldsymbol{v}}_{\mathrm{N}}$ is arbitrary, so it can be written as

$$
\bar{\boldsymbol{A}}_{\mathrm{NN}} \bar{\boldsymbol{u}}_{\mathrm{N}} = \bar{\boldsymbol{l}}_{\mathrm{N}} - \bar{\boldsymbol{u}}_{\mathrm{D}} \bar{\boldsymbol{A}}_{\mathrm{ND}} = \hat{\boldsymbol{l}}. \qquad (6.2.21)
$$

Equation (6.2.21) is a simultaneous linear equation with respect to unknown vector $\bar{\boldsymbol{u}}_{\mathrm{N}}$ and is called the discretized equation of the finite element method. Solving Eq. (6.2.21) for $\bar{\boldsymbol{u}}_{\mathrm{N}}$ gives

$$
\bar{\boldsymbol{u}}_{\mathrm{N}} = \bar{\boldsymbol{A}}_{\mathrm{NN}}^{-1} \hat{\boldsymbol{l}}. \qquad (6.2.22)
$$

From this, the finite element solution $u_h(\bar{\boldsymbol{u}})$ can be obtained by substituting $\bar{\boldsymbol{u}} = \left(\bar{u}_{\mathrm{D}}, \bar{\boldsymbol{u}}_{\mathrm{N}}^{\top}\right)^{\top}$ into Eq. (6.2.1). Moreover, the finite element solution $u_h(\bar{\boldsymbol{u}}_i)$ on Ω_i on the finite element $i \in \mathcal{E}$ domain can be found by Eq. (6.2.9).

6.2.4 Exercise Problem

We solve a one-dimensional Poisson problem using the finite element method in the following exercise.

Exercise 6.2.1 (Finite element method for 1D Poisson problem) Let b be a constant function as in Problem 6.1.1. Here, show the system of linear equations for seeking the approximate solution in the aforementioned one-dimensional finite element method using the finite element mesh shown in Fig. 6.8. Moreover, obtain the approximate solution when we set $b = 1$, $u_{\mathrm{D}} = 0$ and $p_{\mathrm{N}} = 0$. □

Answer Let the size of the finite element be $h = 1/4$. From Eq. (6.2.14), Eq. (6.2.18) and Eq. (6.2.17), we get

$$\bar{\boldsymbol{A}}_i = \frac{1}{h}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad \bar{\boldsymbol{b}}_i = \frac{hb}{2}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \bar{\boldsymbol{p}}_4 = \begin{pmatrix} 0 \\ p_{\mathrm{N}} \end{pmatrix}.$$

Expanding $\bar{\boldsymbol{A}}_1$ and $\bar{\boldsymbol{b}}_1$ to match the node value vectors gives

$$\tilde{\boldsymbol{A}}_1 = \frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{\boldsymbol{b}}_1 = \frac{hb}{2}\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

If $\tilde{\boldsymbol{A}}_2$ and $\tilde{\boldsymbol{b}}_2$ are overlapped with $\tilde{\boldsymbol{A}}_1$ and $\tilde{\boldsymbol{b}}_1$ respectively, we obtain

$$\tilde{\boldsymbol{A}}_1 + \tilde{\boldsymbol{A}}_2 = \frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1+1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{\boldsymbol{b}}_1 + \tilde{\boldsymbol{b}}_2 = \frac{hb}{2}\begin{pmatrix} 1 \\ 1+1 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Similarly, overlapping $\tilde{\boldsymbol{A}}_3$ and $\tilde{\boldsymbol{b}}_3$, $\tilde{\boldsymbol{A}}_4$ and $\tilde{\boldsymbol{b}}_4$ as well as $\tilde{\boldsymbol{p}}_4$ gives

$$\bar{\boldsymbol{A}} = \sum_{i \in \{1,\ldots,4\}} \tilde{\boldsymbol{A}}_i = \frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix},$$

$$\bar{\boldsymbol{l}} = \sum_{i \in \{1,\ldots,4\}} \tilde{\boldsymbol{b}}_i + \tilde{\boldsymbol{p}}_4 = \frac{hb}{2}\begin{pmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ p_{\mathrm{N}} \end{pmatrix}.$$

Substituting fundamental boundary conditions $u_0 = \bar{u}_{\mathrm{D}} = u_{\mathrm{D}}$ and $v_0 = \bar{v}_{\mathrm{D}} = 0$ into Eq. (6.2.19) gives

$$\begin{pmatrix} 0 \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} \cdot \left(\frac{1}{h}\begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} u_{\mathrm{D}} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} - \frac{hb}{2}\begin{pmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ p_{\mathrm{N}} \end{pmatrix} \right) = 0.$$

Rearranging this equation gives

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} \cdot \left(\frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} - \frac{hb}{2}\begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ p_{\mathrm{N}} \end{pmatrix} - \frac{1}{h}\begin{pmatrix} u_{\mathrm{D}} \\ 0 \\ 0 \\ 0 \end{pmatrix} \right) = 0.$$

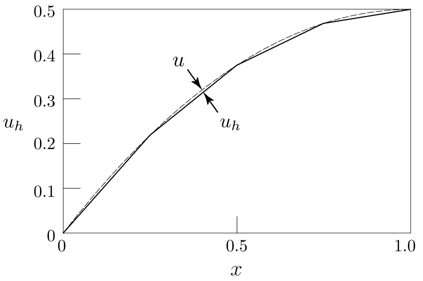


Fig. 6.9: Exact solution u and approximate solution u_h of Exercise 6.2.1.

(v_1, v_2, v_3, v_4) is arbitrary, so

$$\frac{1}{h}\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \frac{hb}{2}\begin{pmatrix} 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ p_{\mathrm{N}} \end{pmatrix} + \frac{1}{h}\begin{pmatrix} u_{\mathrm{D}} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

can be obtained. This equation is a simultaneous linear equation such as

$$\bar{\boldsymbol{A}}_{\mathrm{NN}}\bar{\boldsymbol{u}}_{\mathrm{N}} = \hat{\boldsymbol{l}}$$

with respect to $\bar{\boldsymbol{u}}_{\mathrm{N}}$. Here, when $b = 1$, $u_{\mathrm{D}} = 0$, $p_{\mathrm{N}} = 0$, we get

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} = \frac{1}{4^2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 2 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1/2 \end{pmatrix} = \begin{pmatrix} 7/32 \\ 3/8 \\ 15/32 \\ 1/2 \end{pmatrix}.$$

On the other hand, the exact solution is

$$u = -\frac{1}{2}x^2 + x.$$

Figure 6.9 shows a comparison between the numerical solution u_h and the exact solution u. □

Change the boundary conditions of Exercise 6.2.1 and consider the following problem.

Exercise 6.2.2 (Dirichlet problem of 1D Poisson problem) Consider a problem seeking $u : (0, 1) \to \mathbb{R}$ satisfying

$$-\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = b \quad \text{in } (0,1), \quad u(0) = u_{\mathrm{D}0}, \quad u(1) = u_{\mathrm{D}1}$$

when b, $u_{\mathrm{D}0}$, $u_{\mathrm{D}1}$, $p_{\mathrm{N}} \in \mathbb{R}$ are given. Use the mesh for the finite elements of Fig. 6.8 in order to show the simultaneous linear equations when seeking the approximate solution using the finite element method. Moreover, obtain the numerical solution when $b = 1$ and $u_{\mathrm{D}0} = u_{\mathrm{D}1} = 0$. □

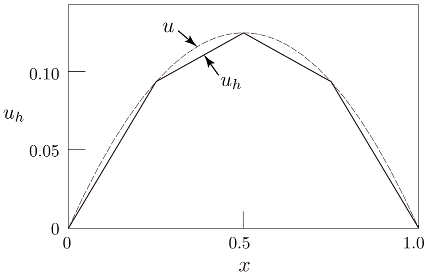


Fig. 6.10: Exact solution u and approximate solution u_h of Exercise 6.2.2.

Answer The calculation of $\bar{\boldsymbol{A}}$ is similar to Exercise 6.2.1. Moreover $\bar{\boldsymbol{l}} = \sum_{i\in\{1,\dots,4\}} \tilde{\boldsymbol{b}}_i$. If the fundamental boundary conditions $u_0 = \bar{u}_{\mathrm{D}0} = u_{\mathrm{D}0}$, $u_4 = \bar{u}_{\mathrm{D}4} = u_{\mathrm{D}1}$, $v_0 = \bar{v}_{\mathrm{D}0} = 0$ and $v_4 = \bar{v}_{\mathrm{D}4} = 0$ are substituted into Eq. (6.2.19), we get

$$\begin{pmatrix} 0 \\ v_1 \\ v_2 \\ v_3 \\ 0 \end{pmatrix} \cdot \left(\frac{1}{h} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} u_{\mathrm{D}0} \\ u_1 \\ u_2 \\ u_3 \\ u_{\mathrm{D}1} \end{pmatrix} - \frac{hb}{2} \begin{pmatrix} 1 \\ 2 \\ 2 \\ 2 \\ 1 \end{pmatrix} \right) = 0.$$

Rearranging this equation gives

$$\begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} \left(\frac{1}{h} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} - \frac{h}{2} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} - \frac{1}{h} \begin{pmatrix} u_{\mathrm{D}0} \\ 0 \\ u_{\mathrm{D}1} \end{pmatrix} \right) = 0.$$

Since (v_1, v_2, v_3) is arbitrary,

$$\frac{1}{h} \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \frac{h}{2} \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} + \frac{1}{h} \begin{pmatrix} u_{\mathrm{D}0} \\ 0 \\ u_{\mathrm{D}1} \end{pmatrix}$$

is obtained. These equations are simultaneous linear equations like

$$\bar{\boldsymbol{A}}_{\mathrm{NN}} \bar{\boldsymbol{u}}_{\mathrm{N}} = \hat{\boldsymbol{l}}.$$

When $b = 1$ and $u_{\mathrm{D}0} = u_{\mathrm{D}1} = 0$, we get

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \frac{1}{4^3} \begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3/32 \\ 1/8 \\ 3/32 \end{pmatrix}.$$

Figure 6.10 shows the comparison between the numerical solution u_h and the exact solution $u = \dfrac{1}{2} x\,(x-1)$. □

The following can be said to be one of the characteristics of the finite element method from Exercise 6.2.1 and Exercise 6.2.2. The Galerkin method seen in Sect. 6.1 required a change of basis functions for a change in fundamental boundary conditions. However, in the finite element method, fundamental boundary conditions can be changed easily as boundary conditions are substituted in after seeking the coefficient matrix and known term vector.

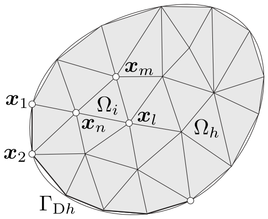


Fig. 6.11: Triangular finite elements and nodes in a 2D domain Ω.

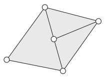


Fig. 6.12: Counterexample of triangular finite element and nodes.

6.3 Two-Dimensional Finite Element Method

Next, we consider a finite element method with respect to a two-dimensional Poisson problem (Problem 6.1.6 with $d = 2$). Here, approximate functions used in the finite element method within the framework of the Galerkin method are also defined. After that, their approximate functions will be viewed as approximate functions defined for each split finite element domain.

6.3.1 Approximate Functions in Galerkin Method

With reference to Fig. 6.11, a two-dimensional domain Ω and Dirichlet boundary Γ_{D} are assumed to be approximated by a polygonal domain Ω_h and line graph $\Gamma_{\mathrm{D}h}$, respectively. Furthermore, Ω_h is split into a set of triangle domains $\{\Omega_i\}_i$. In this case, Ω_i is called the domain of triangular finite elements and the set of finite element numbers i is denoted by $\mathcal{E}$. Here, we assume that there are no overlapping triangular domains Ω_i for all $i \in \mathcal{E}$ and there are no vertices on the boundary of triangular domains other than the vertices of Ω_i as in Fig. 6.12.

Moreover, the vertices $\boldsymbol{x}_j = (x_{j1}, x_{j2})^\top$ of triangles are called nodes and the set of node numbers j is denoted as $\mathcal{N}$. Furthermore, $\mathcal{N}$ is split into two sets that are the set $\mathcal{N}_{\mathrm{D}}$ of node numbers on $\Gamma_{\mathrm{D}h}$ and the set $\mathcal{N}_{\mathrm{N}} = \mathcal{N} \setminus \mathcal{N}_{\mathrm{D}}$ of the other node numbers. $\mathcal{N}$ is reordered so that $\mathcal{N}_{\mathrm{D}}$ comes first.

With respect to a triangular finite element mesh such as this, basis functions in the two-dimensional finite element method with respect to Problem 6.1.6 are defined by a pyramidal function with a unit height such as shown in Fig. 6.13 with respect to a node $j \in \mathcal{N}$. In other words, it is assumed that ϕ_j is a first-order polynomial with support on the finite elements having a node j on the vertex and is a continuous function taking the value 1 at node j and the value 0 at other nodes. These characteristics, as seen in the one-dimensional finite element method, lead to the conclusion that integration after substitution into the weak form is convenient and the unknown multipliers in the approximate

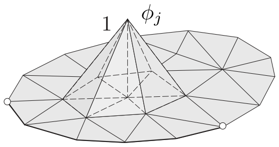


Fig. 6.13: Basis function ϕ_j.

functions become node values as shown next.

Using these basis functions, the finite element method sets the approximate functions to be

$$u_h\left(\bar{\boldsymbol{u}}\right) = \sum_{j\in\mathcal{N}_{\mathrm{D}}} u_j\phi_j + \sum_{j\in\mathcal{N}_{\mathrm{N}}} u_j\phi_j = \begin{pmatrix}\bar{\boldsymbol{u}}_{\mathrm{D}}\\ \bar{\boldsymbol{u}}_{\mathrm{N}}\end{pmatrix}\cdot\begin{pmatrix}\boldsymbol{\phi}_{\mathrm{D}}\\ \boldsymbol{\phi}_{\mathrm{N}}\end{pmatrix} = \bar{\boldsymbol{u}}\cdot\boldsymbol{\phi}, \tag{6.3.1}$$

$$v_h\left(\bar{\boldsymbol{v}}\right) = \sum_{j\in\mathcal{N}_{\mathrm{D}}} v_j\phi_j + \sum_{j\in\mathcal{N}_{\mathrm{N}}} v_j\phi_j = \begin{pmatrix}\bar{\boldsymbol{v}}_{\mathrm{D}}\\ \bar{\boldsymbol{v}}_{\mathrm{N}}\end{pmatrix}\cdot\begin{pmatrix}\boldsymbol{\phi}_{\mathrm{D}}\\ \boldsymbol{\phi}_{\mathrm{N}}\end{pmatrix} = \bar{\boldsymbol{v}}\cdot\boldsymbol{\phi}. \tag{6.3.2}$$

Here, from the fact that $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ represent the node values of the approximate functions u_h and v_h respectively, $\bar{\boldsymbol{u}}$ and $\bar{\boldsymbol{v}}$ are called nodal value vectors. Moreover, $\bar{\boldsymbol{u}}_{\mathrm{D}} = \left(u_{\mathrm{D}}\left(\boldsymbol{x}_j\right)\right)_{j\in\mathcal{N}_{\mathrm{D}}}$ and $\bar{\boldsymbol{v}}_{\mathrm{D}} = \boldsymbol{0}_{\mathbb{R}^{|\mathcal{N}_{\mathrm{D}}|}}$ are called Dirichlet-type nodal value vectors and $\bar{\boldsymbol{u}}_{\mathrm{N}} = \left(u_j\right)_{j\in\mathcal{N}_{\mathrm{N}}}$ and $\bar{\boldsymbol{v}}_{\mathrm{N}} = \left(v_j\right)_{j\in\mathcal{N}_{\mathrm{N}}}$ are called Neumann-type nodal value vectors.

6.3.2 Approximate Functions in Finite Element Method

In the finite element method, the basis functions $\boldsymbol{\phi}$ defined on Ω_h can be rewritten as basis functions defined on Ω_i with respect to all $i \in \mathcal{E}$. Based on that, let the approximate function on Ω_i be

$$u_h\left(\bar{\boldsymbol{u}}_i\right) = \begin{pmatrix}\varphi_{i(1)} & \varphi_{i(2)} & \varphi_{i(3)}\end{pmatrix}\begin{pmatrix}u_{i(1)}\\ u_{i(2)}\\ u_{i(3)}\end{pmatrix} = \boldsymbol{\varphi}_i\cdot\bar{\boldsymbol{u}}_i, \tag{6.3.3}$$

$$v_h\left(\bar{\boldsymbol{v}}_i\right) = \begin{pmatrix}\varphi_{i(1)} & \varphi_{i(2)} & \varphi_{i(3)}\end{pmatrix}\begin{pmatrix}v_{i(1)}\\ v_{i(2)}\\ v_{i(3)}\end{pmatrix} = \boldsymbol{\varphi}_i\cdot\bar{\boldsymbol{v}}_i. \tag{6.3.4}$$

Here, when the three node numbers of the finite element $i \in \mathcal{E}$ are l, m and $n \in \mathcal{N}$, $\boldsymbol{\varphi}_i = \left(\varphi_l, \varphi_m, \varphi_n\right)^{\top} = \left(\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)}\right)^{\top} : \Omega_i \to \mathbb{R}^3$ are referred to as basis functions in the finite element. Moreover,

$$\bar{\boldsymbol{x}}_i = \begin{pmatrix}\boldsymbol{x}_l\\ \boldsymbol{x}_m\\ \boldsymbol{x}_n\end{pmatrix} = \begin{pmatrix}\boldsymbol{x}_{i(1)}\\ \boldsymbol{x}_{i(2)}\\ \boldsymbol{x}_{i(3)}\end{pmatrix},$$

$$\bar{\boldsymbol{u}}_i = \begin{pmatrix}u_l\\ u_m\\ u_n\end{pmatrix} = \begin{pmatrix}u_{i(1)}\\ u_{i(2)}\\ u_{i(3)}\end{pmatrix}, \quad \bar{\boldsymbol{v}}_i = \begin{pmatrix}v_l\\ v_m\\ v_n\end{pmatrix} = \begin{pmatrix}v_{i(1)}\\ v_{i(2)}\\ v_{i(3)}\end{pmatrix}$$

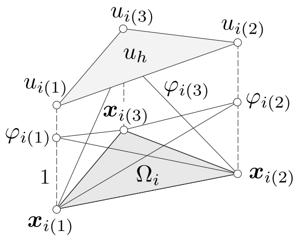


Fig. 6.14: Basis functions $\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)}$ in triangular finite element $i \in \mathcal{E}$.

are called the element node vector and element nodal value vectors with respect to u and v for the finite element $i \in \mathcal{E}$. The suffix $\alpha \in \{1,2,3\}$ used in $\boldsymbol{x}_{i(\alpha)}$, $u_{i(\alpha)}$ and $v_{i(\alpha)}$ is called the local node number of the finite element $i \in \mathcal{E}$. Figure 6.14 shows how the approximate function u_h in the finite element $i \in \mathcal{E}$ is constructed from $\boldsymbol{\varphi}_i$ and $\bar{\boldsymbol{u}}_i$.

Basis functions $\boldsymbol{\varphi}_i$ on $i \in \mathcal{E}$ constructed in this way satisfy

$$\varphi_{i(\alpha)}\left(\boldsymbol{x}_{i(\beta)}\right) = \delta_{\alpha\beta} \tag{6.3.5}$$

with respect to $\alpha, \beta \in \{1,2,3\}$. Moreover,

$$\sum_{\alpha \in \{1,2,3\}} \varphi_{i(\alpha)}(\boldsymbol{x}) = 1$$

holds on all points $\boldsymbol{x} \in \Omega_i$.

Here, obtain the equations of basis functions $\varphi_{i(1)}$, $\varphi_{i(2)}$ and $\varphi_{i(3)}$ on the finite element $i \in \mathcal{E}$. From the fact that $\varphi_{i(\alpha)}$ for $\alpha \in \{1,2,3\}$ is a linear equation with respect to $\boldsymbol{x} = (x_1, x_2)^\top \in \Omega_i$, it is a complete first-order polynomial constructed of three unknown multipliers. Let this be

$$\varphi_{i(\alpha)} = \zeta_\alpha + \eta_\alpha x_1 + \theta_\alpha x_2. \tag{6.3.6}$$

The undetermined multipliers ζ_α, η_α and θ_α can be determined by giving the values of $\varphi_{i(\alpha)}$ at three nodes. Their values can be given by Eq. (6.3.5). In other words, with respect to $\beta \in \{1,2,3\}$, it can be determined by

$$\varphi_{i(\alpha)}\left(\boldsymbol{x}_{i(\beta)}\right) = \zeta_\alpha + \eta_\alpha x_{i(\beta)1} + \theta_\alpha x_{i(\beta)2} = \delta_{\alpha\beta}.$$

This equation can be expanded as

$$\begin{pmatrix} 1 & x_{i(1)1} & x_{i(1)2} \\ 1 & x_{i(2)1} & x_{i(2)2} \\ 1 & x_{i(3)1} & x_{i(3)2} \end{pmatrix} \begin{pmatrix} \zeta_1 & \zeta_2 & \zeta_3 \\ \eta_1 & \eta_2 & \eta_3 \\ \theta_1 & \theta_2 & \theta_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Here, solving for the undetermined multipliers gives

$$\begin{pmatrix} \zeta_1 & \zeta_2 & \zeta_3 \\ \eta_1 & \eta_2 & \eta_3 \\ \theta_1 & \theta_2 & \theta_3 \end{pmatrix} = \frac{1}{\gamma} \left(\begin{matrix} x_{i(2)1}x_{i(3)2} - x_{i(3)1}x_{i(2)2} \\ x_{i(2)2} - x_{i(3)2} \\ x_{i(3)1} - x_{i(2)1} \end{matrix} \right.$$

$$\left.\begin{matrix} x_{i(3)1}x_{i(1)2} - x_{i(1)1}x_{i(3)2} & x_{i(1)1}x_{i(2)2} - x_{i(2)1}x_{i(1)2} \\ x_{i(3)2} - x_{i(1)2} & x_{i(1)2} - x_{i(2)2} \\ x_{i(1)1} - x_{i(3)1} & x_{i(2)1} - x_{i(1)1} \end{matrix}\right), \tag{6.3.7}$$

where

$$\begin{aligned} \gamma &= \begin{vmatrix} x_{i(1)1} & x_{i(1)2} & 1 \\ x_{i(2)1} & x_{i(2)2} & 1 \\ x_{i(3)1} & x_{i(3)2} & 1 \end{vmatrix} \\ &= x_{i(1)1}\left(x_{i(2)2} - x_{i(3)2}\right) + x_{i(2)1}\left(x_{i(3)2} - x_{i(1)2}\right) \\ &\quad + x_{i(3)1}\left(x_{i(1)2} - x_{i(2)2}\right). \end{aligned} \tag{6.3.8}$$

If the three nodes $\boldsymbol{x}_{i(1)}$, $\boldsymbol{x}_{i(2)}$ and $\boldsymbol{x}_{i(3)}$ of the triangular finite element are chosen so that they are anti-clockwise, then γ is equal to twice the area $|\Omega_i|$ of the triangle Ω_i (Practice **6.3**).

The basis functions $\varphi_{i(1)}$, $\varphi_{i(2)}$ and $\varphi_{i(3)}$ of the finite element were obtained by substituting Eq. (6.3.7) and Eq. (6.3.8) into Eq. (6.3.6). Using these, the approximate functions $u_h(\bar{\boldsymbol{u}}_i)$ and $v_h(\bar{\boldsymbol{v}}_i)$ defined by Eq. (6.3.3) and Eq. (6.3.4) can be written as

$$\begin{aligned} u_h(\bar{\boldsymbol{u}}_i) &= \begin{pmatrix} \varphi_{i(1)} & \varphi_{i(2)} & \varphi_{i(3)} \end{pmatrix} \begin{pmatrix} 0 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_l \\ u_m \\ u_n \\ \vdots \\ u_{|\mathcal{N}|} \end{pmatrix} \\ &= \boldsymbol{\varphi}_i \cdot (\boldsymbol{Z}_i \bar{\boldsymbol{u}}), \qquad (6.3.9) \\ v_h(\bar{\boldsymbol{v}}_i) &= \boldsymbol{\varphi}_i \cdot (\boldsymbol{Z}_i \bar{\boldsymbol{v}}) \qquad (6.3.10) \end{aligned}$$

Here, $\boldsymbol{Z}_i$ is a Boolean matrix which links the total nodal value vector $\bar{\boldsymbol{u}}$ and the nodal value vector $\bar{\boldsymbol{u}}_i = (u_l, u_m, u_n)^\top$ of the finite element $i \in \mathcal{E}$.

6.3.3 Discretized Equations

Approximate functions $u_h(\bar{\boldsymbol{u}}_i)$ and $v_h(\bar{\boldsymbol{v}}_i)$ on Ω_i have been defined, so we shall substitute Eq. (6.3.9) and Eq. (6.3.10) into the weak form of two-dimensional Poisson problems (Problem 6.1.7) and see how the discretized equation with $\bar{\boldsymbol{u}}_{\mathrm{N}}$ as an unknown is obtained.

If $u_h(\bar{\boldsymbol{u}})$ and $v_h(\bar{\boldsymbol{v}})$ of Eq. (6.3.1) and Eq. (6.3.2) are substituted into the weak form, we get

$$a\left(u_h(\bar{\boldsymbol{u}}), v_h(\bar{\boldsymbol{v}})\right) = l\left(v_h(\bar{\boldsymbol{v}})\right). \tag{6.3.11}$$

The left-hand side of Eq. (6.3.11) can be written as

$$a\left(u_h(\bar{\boldsymbol{u}}), v_h(\bar{\boldsymbol{v}})\right) = \sum_{i \in \mathcal{E}} \int_{\Omega_i} \boldsymbol{\nabla} u_h(\bar{\boldsymbol{u}}_i) \cdot \boldsymbol{\nabla} v_h(\bar{\boldsymbol{v}}_i) \, \mathrm{d}x$$

$$= \sum_{i \in \mathcal{E}} a_i \left(u_h \left(\bar{\boldsymbol{u}}_i \right), v_h \left(\bar{\boldsymbol{v}}_i \right) \right). \tag{6.3.12}$$

Here, each term on the right-hand side of Eq. (6.3.12) can be summarized as

$$\begin{aligned} & a_i \left(u_h \left(\bar{\boldsymbol{u}}_i \right), v_h \left(\bar{\boldsymbol{v}}_i \right) \right) \\ & \quad = \begin{pmatrix} v_{i(1)} & v_{i(2)} & v_{i(3)} \end{pmatrix} \\ & \qquad \times \begin{pmatrix} \int_{\Omega_i} \boldsymbol{\nabla} \varphi_{i(1)} \cdot \boldsymbol{\nabla} \varphi_{i(1)} \, \mathrm{d}x & \cdots & \int_{\Omega_i} \boldsymbol{\nabla} \varphi_{i(1)} \cdot \boldsymbol{\nabla} \varphi_{i(3)} \, \mathrm{d}x \\ \vdots & \ddots & \vdots \\ \int_{\Omega_i} \boldsymbol{\nabla} \varphi_{i(3)} \cdot \boldsymbol{\nabla} \varphi_{i(1)} \, \mathrm{d}x & \cdots & \int_{\Omega_i} \boldsymbol{\nabla} \varphi_{i(3)} \cdot \boldsymbol{\nabla} \varphi_{i(3)} \, \mathrm{d}x \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \end{pmatrix} \\ & \quad = \begin{pmatrix} v_{i(1)} & v_{i(2)} & v_{i(3)} \end{pmatrix} \\ & \qquad \times \begin{pmatrix} a_i \left(\varphi_{i(1)}, \varphi_{i(1)} \right) & \cdots & a_i \left(\varphi_{i(1)}, \varphi_{i(3)} \right) \\ \vdots & \ddots & \vdots \\ a_i \left(\varphi_{i(3)}, \varphi_{i(1)} \right) & \cdots & a_i \left(\varphi_{i(3)}, \varphi_{i(3)} \right) \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \end{pmatrix} \\ & \quad = \bar{\boldsymbol{v}}_i \cdot \left(\bar{\boldsymbol{A}}_i \bar{\boldsymbol{u}}_i \right) = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^{\top} \bar{\boldsymbol{A}}_i \boldsymbol{Z}_i \bar{\boldsymbol{u}} \right) = \bar{\boldsymbol{v}} \cdot \left(\tilde{\boldsymbol{A}}_i \bar{\boldsymbol{u}} \right). \end{aligned} \tag{6.3.13}$$

$\bar{\boldsymbol{A}}_i$ is called the coefficient matrix of the finite element $i \in \mathcal{E}$. $\tilde{\boldsymbol{A}}_i$ is a matrix which has been expanded with the addition of 0 to match the total node value vectors.

If η_α and θ_α of Eq. (6.3.7) are substituted into Eq. (6.3.6), $\bar{\boldsymbol{A}}_i = \left(\bar{a}_{i(\alpha\beta)} \right)_{\alpha\beta} \in \mathbb{R}^{3 \times 3}$ is calculated as

$$\begin{aligned} \bar{a}_{i(\alpha\beta)} &= \int_{\Omega_i} \left(\frac{\partial \varphi_{i(\alpha)}}{\partial x_1} \frac{\partial \varphi_{i(\beta)}}{\partial x_1} + \frac{\partial \varphi_{i(\alpha)}}{\partial x_2} \frac{\partial \varphi_{i(\beta)}}{\partial x_2} \right) \, \mathrm{d}x \\ &= \int_{\Omega_i} \left(\eta_\alpha \eta_\beta + \theta_\alpha \theta_\beta \right) \, \mathrm{d}x = |\Omega_i| \left(\eta_\alpha \eta_\beta + \theta_\alpha \theta_\beta \right), \end{aligned} \tag{6.3.14}$$

where $|\Omega_i| = \gamma/2$ and γ is given by Eq. (6.3.8).

On the other hand, the right-hand side of Eq. (6.3.11) can also be split into elements as

$$\begin{aligned} l \left(v_h \left(\bar{\boldsymbol{v}} \right) \right) &= \sum_{i \in \mathcal{E}} \int_{\Omega_i} b v_h \left(\bar{\boldsymbol{v}}_i \right) \, \mathrm{d}x + \sum_{i \in \mathcal{E}} \int_{\partial \Omega_i \cap \Gamma_{\mathrm{N}}} p_{\mathrm{N}} v_h \left(\bar{\boldsymbol{v}}_i \right) \, \mathrm{d}\gamma \\ &= \sum_{i \in \mathcal{E}} l_i \left(v_h \left(\bar{\boldsymbol{v}}_i \right) \right), \end{aligned} \tag{6.3.15}$$

where

$$l_i \left(v_h \left(\bar{\boldsymbol{v}}_i \right) \right)$$

$$
\begin{aligned}
&= \begin{pmatrix} v_{i(1)} & v_{i(2)} & v_{i(3)} \end{pmatrix} \left(\begin{pmatrix} \int_{\Omega_i} b\varphi_{i(1)} \, \mathrm{d}x \\ \int_{\Omega_i} b\varphi_{i(2)} \, \mathrm{d}x \\ \int_{\Omega_i} b\varphi_{i(3)} \, \mathrm{d}x \end{pmatrix} + \begin{pmatrix} \int_{\partial\Omega_i \cap \Gamma_\mathrm{N}} p_\mathrm{N}\varphi_{i(1)} \, \mathrm{d}\gamma \\ \int_{\partial\Omega_i \cap \Gamma_\mathrm{N}} p_\mathrm{N}\varphi_{i(2)} \, \mathrm{d}\gamma \\ \int_{\partial\Omega_i \cap \Gamma_\mathrm{N}} p_\mathrm{N}\varphi_{i(3)} \, \mathrm{d}\gamma \end{pmatrix} \right) \\
&= \begin{pmatrix} v_{i(1)} & v_{i(2)} & v_{i(3)} \end{pmatrix} \left(\begin{pmatrix} b_{i(1)} \\ b_{i(2)} \\ b_{i(3)} \end{pmatrix} + \begin{pmatrix} p_{i(1)} \\ p_{i(2)} \\ p_{i(3)} \end{pmatrix} \right) \\
&= \bar{\boldsymbol{v}}_i \cdot \left(\bar{\boldsymbol{b}}_i + \bar{\boldsymbol{p}}_i \right) = \bar{\boldsymbol{v}} \cdot \left\{ \boldsymbol{Z}_i^\top \left(\bar{\boldsymbol{b}}_i + \bar{\boldsymbol{p}}_i \right) \right\} = \bar{\boldsymbol{v}} \cdot \left(\tilde{\boldsymbol{b}}_i + \tilde{\boldsymbol{p}}_i \right) \\
&= \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^\top \bar{\boldsymbol{l}}_i \right) = \bar{\boldsymbol{v}} \cdot \tilde{\boldsymbol{l}}_i
\end{aligned} \tag{6.3.16}
$$

with respect to $i \in \mathcal{E}$. Here, $\bar{\boldsymbol{l}}_i$ is called a known term vector of the finite element $i \in \mathcal{E}$. $\bar{\boldsymbol{b}}_i$ and $\bar{\boldsymbol{p}}_i$ represent components of the known term vectors with respect to b and p_N respectively. $\tilde{\boldsymbol{l}}_i$, $\tilde{\boldsymbol{b}}_i$ and $\tilde{\boldsymbol{p}}_i$ are vectors of $\bar{\boldsymbol{l}}_i$, $\bar{\boldsymbol{b}}_i$ and $\bar{\boldsymbol{p}}_i$ respectively which have been expanded with 0 added in to match the total node value vectors.

If b is a constant function, $\bar{\boldsymbol{b}}_i = \left(b_{i(1)}, b_{i(2)}, b_{i(3)} \right)^\top$ is calculated as

$$
\bar{\boldsymbol{b}}_i = b \begin{pmatrix} \int_{\Omega_i} \varphi_{i(1)} \, \mathrm{d}x \\ \int_{\Omega_i} \varphi_{i(2)} \, \mathrm{d}x \\ \int_{\Omega_i} \varphi_{i(3)} \, \mathrm{d}x \end{pmatrix} = \frac{b\,|\Omega_i|}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{6.3.17}
$$

Here, the following integration formula using area coordinates was used. Area coordinates are coordinates where a point $\boldsymbol{x} \in \Omega_i$ in the triangular finite element is represented by the three-dimensional vector with elements of basis functions $\varphi_{i(1)}(\boldsymbol{x})$, $\varphi_{i(2)}(\boldsymbol{x})$ and $\varphi_{i(3)}(\boldsymbol{x})$ of the triangular finite element (see Sect. 6.4.2 for reference).

Theorem 6.3.1 (Integrals of area coordinates) When $\left(\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)} \right)$ denotes the area coordinates on the two-dimensional triangular domain Ω_i,

$$
\int_{\Omega_i} \left(\varphi_{i(1)} \right)^l \left(\varphi_{i(2)} \right)^m \left(\varphi_{i(3)} \right)^n \, \mathrm{d}x = 2\,|\Omega_i| \frac{l!m!n!}{(l+m+n+2)!}
$$

holds with respect to non-negative integers l, m and n. Here, $|\Omega_i| = \gamma/2$ and γ is given by Eq. (6.3.8). □

Furthermore, when p_N is a constant function, $\bar{\boldsymbol{p}}_i = \left(p_{i(1)}, p_{i(2)}, p_{i(3)} \right)^\top$ becomes

$$
\bar{\boldsymbol{p}}_i = p_\mathrm{N} \begin{pmatrix} 0 \\ \int_{\partial\Omega_i \cap \Gamma_\mathrm{N}} \varphi_{i(2)} \, \mathrm{d}\gamma \\ \int_{\partial\Omega_i \cap \Gamma_\mathrm{N}} \varphi_{i(3)} \, \mathrm{d}\gamma \end{pmatrix} = \frac{p_\mathrm{N} h}{2} \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix},
$$

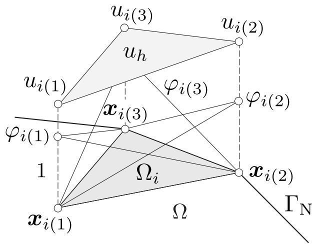


Fig. 6.15: Finite element including the boundary.

where h is the length of $\partial\Omega_i \cap \Gamma_{\mathrm{N}}$ (see Fig. 6.15).

Here, if Eq. (6.3.12) with Eq. (6.3.13) substituted in and Eq. (6.3.15) with Eq. (6.3.16) substituted in are substituted into the weak form (Eq. (6.3.11)), we get

$$\bar{\boldsymbol{v}} \cdot \left(\sum_{i\in\mathcal{E}} \tilde{\boldsymbol{A}}_i \bar{\boldsymbol{u}} \right) = \bar{\boldsymbol{v}} \cdot \sum_{i\in\mathcal{E}} \tilde{\boldsymbol{l}}_i .$$

This equation is written as

$$\bar{\boldsymbol{v}} \cdot \left(\bar{\boldsymbol{A}} \bar{\boldsymbol{u}} \right) = \bar{\boldsymbol{v}} \cdot \bar{\boldsymbol{l}}. \tag{6.3.18}$$

In other words,

$$\begin{aligned} \bar{\boldsymbol{A}} &= \sum_{i\in\mathcal{E}} \tilde{\boldsymbol{A}}_i \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}, \\ \bar{\boldsymbol{l}} &= \sum_{i\in\mathcal{E}} \tilde{\boldsymbol{l}}_i = \sum_{i\in\mathcal{E}} \left(\tilde{\boldsymbol{b}}_i + \tilde{\boldsymbol{p}}_i \right) = \bar{\boldsymbol{b}} + \bar{\boldsymbol{p}} \in \mathbb{R}^{|\mathcal{N}|}. \end{aligned}$$

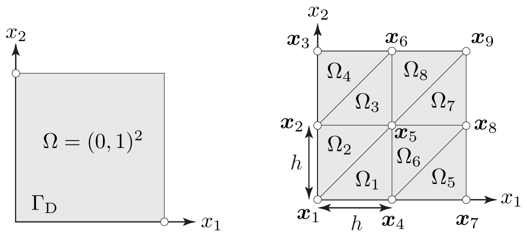
$\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{l}}$ are called the total coefficient matrix and total known term vector, respectively. Moreover, $\bar{\boldsymbol{b}}$ and $\bar{\boldsymbol{p}}$ are called total nodal value vectors of b and p_{N} respectively.

We substitute the fundamental boundary conditions into Eq. (6.3.18). In other words, if $u_j = u_{\mathrm{D}}(\boldsymbol{x}_j)$ and $v_j = 0$ are substituted in $j \in \mathcal{N}_{\mathrm{D}}$,

$$\begin{pmatrix} \boldsymbol{0}^{\top} & \bar{\boldsymbol{v}}_{\mathrm{N}}^{\top} \end{pmatrix} \left(\begin{pmatrix} \bar{\boldsymbol{A}}_{\mathrm{DD}} & \bar{\boldsymbol{A}}_{\mathrm{DN}} \\ \bar{\boldsymbol{A}}_{\mathrm{ND}} & \bar{\boldsymbol{A}}_{\mathrm{NN}} \end{pmatrix} \begin{pmatrix} \bar{\boldsymbol{u}}_{\mathrm{D}} \\ \bar{\boldsymbol{u}}_{\mathrm{N}} \end{pmatrix} - \begin{pmatrix} \bar{\boldsymbol{l}}_{\mathrm{D}} \\ \bar{\boldsymbol{l}}_{\mathrm{N}} \end{pmatrix} \right) = 0 \tag{6.3.19}$$

is obtained. Here, $\bar{\boldsymbol{u}}_{\mathrm{D}}$ and $\bar{\boldsymbol{u}}_{\mathrm{N}}$ are vectors defined in Eq. (6.3.1) and $\bar{\boldsymbol{v}}_{\mathrm{D}}$ and $\bar{\boldsymbol{v}}_{\mathrm{N}}$ are vectors defined in Eq. (6.3.2). Moreover,

$$\begin{aligned} \bar{\boldsymbol{A}}_{\mathrm{DD}} &= \left(\bar{A}_{ij} \right)_{i\in\mathcal{N}_{\mathrm{D}}\ j\in\mathcal{N}_{\mathrm{D}}}, \quad \bar{\boldsymbol{A}}_{\mathrm{DN}} = \left(\bar{A}_{ij} \right)_{i\in\mathcal{N}_{\mathrm{D}}\ j\in\mathcal{N}_{\mathrm{N}}}, \\ \bar{\boldsymbol{A}}_{\mathrm{ND}} &= \left(\bar{A}_{ij} \right)_{i\in\mathcal{N}_{\mathrm{N}}\ j\in\mathcal{N}_{\mathrm{D}}}, \quad \bar{\boldsymbol{A}}_{\mathrm{NN}} = \left(\bar{A}_{ij} \right)_{i\in\mathcal{N}_{\mathrm{N}}\ j\in\mathcal{N}_{\mathrm{N}}}, \\ \bar{\boldsymbol{l}}_{\mathrm{D}} &= \left(l_i \right)_{i\in\mathcal{N}_{\mathrm{D}}}, \quad \bar{\boldsymbol{l}}_{\mathrm{N}} = \left(l_i \right)_{i\in\mathcal{N}_{\mathrm{N}}}. \end{aligned}$$



(a) Domain Ω (b) Finite element partition

Fig. 6.16: An example of a 2D Poisson problem.

Rearranging Eq. (6.3.19) gives

$$\bar{\boldsymbol{v}}_{\mathrm{N}}^{\top}\left(\bar{\boldsymbol{A}}_{\mathrm{NN}}\bar{\boldsymbol{u}}_{\mathrm{N}} - \bar{\boldsymbol{l}}_{\mathrm{N}} + \bar{\boldsymbol{u}}_{\mathrm{D}}\bar{\boldsymbol{A}}_{\mathrm{ND}}\right) = 0.$$

Since $\bar{\boldsymbol{v}}_{\mathrm{N}}$ is arbitrary, we can write

$$\bar{\boldsymbol{A}}_{\mathrm{NN}}\bar{\boldsymbol{u}}_{\mathrm{N}} = \bar{\boldsymbol{l}}_{\mathrm{N}} - \bar{\boldsymbol{u}}_{\mathrm{D}}\bar{\boldsymbol{A}}_{\mathrm{ND}} = \hat{\boldsymbol{l}}. \tag{6.3.20}$$

Equation (6.3.20) is a simultaneous linear equation with respect to the unknown vector $\bar{\boldsymbol{u}}_{\mathrm{N}}$ and is called the discretized equation of the finite element method. If Eq. (6.3.20) is solved for $\bar{\boldsymbol{u}}_{\mathrm{N}}$, we get

$$\bar{\boldsymbol{u}}_{\mathrm{N}} = \bar{\boldsymbol{A}}_{\mathrm{NN}}^{-1}\hat{\boldsymbol{l}}. \tag{6.3.21}$$

From these, the finite element solution $u_h(\bar{\boldsymbol{u}})$ can be obtained by substituting $\bar{\boldsymbol{u}} = \left(\bar{\boldsymbol{u}}_{\mathrm{D}}, \bar{\boldsymbol{u}}_{\mathrm{N}}^{\top}\right)^{\top}$ into Eq. (6.3.1). Moreover, the finite element solution $u_h(\bar{\boldsymbol{u}}_i)$ in the domain Ω_i of the finite element $i \in \mathcal{E}$ can be sought via Eq. (6.3.9).

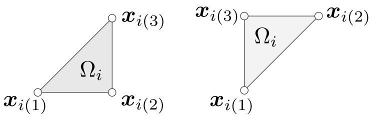
6.3.4 Exercise Problem

Let us look in detail at how an approximate solution of a two-dimensional Poisson problem can be obtained using the triangular finite element shown above ([3, Section 5.3, p. 67]).

Exercise 6.3.2 (Finite element method for 2D Poisson problem)
Let the domain Ω be $(0,1)^2$ and define the boundaries $\Gamma_{\mathrm{D}} = \{\boldsymbol{x} \in \partial\Omega \mid x_1 = 0,\ x_2 = 0\}$ and $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \bar{\Gamma}_{\mathrm{D}}$. In this case, obtain the approximate solution from the finite element method of $u : (0,1)^2 \to \mathbb{R}$ which satisfies

$$-\Delta u = 1 \quad \text{in } \Omega, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma_{\mathrm{N}}, \quad u = 0 \quad \text{on } \Gamma_{\mathrm{D}}.$$

Here, use the element partition in Fig. 6.16. □



(a) Type 1 (b) Type 2

Fig. 6.17: Finite element types.

Answer Let the size of the finite element be $h = 1/2$. In this case, from Eq. (6.3.8), we get $\gamma = h^2$ and $|\Omega_i| = \gamma/2 = h^2/2$. Let us calculate the coefficient matrix $\bar{\boldsymbol{A}}_i$ and $\bar{\boldsymbol{b}}_i$ using Eq. (6.3.14) and Eq. (6.3.17). Think about the finite element split into two types. Let the finite element $i \in \{1, 3, 5, 7\}$ of the form such as in Fig. 6.17 (a) be as type 1. The basis function $\varphi_{i(\alpha)}$ with respect to $\alpha \in \{1, 2, 3\}$ of type 1 was given by Eq. (6.3.6). The undetermined multipliers which are required in the calculation of the coefficient matrix $\bar{\boldsymbol{A}}_i$ are η_α and θ_α. These values with respect to type 1 are

$$\begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix} = \frac{1}{\gamma}\begin{pmatrix}x_{i(2)2} - x_{i(3)2}\\ x_{i(3)2} - x_{i(1)2}\\ x_{i(1)2} - x_{i(2)2}\end{pmatrix} = \frac{1}{h^2}\begin{pmatrix}-h\\ h\\ 0\end{pmatrix},$$

$$\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\end{pmatrix} = \frac{1}{\gamma}\begin{pmatrix}x_{i(3)1} - x_{i(2)1}\\ x_{i(1)1} - x_{i(3)1}\\ x_{i(2)1} - x_{i(1)1}\end{pmatrix} = \frac{1}{h^2}\begin{pmatrix}0\\ -h\\ h\end{pmatrix}$$

from Eq. (6.3.7). Substituting these into Eq. (6.3.14) and Eq. (6.3.17) gives the coefficient matrix and the known term vector as

$$\bar{\boldsymbol{A}}_i = \frac{1}{2}\begin{pmatrix}1 & -1 & 0\\ -1 & 2 & -1\\ 0 & -1 & 1\end{pmatrix}, \quad \bar{\boldsymbol{l}}_i = \bar{\boldsymbol{b}}_i = \frac{h^2}{6}\begin{pmatrix}1\\ 1\\ 1\end{pmatrix}.$$

On the other hand, the finite element $i \in \{2, 4, 6, 8\}$ in the form such as the one in Fig. 6.17 (b) is type 2. Similarly, with respect to these, the undetermined multipliers of the basis function can be obtained as

$$\begin{pmatrix}\eta_1\\ \eta_2\\ \eta_3\end{pmatrix} = \frac{1}{\gamma}\begin{pmatrix}x_{i(2)2} - x_{i(3)2}\\ x_{i(3)2} - x_{i(1)2}\\ x_{i(1)2} - x_{i(2)2}\end{pmatrix} = \frac{1}{h^2}\begin{pmatrix}0\\ h\\ -h\end{pmatrix},$$

$$\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\end{pmatrix} = \frac{1}{\gamma}\begin{pmatrix}x_{i(3)1} - x_{i(2)1}\\ x_{i(1)1} - x_{i(3)1}\\ x_{i(2)1} - x_{i(1)1}\end{pmatrix} = \frac{1}{h^2}\begin{pmatrix}-h\\ 0\\ h\end{pmatrix}.$$

Substituting these into Eq. (6.3.14) and Eq. (6.3.17) gives a type 2 coefficient matrix and known term vector as

$$\bar{\boldsymbol{A}}_i = \frac{1}{2}\begin{pmatrix}1 & 0 & -1\\ 0 & 1 & -1\\ -1 & -1 & 2\end{pmatrix}, \quad \bar{\boldsymbol{l}}_i = \bar{\boldsymbol{b}}_i = \frac{h^2}{6}\begin{pmatrix}1\\ 1\\ 1\end{pmatrix}.$$

The relationships of $\boldsymbol{x}_j$ with respect to the local nodes $\boldsymbol{x}_{i(1)}$, $\boldsymbol{x}_{i(2)}$, $\boldsymbol{x}_{i(3)}$ and total nodes $j \in \mathcal{N} = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ for the finite element $i \in \mathcal{E} = \{1, 2, 3, 4, 5, 6, 7, 8\}$ are shown in Table 6.1.

Table 6.1: Relationship between local nodes $\boldsymbol{x}_{i(1)}$, $\boldsymbol{x}_{i(2)}$, $\boldsymbol{x}_{i(3)}$ and the total nodes $\boldsymbol{x}_j$ in Exercise 6.3.2.

$i \in \mathcal{E}$	1	2	3	4	5	6	7	8
$\boldsymbol{x}_{i(1)}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_2$	$\boldsymbol{x}_4$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_5$
$\boldsymbol{x}_{i(2)}$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$	$\boldsymbol{x}_8$	$\boldsymbol{x}_9$
$\boldsymbol{x}_{i(3)}$	$\boldsymbol{x}_5$	$\boldsymbol{x}_2$	$\boldsymbol{x}_6$	$\boldsymbol{x}_3$	$\boldsymbol{x}_8$	$\boldsymbol{x}_5$	$\boldsymbol{x}_9$	$\boldsymbol{x}_6$
Type	1	2	1	2	1	2	1	2

We expand $\bar{\boldsymbol{A}}_1$ and $\bar{\boldsymbol{l}}_1$ in conjunction with the total node value vectors as

$$\tilde{\boldsymbol{A}}_1 = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{\boldsymbol{l}}_1 = \frac{h^2}{6}\begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Similarly, if $\tilde{\boldsymbol{A}}_i$ and $\tilde{\boldsymbol{l}}_i$ are formed using $\bar{\boldsymbol{A}}_i$ and $\bar{\boldsymbol{l}}_i$ with respect to $i \in \{2, \ldots, 8\}$ and superimposed, one obtains

$$\bar{\boldsymbol{A}} = \frac{1}{2}\begin{pmatrix} 2 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -2 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -2 & 0 & -1 & 0 & 0 \\ 0 & -2 & 0 & -2 & 8 & -2 & 0 & -2 & 0 \\ 0 & 0 & -1 & 0 & -2 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & -2 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 2 \end{pmatrix}, \quad \bar{\boldsymbol{l}} = \frac{h^2}{6}\begin{pmatrix} 2 \\ 3 \\ 1 \\ 3 \\ 6 \\ 3 \\ 1 \\ 3 \\ 2 \end{pmatrix}.$$

Moreover, if the fundamental boundary conditions $u_1 = u_2 = u_3 = u_4 = u_7 = 0$, $v_1 = v_2 = v_3 = v_4 = v_7 = 0$ are substituted in as Eq. (6.3.19), we get

$$\frac{1}{2}\begin{pmatrix} 8 & -2 & -2 & 0 \\ -2 & 4 & 0 & -1 \\ -2 & 0 & 4 & -1 \\ 0 & -1 & -1 & 2 \end{pmatrix}\begin{pmatrix} u_5 \\ u_6 \\ u_8 \\ u_9 \end{pmatrix} = \frac{1}{24}\begin{pmatrix} 6 \\ 3 \\ 3 \\ 2 \end{pmatrix}.$$

Solving this simultaneous linear equation,

$$\begin{pmatrix} u_5 \\ u_6 \\ u_8 \\ u_9 \end{pmatrix} = \frac{1}{192}\begin{pmatrix} 3 & 2 & 2 & 2 \\ 2 & 6 & 2 & 4 \\ 2 & 2 & 6 & 4 \\ 2 & 4 & 4 & 12 \end{pmatrix}\begin{pmatrix} 6 \\ 3 \\ 3 \\ 2 \end{pmatrix} = \frac{1}{96}\begin{pmatrix} 17 \\ 22 \\ 22 \\ 30 \end{pmatrix}$$

is obtained. Figure 6.18 shows the approximate solution $u_h(\bar{\boldsymbol{u}})$ with this result as a node value. □

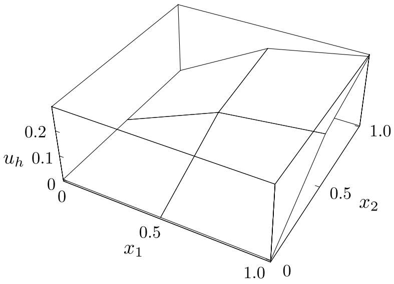


Fig. 6.18: Approximate solution of Exercise 6.3.2.

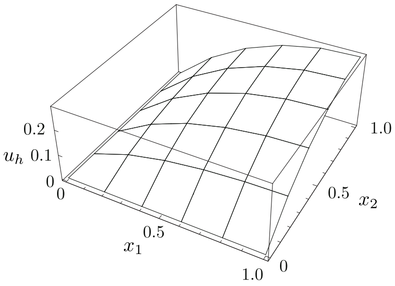


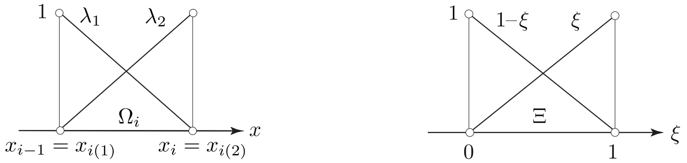
Fig. 6.19: Approximate solution when increasing the division number of Exercise 6.3.2.

Figure 6.19 shows the result when the division number in Exercise 6.3.2 is increased to 36 nodes and 50 finite elements.

6.4 Various Finite Elements

The finite element methods in one and two dimensions when the basis function is constructed of linear functions were seen in Sections 6.2 and 6.3. Next, let us show how to change the basis function to a higher order and how to change the shape of a finite element into a square, and give an overview of the three-dimensional finite element method. Other than the method of construction of the basis function being different, their procedures, such as the Galerkin method seeking the undetermined multipliers by substituting them into the weak form, are identical. So let us look at how the basis function is defined for a finite element and how the approximate function is constructed.

In this chapter, the domain of the basis function is changed to a domain whose size is normalized (normal domain), and a finite element defined on such a domain is referred to as normal element. A finite element of arbitrary size is thought to be given by a mapping from a normal element.



(a) Length coordinates (λ_1, λ_2). (b) Normal coordinates $\xi = \lambda_2 = 1 - \lambda_1$.

Fig. 6.20: Length coordinates and normal coordinates of 1D finite elements.

6.4.1 One-Dimensional Higher-Order Finite Elements

Firstly, let us think about making the one-dimensional finite element a higher-order. Let the normal domain with respect to the one-dimensional finite element be $\Xi = (0, 1)$. The point $x \in \Omega_i$ on the domain $\Omega_i = (x_{i-1}, x_i)$ of the one-dimensional finite element $i \in \mathcal{E}$ can be changed to be a point on the normal domain $\xi \in \Xi$ by using the basis functions $\varphi_{i(1)}$ and $\varphi_{i(2)}$ of the one-dimensional finite element (first-order element) defined by Eq. (6.2.3) and Eq. (6.2.4) via $\xi = \varphi_{i(2)}(x) = 1 - \varphi_{i(1)}(x)$. Here, $\varphi_{i(2)}(x) = 1 - \varphi_{i(1)}(x)$ and $\xi \in \Xi$ are viewed as the same and defined $\left(\varphi_{i(1)}(x), \varphi_{i(2)}(x)\right)$ as the length coordinate. It is a two-dimensional vector to express length, but it should be noted that the condition $\varphi_{i(1)}(x) + \varphi_{i(2)}(x) = 1$ is imposed. Furthermore, using functions of $\left(\varphi_{i(1)}(x), \varphi_{i(2)}(x)\right)$ as coordinates may bring confusion so the length coordinate is written as (λ_1, λ_2). Figure 6.20 shows the relationship between the length coordinates (λ_1, λ_2) and normal coordinates ξ.

Hereafter, we consider the basis functions defined on the normal domain $\Xi = (0, 1)$ and the basis functions will be made higher-order. As a preparation for this, let us confirm how the basis functions $\left(\varphi_{i(1)}(x), \varphi_{i(2)}(x)\right)$ defined with respect to $x \in \Omega_i$ seen in Sect. 6.2 is expressed using the basis functions defined with respect to $\xi \in \Xi$. In this chapter, the node and basis function defined on the normal domain are to be expressed as $\xi_{(\cdot)}$ and $\hat{\varphi}_{(\cdot)}$, respectively. In other words, let the node with respect to first-order basis function be

$$\begin{pmatrix}\xi_{(1)} & \xi_{(2)}\end{pmatrix} = \begin{pmatrix}0 & 1\end{pmatrix} \tag{6.4.1}$$

and the basis function be

$$\begin{pmatrix}\hat{\varphi}_{(1)}(\xi) & \hat{\varphi}_{(2)}(\xi)\end{pmatrix} = \begin{pmatrix}\lambda_1 & \lambda_2\end{pmatrix} = \begin{pmatrix}1 - \xi & \xi\end{pmatrix}. \tag{6.4.2}$$

On the other hand, the mapping $f_i : \Xi \to \Omega_i$ from normal coordinate $\xi \in \Xi$ to the global coordinate $x \in \Omega_i$ is given by

$$x = f_i(\xi) = x_{i(1)} + \xi\left(x_{i(2)} - x_{i(1)}\right). \tag{6.4.3}$$

Here, between the first-order basis functions $\left(\varphi_{i(1)}, \varphi_{i(2)}\right)$ defined on Ω_i and the first-order basis functions $\left(\hat{\varphi}_{(1)}, \hat{\varphi}_{(2)}\right)$ defined on the normal domain,

$$\begin{pmatrix}\varphi_{i(1)}(f_i(\xi)) & \varphi_{i(2)}(f_i(\xi))\end{pmatrix} = \begin{pmatrix}\hat{\varphi}_{(1)}(\xi) & \hat{\varphi}_{(2)}(\xi)\end{pmatrix} \tag{6.4.4}$$

holds.

Let us define second-order basis functions corresponding to the definition of first-order basis functions. Second-order functions have three undetermined multipliers. Hence, adding a mid-node, we define the nodes as

$$\begin{pmatrix}\xi_{i(1)} & \xi_{i(2)} & \xi_{i(3)}\end{pmatrix} = \begin{pmatrix}0 & 1 & 1/2\end{pmatrix}. \tag{6.4.5}$$

The basis functions $\hat{\varphi}_{(1)}$, $\hat{\varphi}_{(2)}$ and $\hat{\varphi}_{(3)}$ of normal element are determined so that they satisfy the boundary conditions at $\xi_{(1)}$, $\xi_{(2)}$ and $\xi_{(3)}$, respectively. These conditions can be given by

$$\hat{\varphi}_{(\alpha)}\left(\xi_{(\beta)}\right) = \delta_{\alpha\beta} \tag{6.4.6}$$

with respect to $\alpha, \beta \in \{1, 2, 3\}$. Equation (6.4.6) shows conditions for the undetermined multipliers to be the node values of an approximate function. From this condition, the three undetermined multipliers of a second-order function are determined as

$$\begin{pmatrix}\hat{\varphi}_{(1)} & \hat{\varphi}_{(2)} & \hat{\varphi}_{(3)}\end{pmatrix} = \begin{pmatrix}\lambda_1(2\lambda_1 - 1) & \lambda_2(2\lambda_2 - 1) & 4\lambda_1\lambda_2\end{pmatrix}. \tag{6.4.7}$$

Basis functions determined in this way satisfies

$$\sum_{\alpha \in \{1,2,3\}} \hat{\varphi}_{(\alpha)}(\xi) = 1 \tag{6.4.8}$$

with respect to all $\xi \in \Xi$. Equation (6.4.8) is the condition at which the approximate solution matches the exact solution when the exact solution is $u = 1$. Figure 6.21 shows these basis functions and the basis functions defined on Ω_i. Here, let l, m, and n in the figure be the node numbers given in total nodes.

Therefore, the approximate function used in a second-order one-dimensional finite element is constructed as

$$\hat{u}_h(\xi) = \begin{pmatrix}\hat{\varphi}_{(1)}(\xi) & \hat{\varphi}_{(2)}(\xi) & \hat{\varphi}_{(3)}(\xi)\end{pmatrix}\begin{pmatrix}u_{i(1)} \\ u_{i(2)} \\ u_{i(3)}\end{pmatrix} = \hat{\boldsymbol{\varphi}}(\xi) \cdot \boldsymbol{u}_i$$

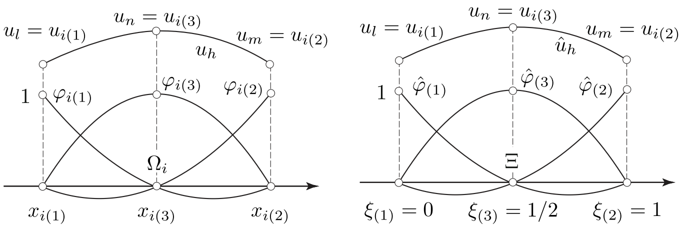
on the normal domain. Here, in this chapter the approximate function defined on a normal domain will be expressed as $\hat{u}_h$.

Similarly, the one-dimensional finite element of $m(\in \mathbb{N})$-th order is constructed in the following way.

Definition 6.4.1 (m-th order one-dimensional finite element) Let $\Xi = (0, 1)$ be a normal domain. With respect to $m \in \mathbb{N}$, place the nodes as

$$\begin{pmatrix}\xi_{i(1)} & \xi_{i(2)} & \cdots & \xi_{i(m+1)}\end{pmatrix} = \begin{pmatrix}0 & 1/m & \cdots & 1\end{pmatrix}.$$

Construct the basis functions $\hat{\varphi}_{(\alpha)}$ with respect to $\alpha \in \{1, \dots, m+1\}$ as an m-th order polynomial and select the undetermined multipliers so that Eq. (6.4.6) with respect to $\beta \in \{1, \dots, m+1\}$ is satisfied. A finite element using basis functions constructed in this way is called a one-dimensional m-th order finite element. □



(a) Basis functions on Ω_i. (b) Basis functions on Ξ.

Fig. 6.21: Basis functions and an approximate function used in a second-order 1D finite element.

When basis functions $\hat{\varphi}_{(\alpha)}$ are defined as in Definition 6.4.1, an approximate function is constructed by

$$\hat{u}_h = \sum_{\alpha \in \mathcal{N}_i} \hat{\varphi}_{(\alpha)} u_{i(\alpha)} = \hat{\boldsymbol{\varphi}} \cdot \bar{\boldsymbol{u}}_i \tag{6.4.9}$$

on the normal domain, where $\mathcal{N}_i$ is a set of local node numbers. Equation (6.4.9) expresses the relationship which holds with respect to the approximate function, not limited to an m-th order one-dimensional finite element.

In this way, if approximate functions are given on the normal coordinates $\xi \in \Xi$, the bilinear form for each finite element in the weak form (Eq. (6.2.13)) is

$$\begin{aligned} & a_i\left(u_h\left(\bar{\boldsymbol{u}}_i\right), v_h\left(\bar{\boldsymbol{v}}_i\right)\right) \\ & \quad = \begin{pmatrix} v_{i(1)} & \cdots & v_{i(m+1)} \end{pmatrix} \\ & \quad \times \begin{pmatrix} a_i\left(\varphi_{i(1)}, \varphi_{i(1)}\right) & \cdots & a_i\left(\varphi_{i(1)}, \varphi_{i(m+1)}\right) \\ \vdots & \ddots & \vdots \\ a_i\left(\varphi_{i(m+1)}, \varphi_{i(1)}\right) & \cdots & a_i\left(\varphi_{i(m+1)}, \varphi_{i(m+1)}\right) \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ \vdots \\ u_{i(m+1)} \end{pmatrix} \\ & \quad = \bar{\boldsymbol{v}}_i \cdot \left(\bar{\boldsymbol{A}}_i \bar{\boldsymbol{u}}_i\right) = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^{\top} \bar{\boldsymbol{A}}_i \boldsymbol{Z}_i \bar{\boldsymbol{u}}\right) = \bar{\boldsymbol{v}} \cdot \left(\tilde{\boldsymbol{A}}_i \bar{\boldsymbol{u}}\right). \end{aligned} \tag{6.4.10}$$

Here, we can write

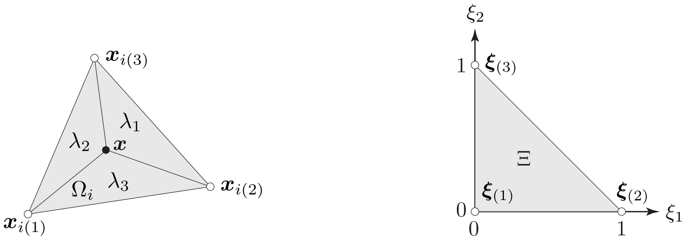
$$\frac{\mathrm{d}\varphi_{i(\alpha)}}{\mathrm{d}x} = \frac{\mathrm{d}\hat{\varphi}_{(\alpha)}}{\mathrm{d}\xi} \frac{\mathrm{d}\xi}{\mathrm{d}x} = \frac{1}{\omega_i} \frac{\mathrm{d}\hat{\varphi}_{(\alpha)}}{\mathrm{d}\xi},$$

where

$$\omega_i = \frac{\mathrm{d}x}{\mathrm{d}\xi} = \frac{\mathrm{d}f_i}{\mathrm{d}\xi} = x_{i(2)} - x_{i(1)}$$

from Eq. (6.4.3). Hence, each element of $\bar{\boldsymbol{A}}_i = \left(\bar{a}_{i(\alpha\beta)}\right)_{\alpha\beta} \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$ can be calculated from

$$\bar{a}_{i(\alpha\beta)} = \int_{\Omega_i} \frac{\mathrm{d}\varphi_{i(\alpha)}}{\mathrm{d}x} \frac{\mathrm{d}\varphi_{i(\beta)}}{\mathrm{d}x} \,\mathrm{d}x = \frac{1}{\omega_i} \int_0^1 \frac{\mathrm{d}\hat{\varphi}_{(\alpha)}}{\mathrm{d}\xi} \frac{\mathrm{d}\hat{\varphi}_{(\beta)}}{\mathrm{d}\xi} \,\mathrm{d}\xi.$$



(a) Area coordinates $(\lambda_1, \lambda_2, \lambda_3)$. (b) Normal coordinates $(\xi_1, \xi_2) = (\lambda_2, \lambda_3)$.

Fig. 6.22: Area coordinates and normal coordinates for a triangular finite element.

Moreover, even with respect to the linear form (Eq. (6.2.16)) for each finite element in the weak form, we have

$$\begin{aligned} l_i\left(v_h\left(\bar{\boldsymbol{v}}_i\right)\right) &= \begin{pmatrix} v_{i(1)} & \cdots & v_{i(m+1)} \end{pmatrix} \begin{pmatrix} b_{i(1)} \\ \vdots \\ b_{i(m+1)} \end{pmatrix} \\ &= \bar{\boldsymbol{v}}_i \cdot \bar{\boldsymbol{b}}_i = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^\top \bar{\boldsymbol{b}}_i\right) = \bar{\boldsymbol{v}} \cdot \tilde{\boldsymbol{b}}_i \\ &= \bar{\boldsymbol{v}}_i \cdot \bar{\boldsymbol{l}}_i = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^\top \bar{\boldsymbol{l}}_i\right) = \bar{\boldsymbol{v}} \cdot \tilde{\boldsymbol{l}}_i. \end{aligned} \tag{6.4.11}$$

Here, each element of $\bar{\boldsymbol{b}}_i = \left(\bar{b}_{i(\alpha)}\right)_\alpha \in \mathbb{R}^{|\mathcal{N}_i|}$ can be calculated by

$$\bar{b}_{i(\alpha)} = \int_{\Omega_i} b_0 \varphi_{i(\alpha)} \, \mathrm{d}x = \omega_i \int_0^1 b_0 \hat{\varphi}_{(\alpha)} \, \mathrm{d}\xi.$$

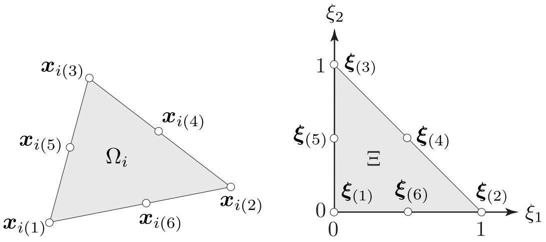
A similar relationship holds with respect to Eq. (6.2.17) too.

6.4.2 Triangular Higher-Order Finite Elements

Next, let us think about the higher-order triangular finite element used with respect to a two-dimensional problem. Basis functions $\varphi_{i(1)}$, $\varphi_{i(2)}$ and $\varphi_{i(3)}$ of the first-order triangular finite element defined by Eq. (6.3.6) are called area coordinates. The reason for this is because when a point $\boldsymbol{x} \in \Omega_i$ on the domain Ω_i of finite element $i \in \mathcal{E}$ is selected as in Fig. 6.22 and λ_1, λ_2 and λ_3 are the area ratios of three small triangles with $\boldsymbol{x}$ as a vertex against $|\Omega_i|$, $(\lambda_1, \lambda_2, \lambda_3) = \left(\varphi_{i(1)}, \varphi_{i(2)}, \varphi_{i(3)}\right)$ is established. From this, $\boldsymbol{\xi} = (\xi_1, \xi_2)^\top = (\lambda_2, \lambda_3)^\top \in \mathbb{R}^2$ is called the normal coordinates and $\Xi = \left\{ \boldsymbol{\xi} \in (0,1)^2 \,\middle|\, \xi_1 + \xi_2 < 1 \right\}$ is called a normal domain.

Let us again think about the higher-order after defining linear basis functions on the normal domain. Let nodes with respect to normal element be

$$\begin{pmatrix} \boldsymbol{\xi}_{(1)} & \boldsymbol{\xi}_{(2)} & \boldsymbol{\xi}_{(3)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{6.4.12}$$



(a) Nodes on Ω_i (b) Nodes on Ξ

Fig. 6.23: Nodes used in second-order triangular finite element.

The basis functions looked at in Sect. 6.3 were

$$\begin{pmatrix}\hat{\varphi}_{(1)}(\boldsymbol{\xi}) & \hat{\varphi}_{(2)}(\boldsymbol{\xi}) & \hat{\varphi}_{(3)}(\boldsymbol{\xi})\end{pmatrix} = \begin{pmatrix}\lambda_1 & \lambda_2 & \lambda_3\end{pmatrix}. \tag{6.4.13}$$

Let us define the second-order basis functions corresponding to the definition of the first-order basis functions. Complete second-order polynomials with respect to ξ_1 and ξ_2 have six undetermined multipliers $a_1, \ldots, a_6$ given as

$$a_1 + a_2\xi_1 + a_3\xi_2 + a_4\xi_1^2 + a_5\xi_1\xi_2 + a_6\xi_2^2.$$

Then, mid-point nodes can be added to the three sides of the triangle as shown in Fig. 6.23. Let the nodes be

$$\begin{pmatrix}\boldsymbol{\xi}_{(1)} & \boldsymbol{\xi}_{(2)} & \boldsymbol{\xi}_{(3)} & \boldsymbol{\xi}_{(4)} & \boldsymbol{\xi}_{(5)} & \boldsymbol{\xi}_{(6)}\end{pmatrix} = \begin{pmatrix}0 & 1 & 0 & 1/2 & 0 & 1/2\\ 0 & 0 & 1 & 1/2 & 1/2 & 0\end{pmatrix}. \tag{6.4.14}$$

Here, if the basis functions $\hat{\varphi}_{(\alpha)}$ of the normal element with respect to $\alpha \in \{1, \ldots, 6\}$ are determined so that $\left(\boldsymbol{\xi}_{(\beta)}\right) = \delta_{\alpha\beta}$ is satisfied with respect to $\beta \in \{1, \ldots, 6\}$, we get

$$\begin{pmatrix}\hat{\varphi}_{(1)}\\ \hat{\varphi}_{(2)}\\ \hat{\varphi}_{(3)}\end{pmatrix} = \begin{pmatrix}\lambda_1(2\lambda_1 - 1)\\ \lambda_2(2\lambda_2 - 1)\\ \lambda_3(2\lambda_3 - 1)\end{pmatrix}, \quad \begin{pmatrix}\hat{\varphi}_{(4)}\\ \hat{\varphi}_{(5)}\\ \hat{\varphi}_{(6)}\end{pmatrix} = \begin{pmatrix}4\lambda_2\lambda_3\\ 4\lambda_1\lambda_3\\ 4\lambda_1\lambda_2\end{pmatrix}.$$

Figure 6.24 shows these functions. An approximate function can be constructed using these as Eq. (6.4.9).

Similarly, an $m(\in \mathbb{N})$-th-order triangular finite element can be constructed in the following way from a two-dimensional m-th order complete polynomial.

Definition 6.4.2 (m-th order triangular finite element)
Let $\Xi = \left\{\boldsymbol{\xi} \in (0,1)^2 \,\middle|\, \xi_1 + \xi_2 < 1\right\}$ be a normal domain. For $m \in \mathbb{N}$, place the nodes $\boldsymbol{\xi}_{(\alpha)}$ with respect to $\alpha \in \mathcal{N}_i$ as shown in Fig. 6.25. Construct the basis functions $\hat{\varphi}_{(\alpha)}$ with a complete m-th order polynomial with respect to ξ_1 and ξ_2 and determine the undetermined multipliers so that $\hat{\varphi}_{(\alpha)}\left(\boldsymbol{\xi}_{(\beta)}\right) = \delta_{\alpha\beta}$ is

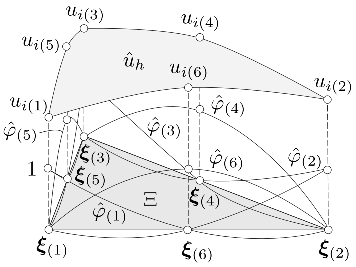
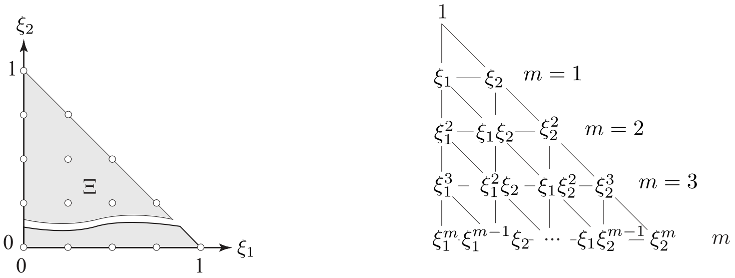


Fig. 6.24: Basis functions and an approximate function used in the second-order triangular finite element.


ξ_2
1
Ξ
0
0
1
ξ_1
1
ξ_1 ξ_2 $m = 1$
ξ_1^2 $\xi_1\xi_2$ ξ_2^2 $m = 2$
ξ_1^3 $\xi_1^2\xi_2$ $\xi_1\xi_2^2$ ξ_2^3 $m = 3$
ξ_1^m $\xi_1^{m-1}\xi_2$ $\cdots$ $\xi_1\xi_2^{m-1}$ ξ_2^m m

(a) Nodes $\boldsymbol{\xi}_{(\alpha)}$ for $\alpha \in \mathcal{N}_i$. (b) Complete m-th order polynomial terms of $|\mathcal{N}_i|$.

Fig. 6.25: Node placement and polynomial terms used in the m-th order triangular finite element.

satisfied with respect to $\alpha, \beta \in \mathcal{N}_i$. The finite element using the basis functions constructed in this way is called triangular m-th order finite element. □

When using the basis functions $\hat{\varphi}_{(\alpha)}$ constructed as in Definition 6.4.2, an approximate function can be constructed on the normal domain as Eq. (6.4.9). If approximate functions are given, the bilinear form (Eq. (6.3.14)) for each finite element in the weak form can be calculated by

$$\begin{aligned}\bar{a}_{i(\alpha\beta)} &= \int_{\Omega_i} \partial_{\boldsymbol{x}} \varphi_{i(\alpha)}(\boldsymbol{x}) \cdot \partial_{\boldsymbol{x}} \varphi_{i(\beta)}(\boldsymbol{x}) \;\mathrm{d}x \\ &= \int_{\Xi} \left\{ \left(\boldsymbol{F}_i^{\top} \right)^{-1} \partial_{\boldsymbol{\xi}} \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi}) \right\} \cdot \left\{ \left(\boldsymbol{F}_i^{\top} \right)^{-1} \partial_{\boldsymbol{\xi}} \hat{\varphi}_{(\beta)}(\boldsymbol{\xi}) \right\} \omega_i \;\mathrm{d}\xi.\end{aligned} \tag{6.4.15}$$

Here, the mapping from the normal domain Ξ to the domain Ω_i of the finite element $i \in \mathcal{E}$ was assumed to be given by

$$\begin{aligned}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} x_{i(2)1} - x_{i(1)1} & x_{i(3)1} - x_{i(1)1} \\ x_{i(2)2} - x_{i(1)2} & x_{i(3)2} - x_{i(1)2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} x_{i(1)1} \\ x_{i(1)2} \end{pmatrix} \\ &= \boldsymbol{F}_i \boldsymbol{\xi} + \boldsymbol{x}_{i(1)}.\end{aligned} \tag{6.4.16}$$

Here, $\boldsymbol{x}_{i(\alpha)} = \left(x_{i(\alpha)1}, x_{i(\alpha)2}\right)^{\top}$ are the coordinate values of local nodes $\alpha \in \{1,2,3\}$ of the finite element $i \in \mathcal{E}$. A mapping such as combining a linear mapping and a translation with a fixed element is called affine mapping. In this mapping, $\boldsymbol{F}_i$ represents a Jacobi matrix and ω_i represents a Jacobian $\det \boldsymbol{F}_i$. Moreover,

$$\begin{aligned}\partial_{\boldsymbol{\xi}} \hat{\varphi}_{(\alpha)}\left(\boldsymbol{\xi}\right) = \begin{pmatrix} \partial\hat{\varphi}_{(\alpha)}/\partial\xi_1 \\ \partial\hat{\varphi}_{(\alpha)}/\partial\xi_2 \end{pmatrix} &= \begin{pmatrix} \partial x_1/\partial\xi_1 & \partial x_2/\partial\xi_1 \\ \partial x_1/\partial\xi_2 & \partial x_2/\partial\xi_2 \end{pmatrix} \begin{pmatrix} \partial\hat{\varphi}_{(\alpha)}/\partial x_1 \\ \partial\hat{\varphi}_{(\alpha)}/\partial x_2 \end{pmatrix} \\ &= \boldsymbol{F}_i^{\top} \partial_{\boldsymbol{x}} \hat{\varphi}_{(\alpha)}\left(\boldsymbol{\xi}\right)\end{aligned}$$

was used to get Eq. (6.4.15).

Furthermore, the linear form (Eq. (6.3.16)) for each finite element in the weak form becomes

$$\begin{aligned} l_i\left(v_h\left(\bar{\boldsymbol{v}}_i\right)\right) &= \bar{\boldsymbol{v}}_i \cdot \left(\bar{\boldsymbol{b}}_i + \bar{\boldsymbol{p}}_i\right) = \bar{\boldsymbol{v}} \cdot \left\{\boldsymbol{Z}_i^{\top}\left(\bar{\boldsymbol{b}}_i + \bar{\boldsymbol{p}}_i\right)\right\} = \bar{\boldsymbol{v}} \cdot \left(\tilde{\boldsymbol{b}}_i + \tilde{\boldsymbol{p}}_i\right) \\ &= \bar{\boldsymbol{v}}_i \cdot \bar{\boldsymbol{l}}_i = \bar{\boldsymbol{v}} \cdot \left(\boldsymbol{Z}_i^{\top} \bar{\boldsymbol{l}}_i\right) = \bar{\boldsymbol{v}} \cdot \tilde{\boldsymbol{l}}_i. \end{aligned} \tag{6.4.17}$$

Here, each of the elements in $\bar{\boldsymbol{b}}_i = \left(\bar{b}_{i(\alpha)}\right)_{\alpha} \in \mathbb{R}^{|\mathcal{N}_i|}$ and $\bar{\boldsymbol{p}}_i = \left(\bar{p}_{i(\alpha)}\right)_{\alpha} \in \mathbb{R}^{|\mathcal{N}_i|}$ are calculated from

$$\bar{b}_{i(\alpha)} = \int_{\Omega_i} b_0 \varphi_{i(\alpha)} \, \mathrm{d}x = \omega_i \int_{\Xi} b_0 \hat{\varphi}_{(\alpha)} \, \mathrm{d}\xi, \tag{6.4.18}$$

$$\bar{p}_{i(\alpha)} = \int_{\partial\Omega_i \cap \Gamma_{\mathrm{N}}} \varphi_{i(2)} \, \mathrm{d}\gamma = \omega_{i1\mathrm{D}} \int_0^1 p_{\mathrm{N}} \hat{\varphi}_{(\alpha)} \, \mathrm{d}\xi_1, \tag{6.4.19}$$

where $\omega_{i1\mathrm{D}} = \mathrm{d}\gamma/\mathrm{d}\xi_1 = |\partial\Omega_i \cap \Gamma_{\mathrm{N}}|$.

The integral on the triangular normal domain Ξ in the above equation can be calculated using the formula in Theorem 6.3.1.

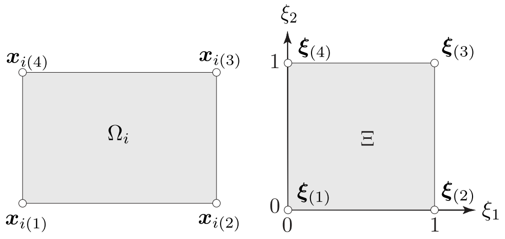
6.4.3 Rectangular Finite Elements

A rectangular finite element can also be considered with respect to two-dimensional problems. Consider two-dimensional domain Ω that can be divided with rectangular domains Ω_i $(i \in \mathcal{E})$ such as the one in Fig. 6.26 (a). Let the normal domain Ξ be $(0,1)^2$ such as the one in Fig. 6.26 (b). The change from $\boldsymbol{x} \in \Omega_i$ to $\boldsymbol{\xi} \in \Xi$ is given by $\boldsymbol{\xi} = (\xi_1, \xi_2)^{\top} = (\lambda_{12}, \lambda_{22})^{\top} \in \Xi$ based on length coordinate in the x_1 direction $(\lambda_{11}, \lambda_{12})$ and length coordinates in the x_2 direction $(\lambda_{21}, \lambda_{22})$ which are defined as

$$\lambda_{11}\left(\boldsymbol{x}\right) = \frac{x_{i(2)1} - x_1}{x_{i(2)1} - x_{i(1)1}}, \tag{6.4.20}$$

$$\lambda_{12}\left(\boldsymbol{x}\right) = \frac{x_1 - x_{i(1)1}}{x_{i(2)1} - x_{i(1)1}}, \tag{6.4.21}$$

$$\lambda_{21}\left(\boldsymbol{x}\right) = \frac{x_{i(4)2} - x_2}{x_{i(4)2} - x_{i(1)2}}, \tag{6.4.22}$$



(a) Node on Ω_i (b) node on Ξ

Fig. 6.26: Node of a rectangular finite element.

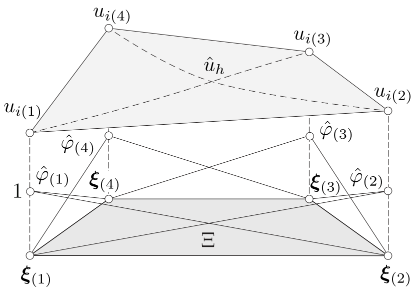


Fig. 6.27: Basis functions and approximate function used in a first-order rectangular finite element.

$$\lambda_{22}(\boldsymbol{x}) = \frac{x_2 - x_{i(1)2}}{x_{i(4)2} - x_{i(1)2}}. \tag{6.4.23}$$

In a first-order rectangular finite element, the bilinear polynomial

$$a_1 + a_2\xi_1 + a_3\xi_2 + a_4\xi_1\xi_2$$

with respect to ξ_1 and ξ_2 is used for the basis function. Four undetermined multipliers $a_1, \ldots, a_4$ are determined by the boundary conditions of four nodes of the normal element:

$$\begin{pmatrix} \boldsymbol{\xi}_{(1)} & \boldsymbol{\xi}_{(2)} & \boldsymbol{\xi}_{(3)} & \boldsymbol{\xi}_{(4)} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{6.4.24}$$

In reality, using the conditions $\hat{\varphi}_{(\alpha)}\left(\boldsymbol{\xi}_{(\beta)}\right) = \delta_{\alpha\beta}$ with respect to $\alpha, \beta \in \{1, 2, 3, 4\}$,

$$\begin{pmatrix} \hat{\varphi}_{(1)} & \hat{\varphi}_{(2)} & \hat{\varphi}_{(3)} & \hat{\varphi}_{(4)} \end{pmatrix} = \begin{pmatrix} \lambda_{11}\lambda_{21} & \lambda_{12}\lambda_{21} & \lambda_{12}\lambda_{22} & \lambda_{11}\lambda_{22} \end{pmatrix}$$

can be obtained as the basis functions of a normal element. Figure 6.27 shows these basis functions. Using these, the approximate function of the normal element can be constructed as Eq. (6.4.9).

Two methods are known for forming higher-order rectangular finite elements. One is a method such as the one below using bi-m-th order polynomials.

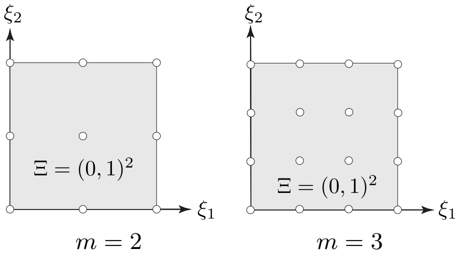


Fig. 6.28: Nodes of a Lagrange family rectangular finite element.

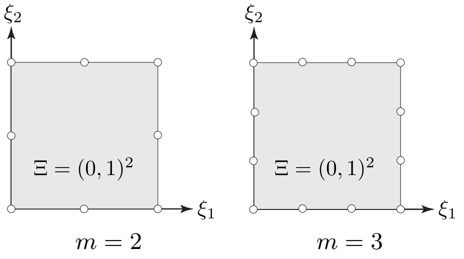


Fig. 6.29: Nodes of a serendipity group rectangular finite element.

Definition 6.4.3 (Lagrange family rectangular finite element) Let $\Xi = (0,1)^2$ be a normal domain. For $m \in \mathbb{N}$, nodes $\boldsymbol{\xi}_{(\alpha)}$ with respect to $\alpha \in \mathcal{N}_i$ are placed as in Fig. 6.28. A basis function $\hat{\varphi}_{(\alpha)}$ is constructed using bi-m-th order polynomials with respect to ξ_1 and ξ_2 and undetermined multipliers are determined so that $\hat{\varphi}_{(\alpha)}\left(\boldsymbol{\xi}_{(\beta)}\right) = \delta_{\alpha\beta}$ is satisfied with respect to $\alpha, \beta \in \mathcal{N}_i$. A finite element using the basis functions constructed in this way is called a Lagrange family rectangular finite element. □

Another method for forming a higher order is to use just the nodes on the finite element boundary. This method is different from the Lagrange family which was obtained by deduction using the bi-m-th order polynomials and, from the fact that the method was found accidentally, it is called the serendipity group.

Definition 6.4.4 (Serendipity group rectangular finite element) Let $\Xi = (0,1)^2$ be a normal domain. For $m \in \mathbb{N}$, place the nodes $\boldsymbol{\xi}_{(\alpha)}$ with respect to $\alpha \in \mathcal{N}_i$ as shown in Fig. 6.29. Construct the basis functions $\hat{\varphi}_{(\alpha)}$ using polynomials of ξ_1 and ξ_2 and determine the undetermined multipliers so that $\hat{\varphi}_{(\alpha)}\left(\boldsymbol{\xi}_{(\beta)}\right) = \delta_{\alpha\beta}$ is satisfied with respect to $\alpha, \beta \in \mathcal{N}_i$. A finite element using these basis functions created in this way is called a serendipity group rectangular finite element. These polynomials are constructed by terms which are complete second-order polynomials with respect to ξ_1 and ξ_2 with $\xi_1^2\xi_2$ and $\xi_1\xi_2^2$ added on when $m = 2$. Moreover, when $m = 3$, these are constructed by $\xi_1^3\xi_2$ and $\xi_1\xi_2^3$

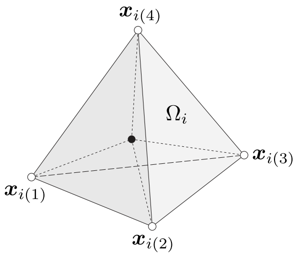


Fig. 6.30: First-order tetrahedral finite element.

being added to complete third-order polynomials with respect to ξ_1 and ξ_2. □

As shown above, if the approximate function $\hat{u}_h(\boldsymbol{\xi})$ is given on the standard coordinate $\boldsymbol{\xi} \in \Xi$, the affine mapping from normal coordinates Ξ to rectangular finite element Ω_i can be used to obtain an element coefficient matrix or known term vector of Ω_i via Eq. (6.4.15) and Eq. (6.4.17). In this case, when the rectangular element is as shown in Fig. 6.26 and $\boldsymbol{x}_{i(\alpha)} = \left(x_{i(\alpha)1}, x_{i(\alpha)2}\right)^\top$ with respect to $\alpha \in \{1,2,3,4\}$ are taken to be the coordinate values of local nodes of the finite element $i \in \mathcal{E}$, the affine mapping is given by

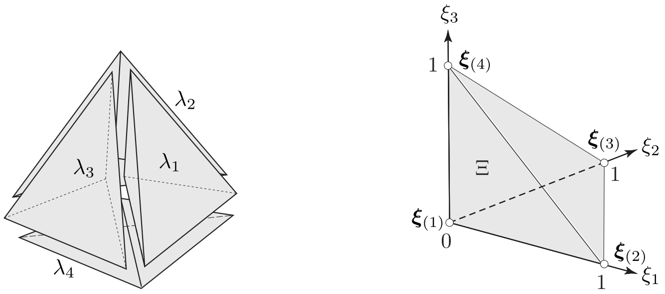
$$\begin{aligned}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= \begin{pmatrix} x_{i(2)1} - x_{i(1)1} & 0 \\ 0 & x_{i(4)2} - x_{i(1)2} \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} x_{i(1)1} \\ x_{i(1)2} \end{pmatrix} \\ &= \boldsymbol{F}_i \boldsymbol{\xi} + \boldsymbol{x}_{i(1)}. \end{aligned} \tag{6.4.25}$$

6.4.4 Tetrahedral finite elements

A three-dimensional finite element can also be constructed in a similar way to the two-dimensional case. First, let us consider a tetrahedron finite element such as the one in Fig. 6.30. Area coordinates were used in triangular finite elements. In a tetrahedral finite element, volume coordinates $(\lambda_1, \ldots, \lambda_4)$ such as that in Fig. 6.31 (a) and a normal domain $\Xi = \left\{ \boldsymbol{\xi} \in (0,1)^3 \;\middle|\; \xi_1 + \xi_2 + \xi_3 < 1 \right\}$ such as the one in Fig. 6.31 (b) are used. In first-order tetrahedral finite elements, $\hat{\varphi}_{(1)} = \lambda_1$, ..., $\hat{\varphi}_{(1)} = \lambda_4$ with respect to the nodes $\boldsymbol{\xi}_{(1)}$, ..., $\boldsymbol{\xi}_{(4)}$ such as the one in Fig. 6.31 (b) are chosen to be basis functions. The m-th order tetrahedron finite element has basis functions using complete m-th order polynomials in a similar way to triangular finite elements.

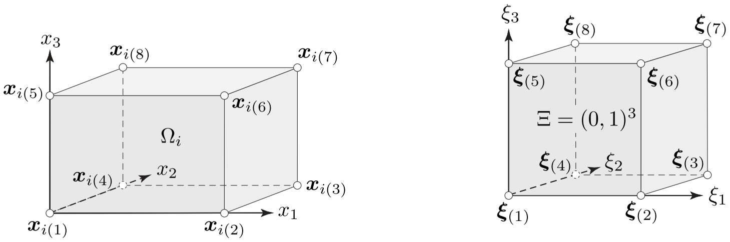
6.4.5 Hexahedral Finite Elements

A hexahedral finite element is also constructed by expanding the rectangular finite element into a three-dimensional space. Figure 6.32 shows nodes of a first-order hexahedral finite element. Here, it is also possible, in a similar way to Eq. (6.4.20) to Eq. (6.4.23), to define the length coordinates $\lambda_{11}, \ldots, \lambda_{33}$ and normal coordinates $\boldsymbol{\xi} = (\xi_1, \xi_2, \xi_3)^\top = (\lambda_{12}, \lambda_{22}, \lambda_{32})^\top \in \Xi$ with respect



(a) Volume coordinates $(\lambda_1, \ldots, \lambda_4)$. (b) Normal coordinates (ξ_1, ξ_2, ξ_3).

Fig. 6.31: Volume coordinates and normal coordinates for a tetrahedral finite element.



(a) Hexahedral first-order finite element. (b) Normal coordinates (ξ_1, ξ_2, ξ_3).

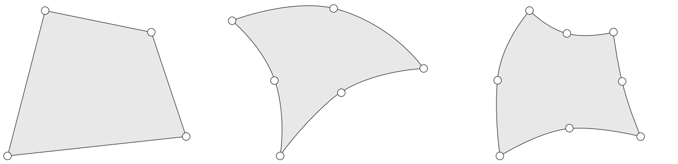
Fig. 6.32: First-order hexahedral finite element and normal coordinates.

to the node coordinates $\boldsymbol{x}_{i(1)}, \ldots, \boldsymbol{x}_{i(8)} \in \mathbb{R}^3$. With respect to a normal element, the Lagrange family m-th order hexahedral finite element has nodes uniformly placed on the normal domain and basis functions $\hat{\varphi}_{(1)}, \ldots, \hat{\varphi}_{((m+1)^3)}$ are constructed using tri-m-th order polynomials with respect to normal coordinates. In serendipity group m-th order hexahedral finite elements, nodes are placed uniformly on the finite element boundary and basis functions constructed using m-th order polynomials.

6.5 Isoparametric Finite Elements

The domain of a finite element seen in Sect. 6.4 is assumed to be triangular or rectangular in two dimensions and tetrahedral or hexadral in three dimensions. Here, let us think about finite elements in the shape of a quadrangle, triangle or quadrangle formed of second-order curves such as shown in Fig. 6.33 and those extended to the three dimensions.

As seen in Sect. 6.4.2, when the basis functions of a triangular finite element are given by area coordinates and the approximate functions constructed from these are substituted into the weak form, the domain integrals are calculated using Theorem 6.3.1. In the case of rectangular finite elements, they can be calculated using Gaussian quadrature, as shown in Sect. 6.5.2. These formulae



(a) Quadrangle (b) 2nd-order curve triangular (c) 2nd-order curve quadrangle

Fig. 6.33: Examples of isoparametric finite elements.

of integration are also effective when the order number of the integrand is increased by making the basis functions a higher order. However, if the integral domain is as shown in Fig. 6.33, such integral formulae can no longer be used.

Hence, if a function (mapping) is used for changing the finite element domain Ω_i into a triangular or rectangular normal domain Ξ if two-dimensional and into a tetrahedral or hexahedral normal coordinates Ξ if three-dimensional, such a change makes integration possible using Theorem 6.3.1 or Gaussian quadrature on Ξ. When the approximate function $\hat{\boldsymbol{x}}_h$ with respect to the mapping $\hat{\boldsymbol{x}} : \Xi \to \Omega_i$ is constructed of the same basis functions as the approximate function of u, the finite element is called an isoparametric finite element. In other words, it is defined in the following way.

Definition 6.5.1 (Isoparametric finite element) Let the normal domain with respect to the domain $\Omega_i \subset \mathbb{R}^d$ of finite element $i \in \mathcal{E}$ be $\Xi \subset \mathbb{R}^d$. With respect to local node number $i \in \mathcal{N}_i = \{1, \ldots, |\mathcal{N}_i|\}$, the basis function of the normal element is $\hat{\boldsymbol{\varphi}} = \left(\hat{\varphi}_{(1)}, \ldots, \hat{\varphi}_{(|\mathcal{N}_i|)}\right)$. In this case, the finite element when the approximate functions of u and v and coordinate values on Ω_i are constructed using

$$\begin{aligned}
\hat{u}_h\left(\boldsymbol{\xi}\right) &= \hat{\boldsymbol{\varphi}}\left(\boldsymbol{\xi}\right) \cdot \bar{\boldsymbol{u}}_i, \\
\hat{v}_h\left(\boldsymbol{\xi}\right) &= \hat{\boldsymbol{\varphi}}\left(\boldsymbol{\xi}\right) \cdot \bar{\boldsymbol{v}}_i, \\
\hat{x}_{h1}\left(\boldsymbol{\xi}\right) &= \hat{\boldsymbol{\varphi}}\left(\boldsymbol{\xi}\right) \cdot \bar{\boldsymbol{x}}_{i1}, \\
&\vdots \\
\hat{x}_{hd}\left(\boldsymbol{\xi}\right) &= \hat{\boldsymbol{\varphi}}\left(\boldsymbol{\xi}\right) \cdot \bar{\boldsymbol{x}}_{id}
\end{aligned}$$

with respect to $\boldsymbol{\xi} \in \Xi$ is called an isoparametric finite element. Here, $\bar{\boldsymbol{u}}_i$ and $\bar{\boldsymbol{v}}_i \in \mathbb{R}^{|\mathcal{N}_i|}$ are taken to be local node value vectors of u and v, and $\bar{\boldsymbol{x}}_{i1}, \ldots, \bar{\boldsymbol{x}}_{id} \in \mathbb{R}^{|\mathcal{N}_i|}$ are taken to be local node coordinate value vectors on Ω_i. □

In an isoparametric finite element, all functions appearing in the weak form with respect to a finite element are given by the normal coordinates $\boldsymbol{\xi} \in \Xi$ as parameters. As a result, the integral domain can be changed to a normal domain and the formula of integration can be used. However, in contrast, the calculation of the partial differential of u with respect to $\boldsymbol{x} \in \Omega_i$ and the Jacobian may be difficult. Let us look at this in the next section.

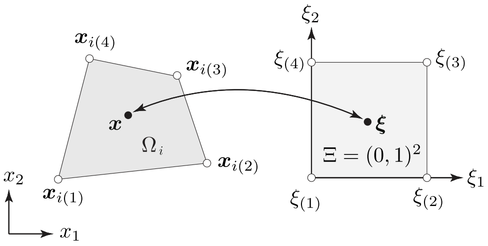


Fig. 6.34: Coordinate transformation of four-node isoparametric finite element.

6.5.1 Two-Dimensional Four-Node Isoparametric Finite Elements

As an example of an isoparametric finite element, let us think about a four-node isoparametric finite element such as the one in Fig. 6.34. With respect to $i \in \mathcal{E}$, assume Ω_i to be a quadrangle domain and $\Xi = (0,1)^2$ a normal domain. Here, with respect to $\boldsymbol{\xi} \in \Xi$,

$$\hat{u}_h(\boldsymbol{\xi}) = \begin{pmatrix} \hat{\varphi}_{(1)}(\boldsymbol{\xi}) & \hat{\varphi}_{(2)}(\boldsymbol{\xi}) & \hat{\varphi}_{(3)}(\boldsymbol{\xi}) & \hat{\varphi}_{(4)}(\boldsymbol{\xi}) \end{pmatrix} \begin{pmatrix} u_{i(1)} \\ u_{i(2)} \\ u_{i(3)} \\ u_{i(4)} \end{pmatrix} = \hat{\boldsymbol{\varphi}}(\boldsymbol{\xi}) \cdot \bar{\boldsymbol{u}}_i,$$

$$\hat{v}_h(\boldsymbol{\xi}) = \hat{\boldsymbol{\varphi}}(\boldsymbol{\xi}) \cdot \bar{\boldsymbol{v}}_i,$$
$$\hat{x}_{h1}(\boldsymbol{\xi}) = \hat{\boldsymbol{\varphi}}(\boldsymbol{\xi}) \cdot \bar{\boldsymbol{x}}_{i1},$$
$$\hat{x}_{h2}(\boldsymbol{\xi}) = \hat{\boldsymbol{\varphi}}(\boldsymbol{\xi}) \cdot \bar{\boldsymbol{x}}_{i2}$$

is defined, where

$$\hat{\boldsymbol{\varphi}} = \begin{pmatrix} \hat{\varphi}_{(1)} \\ \hat{\varphi}_{(2)} \\ \hat{\varphi}_{(3)} \\ \hat{\varphi}_{(4)} \end{pmatrix} = \begin{pmatrix} (1-\xi_1)(1-\xi_2) \\ \xi_1(1-\xi_2) \\ \xi_1\xi_2 \\ (1-\xi_1)\xi_2 \end{pmatrix}.$$

Here, let us think about the calculation of partial differentials with respect to x_1 and x_2 of $\hat{\boldsymbol{\varphi}}$. However, $\hat{\boldsymbol{\varphi}}$ is a function of $\boldsymbol{\xi}$. Here, if the chain rule of differentiation is used, with respect to $\alpha \in \{1,2,3,4\}$,

$$\begin{aligned} \boldsymbol{\nabla}_\xi \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi}) &= \begin{pmatrix} \partial\hat{\varphi}_{(\alpha)}/\partial\xi_1 \\ \partial\hat{\varphi}_{(\alpha)}/\partial\xi_2 \end{pmatrix} = \begin{pmatrix} \partial\hat{x}_1/\partial\xi_1 & \partial\hat{x}_2/\partial\xi_1 \\ \partial\hat{x}_1/\partial\xi_2 & \partial\hat{x}_2/\partial\xi_2 \end{pmatrix} \begin{pmatrix} \partial\hat{\varphi}_{(\alpha)}/\partial x_1 \\ \partial\hat{\varphi}_{(\alpha)}/\partial x_2 \end{pmatrix} \\ &= \left(\boldsymbol{\nabla}_\xi \hat{\boldsymbol{x}}^\top \right) \boldsymbol{\nabla}_x \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi}) \end{aligned}$$

is established. Then,

$$\boldsymbol{\nabla}_x \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi}) = \begin{pmatrix} \partial\hat{\varphi}_{(\alpha)}/\partial x_1 \\ \partial\hat{\varphi}_{(\alpha)}/\partial x_2 \end{pmatrix}$$

$$= \frac{1}{\omega_i(\boldsymbol{\xi})} \begin{pmatrix} \partial \hat{x}_2/\partial \xi_2 & -\partial \hat{x}_2/\partial \xi_1 \\ -\partial \hat{x}_1/\partial \xi_2 & \partial \hat{x}_1/\partial \xi_1 \end{pmatrix} \begin{pmatrix} \partial \hat{\varphi}_{(\alpha)}/\partial \xi_1 \\ \partial \hat{\varphi}_{(\alpha)}/\partial \xi_2 \end{pmatrix}$$

$$= \left(\boldsymbol{\nabla}_\xi \hat{\boldsymbol{x}}^\top \right)^{-1} \boldsymbol{\nabla}_\xi \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi}) \tag{6.5.1}$$

is established. Here, $\left(\boldsymbol{\nabla}_\xi \hat{\boldsymbol{x}}^\top \right)^\top$ and

$$\omega_i(\boldsymbol{\xi}) = \det \left(\boldsymbol{\nabla}_\xi \hat{\boldsymbol{x}}^\top \right) \tag{6.5.2}$$

are the Jacobi matrix and the Jacobian, respectively, of the mapping $\hat{\boldsymbol{x}} : \Xi \to \Omega_i$.

Using these results, the element coefficient matrix $\left(a_i \left(\varphi_{i(\alpha)}, \varphi_{i(\beta)} \right) \right)_{\alpha,\beta} \in \mathbb{R}^{4\times 4}$ can be changed in the following way:

$$\begin{aligned} a_i\left(\varphi_{i(\alpha)}, \varphi_{i(\beta)}\right) &= \int_{\Omega_i} \left(\frac{\partial \varphi_{i(\alpha)}}{\partial x_1} \frac{\partial \varphi_{i(\beta)}}{\partial x_1} + \frac{\partial \varphi_{i(\alpha)}}{\partial x_2} \frac{\partial \varphi_{i(\beta)}}{\partial x_2} \right) \, \mathrm{d}x \\ &= \int_{\Omega_i} \boldsymbol{\nabla}_x \varphi_{i(\alpha)}(\boldsymbol{x}) \cdot \boldsymbol{\nabla}_x \varphi_{i(\beta)}(\boldsymbol{x}) \, \mathrm{d}x \\ &= \int_{(0,1)^2} \boldsymbol{\nabla}_x \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi}) \cdot \boldsymbol{\nabla}_x \hat{\varphi}_{(\beta)}(\boldsymbol{\xi}) \, \omega_i(\boldsymbol{\xi}) \, \mathrm{d}\xi. \end{aligned} \tag{6.5.3}$$

In the integrand on the right-hand side of Eq. (6.5.3), $\boldsymbol{\nabla}_x \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi})$ and $\omega_i(\boldsymbol{\xi})$ can be calculated by Eq. (6.5.1) and Eq. (6.5.2), respectively. The integral can be calculated with the Gaussian quadrature shown next.

6.5.2 Gaussian Quadrature

Let us show the formula of Gaussian quadrature specifically. When $f_n : (-1,1) \to \mathbb{R}$ is n-th order function with respect to $n \in \mathbb{N}$, using the fact that

$$\begin{aligned} &\int_{-1}^{1} f_1(y) \, \mathrm{d}y = 2 f_1(0), \\ &\int_{-1}^{1} f_3(y) \, \mathrm{d}y = f_3\left(-\frac{1}{\sqrt{3}}\right) + f_3\left(\frac{1}{\sqrt{3}}\right), \\ &\int_{-1}^{1} f_5(y) \, \mathrm{d}y = \frac{5}{9} f_5\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f_5(0) + \frac{5}{9} f_5\left(\sqrt{\frac{3}{5}}\right), \\ &\qquad \vdots \end{aligned}$$

holds for $n \in \{1, 3, 5, \ldots\}$, the method for calculating the integration on the left-hand side using the right-hand side is called the Gaussian quadrature. Here, when the term on the right-hand side is written with respect to $i \in \{1, 2, \ldots, (n+1)/2\}$ as $w_i f_n(\eta_i)$, η_i is called the Gaussian node. Figure 6.35 shows the relationship between f_n and Gaussian nodes. Let us see the basis on which these formulae hold.

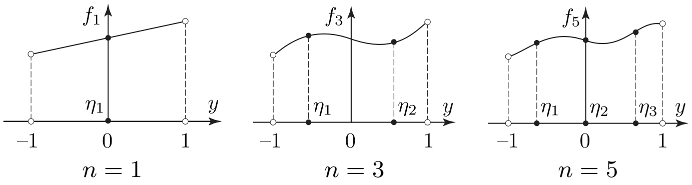


Fig. 6.35: Gaussian quadrature of a one-dimensional function.

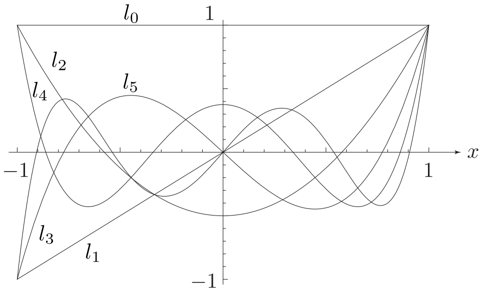


Fig. 6.36: Legendre polynomial l_n.

First, in order to use the Gaussian quadrature theorem to be shown later, let us define Legendre polynomials. Here, n and m are non-negative integers.

Definition 6.5.2 (Legendre polynomials) When the function $l_n : (-1, 1) \to \mathbb{R}$ satisfies the Legendre differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}x}\left\{\left(1 - x^2\right)\frac{\mathrm{d}}{\mathrm{d}x}l_n\right\} + n\left(n+1\right)l_n = 0, \tag{6.5.4}$$

l_n is called a Legendre polynomial. □

Using Rodrigues' formula, the Legendre polynomial can be written as

$$l_n\left(x\right) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\left\{\left(x^2 - 1\right)^n\right\}. \tag{6.5.5}$$

From this, it can be obtained specifically as

$$l_0\left(x\right) = 1, \quad l_1\left(x\right) = x, \quad l_2\left(x\right) = x^2 - \frac{1}{3}, \quad l_3\left(x\right) = x^3 - \frac{3}{5}x, \quad \cdots .$$

Figure 6.36 shows l_0 to l_5. Therefore, it becomes apparent that l_n is an n-th order polynomial. Furthermore, if the function $f : (-1, 1) \to \mathbb{R}$ is a polynomial of less than n-th order then

$$\int_{-1}^{1} f\left(x\right) l_n\left(x\right) \mathrm{d}x = 0$$

holds. This is because

$$\begin{aligned}\int_{-1}^{1} f(x)\, l_n(x)\, \mathrm{d}x &= \left[f \left\{ -\frac{1}{n(n+1)} \left(1-x^2\right) \frac{\mathrm{d}l_n}{\mathrm{d}x} \right\} \right]_{-1}^{1} \\ &\quad - \frac{1}{2^n n!} \int_{-1}^{1} \frac{\mathrm{d}f}{\mathrm{d}x} \frac{\mathrm{d}^{n-1}}{\mathrm{d}x^{n-1}} \left\{ \left(x^2-1\right)^n \right\} \mathrm{d}x \\ &= \frac{(-1)^n}{2^n n!} \int_{-1}^{1} \frac{\mathrm{d}^n f}{\mathrm{d}x^n} \left(x^2-1\right)^n \mathrm{d}x = 0\end{aligned}$$

holds from Eq. (6.5.4) and Eq. (6.5.5). Using these properties, the orthogonality of $\{l_n\}_n$:

$$\int_{-1}^{1} l_n(x)\, l_m(x)\, \mathrm{d}x = \frac{2}{2n+1} \delta_{nm}$$

is established.

Moreover, in general, with respect to $x_0 < x_1 < \cdots < x_n$,

$$\begin{aligned}\phi_i(x) &= \prod_{j \in \{1,\ldots,n\},\ j \neq i} \frac{x - x_j}{x_i - x_j} \\ &= \frac{(x-x_0)(x-x_1)\cdots(x-x_{i-1})(x-x_{i+1})\cdots(x-x_n)}{(x_i-x_0)(x_i-x_1)\cdots(x_i-x_{i-1})(x_i-x_{i+1})\cdots(x_i-x_n)}\end{aligned}$$

satisfies $\phi_i(x_j) = \delta_{ij}$, and with respect to some $f : \mathbb{R} \to \mathbb{R}$, if we set

$$\hat{f}(x) = \sum_{i \in \{1,\ldots,n\}} \phi_i(x)\, f(x_i),$$

then the equation $\hat{f}(x_i) = f(x_i)$ is satisfied. Here, $\phi_i(x)$ is called the Lagrange basis polynomials and $\hat{f}(x)$ is Lagrange interpolation. At this time, we write the Lagrange basis polynomials with respect to the roots $\eta_1, \ldots, \eta_n$ of Legendre polynomial l_n as

$$\varphi_i(x) = \prod_{j \in \{1,\ldots,n\},\ j \neq i} \frac{x - \eta_j}{\eta_i - \eta_j}. \tag{6.5.6}$$

Figure 6.37 shows φ_1 to φ_5. If we consider Lagrange interpolation using $\varphi_i(x)$, a formula of Gaussian quadrature can be obtained as follows.

Theorem 6.5.3 (Gaussian quadrature) Let $\eta_1, \ldots, \eta_n$ be the roots of an n-th order Legendre polynomial l_n. Let $f : (-1, 1) \to \mathbb{R}$ be a polynomial of less than $2n$-th order. In this case,

$$\int_{-1}^{1} f(x)\, \mathrm{d}x = \sum_{i \in \{1,\ldots,n\}} w_i f(\eta_i)$$

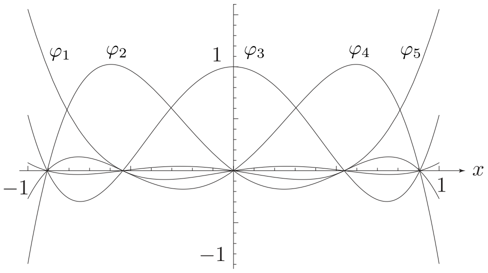


Fig. 6.37: Functions $\varphi_i(x)$ with the roots of a fifth-order Legendre polynomial as nodes.

holds. Here,

$$w_i = \int_{-1}^{1} \varphi_i(x)\, \mathrm{d}x$$

with respect to $\varphi_i(x)$ of Eq. (6.5.6). □

Proof Let us suppose that $f(x)$ is a polynomial of $(n-1)$-th order. In this case,

$$f(x) = \sum_{i\in\{1,\ldots,n\}} \varphi_i(x) f(\eta_i)$$

holds. This is because although the difference between both sides includes an n-th order differential but f's n-th order differential is 0. Hence, the following holds:

$$\int_{-1}^{1} f(x)\, \mathrm{d}x = \int_{-1}^{1} \left(\sum_{i\in\{1,\ldots,n\}} \varphi_i(x) f(\eta_i) \right) \mathrm{d}x = \sum_{i\in\{1,\ldots,n\}} w_i f(\eta_i).$$

Next, let us suppose f is a polynomial of order greater than n but less than $2n$. In this case we can write

$$f(x) = l_n(x)\, g(x) + r(x).$$

However, $g(x)$ and $r(x)$ are polynomials of less than n-th order. Here, the qualities of the Legendre polynomial give

$$\int_{-1}^{1} l_n(x)\, g(x)\, \mathrm{d}x = 0.$$

Moreover, from $l_n(\eta_i) = 0$,

$$f(\eta_i) = r(\eta_i)$$

holds. Furthermore, since $r(x)$ is a polynomial of less than n-th order,

$$\int_{-1}^{1} r(x)\, \mathrm{d}x = \sum_{i\in\{1,\ldots,n\}} w_i r(\eta_i)$$

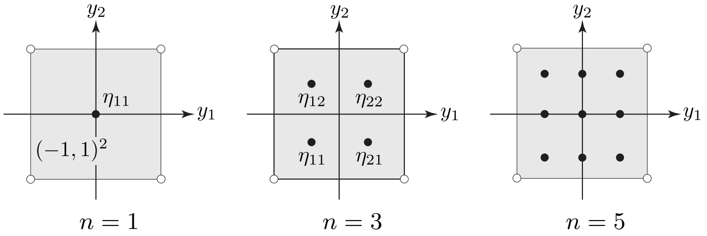


Fig. 6.38: Gaussian quadrature for functions defined on a two-dimensional domain.

holds. Therefore we get

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \int_{-1}^{1} r(x)\,\mathrm{d}x = \sum_{i\in\{1,\ldots,n\}} w_i r(\eta_i) = \sum_{i\in\{1,\ldots,n\}} w_i f(\eta_i)\,.$$

□

In Theorem 6.5.3, when the integral domain is changed to $(0,1)$, the integral equation can be changed by a change in variables to

$$\int_{0}^{1} f_{2n-1}(y)\,\mathrm{d}y = \frac{1}{2}\sum_{i\in\{1,\ldots,n\}} w_i f\left(\frac{\eta_i-1}{2}\right).$$

Furthermore, with respect to a bi-n-th order function on two-dimensional domain $(-1,1)^2$,

$$\begin{aligned}\int_{(-1,1)^2} f_{2n-1}(\boldsymbol{\xi})\,\mathrm{d}\xi &= \sum_{(i,j)\in\{1,\ldots,n\}^2} w_i w_j f(\boldsymbol{\eta}_{ij})\,,\\ \int_{(0,1)^2} f_{2n-1}(\boldsymbol{\xi})\,\mathrm{d}\xi &= \frac{1}{4}\sum_{(i,j)\in\{1,\ldots,n\}^2} w_i w_j f\left(\left(\boldsymbol{\eta}_{ij}-\begin{pmatrix}1\\1\end{pmatrix}\right)\Big/2\right)\end{aligned} \tag{6.5.7}$$

holds with respect to Gaussian nodes such as in Fig. 6.38. These formulae are used in numerical integration of rectangular isoparametric finite elements. For example, with respect to Eq. (6.5.3), Eq. (6.5.7) is used. Here, the value of n is not chosen exactly due to the fact that the inverse calculation of a matrix with polynomials as elements is included. Hence a value as small as possible such that no practical issues arise can be looked for with a numerical experiment. Moreover, with respect to linear elastic problems, it is known that the use of selected reduced integration suppresses the generation of hourglass modes (deformation such that strain becomes 0). We refer the readers to literature focusing on these topics.

6.6 Error Estimation

In Theorem 6.1.13, it was seen that the approximate solution from the Galerkin method was the best element in the set U_h of approximate functions. Hence, the error in the approximate solution from the Galerkin method depends on how close an element of U_h gets to an element in the function space containing the exact solution (approximation ability). Here, the results of Theorem 6.1.13 will be used to think about the error evaluation of an approximate solution (finite element solution) obtained from the finite element method. Results shown here will be used for error evaluation in numerical solutions for shape optimization problems in Chaps. 8 and 9. Here, based on what we have seen so far in this chapter, let us give an abstraction of the finite element method to some extent to look at basic theorems.

6.6.1 Finite Element Division Sequence

Let us define the finite element division. Let $\Omega \subset \mathbb{R}^d$ be a $d \in \{1,2,3\}$-dimensional bounded domain of a polygon in two dimensions and a polyhedron in three dimensions in order to be able to ignore the error due to domain division, and call it polyhedron generally. With respect to Ω, $\mathcal{T} = \{\Omega_i\}_{i\in\mathcal{E}}$ is called a finite element division. Here, $\mathcal{E}$ is the finite set of element numbers. Moreover, Ω_i are convex polyhedrons such that $\bar{\Omega} = \bigcup_{i\in\mathcal{E}} \bar{\Omega}_i$ and

$$\Omega_i \cap \Omega_j = \emptyset, \quad \bar{\Omega}_i \cap \bar{\Omega}_j \subset \mathbb{R}^{d-1}$$

for $i \neq j$ with respect to an arbitrary $i, j \in \mathcal{E}$ is satisfied. With respect to each Ω_i, we call

$$\operatorname{diam}\Omega_i = \sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega_i} \|\boldsymbol{x}-\boldsymbol{y}\|_{\mathbb{R}^d}$$

the diameter of Ω_i and write it as h_i. Moreover, write the diameter of the inscribed sphere in Ω_i as $\operatorname{inscr}\Omega_i$. When some positive real number σ exists and

$$\frac{\operatorname{inscr}\Omega_i}{\operatorname{diam}\Omega_i} \geq \sigma \tag{6.6.1}$$

with respect to all $i \in \mathcal{E}$ holds, $\mathcal{T}$ is said to be a regular finite element division. Furthermore,

$$h(\mathcal{T}) = \max_{i\in\mathcal{E}} h_i \tag{6.6.2}$$

is referred to as the maximum diameter of $\mathcal{T}$ and a finite element division sequence such that it becomes $h(\mathcal{T}) \to 0$ will be written as $\{\mathcal{T}_h\}_{h\to 0}$.

6.6.2 Affine-Equivalent Finite Element Division Sequence

Choose some finite element division $\mathcal{T}_h$ from a regular finite element division sequence $\{\mathcal{T}_h\}_{h \to 0}$ and let the set of node numbers in this case be $\mathcal{N}$ and $\boldsymbol{x}_j$ a node with respect to $j \in \mathcal{N}$. Here, the basis functions $\phi_j : \Omega \to \mathbb{R}$ of the finite element method when viewed as a Galerkin method are assumed to satisfy the following conditions:

(1) With respect to an arbitrary $j, l \in \mathcal{N}$,

$$\phi_j\left(\boldsymbol{x}_l\right) = \delta_{jl}$$

holds.

(2) ϕ_j has support on a domain of finite elements with node $j \in \mathcal{N}$.

The set of these basis functions ϕ_j, as seen in Sections 6.2 to 6.5, can be rewritten as a set of basis functions $\varphi_{i(\alpha)}$ $(\alpha \in \mathcal{N}_i)$ defined on the domain Ω_i of finite elements $i \in \mathcal{E}$. Furthermore, we set the normal domain to be Ξ, basis functions on Ξ given by $\hat{\varphi}_{(\alpha)} : \Xi \to \mathbb{R}$, the mapping of normal domain to finite element domain is written as $\boldsymbol{f}_i : \Xi \to \Omega_i$, and

$$\varphi_{i(\alpha)}\left(\boldsymbol{f}_i\left(\boldsymbol{\xi}\right)\right) = \hat{\varphi}\left(\boldsymbol{\xi}_{(\alpha)}\right).$$

In this section, for simplicity, all normal elements corresponding to all finite elements are taken to be common as

$$\Xi = (0,1)^d \quad \text{or} \quad \Xi = \left\{ \boldsymbol{\xi} \in (0,1)^d \;\middle|\; \xi_1 + \cdots + \xi_d < 1 \right\}. \tag{6.6.3}$$

Moreover, $\boldsymbol{f}_i$ is assumed to be given by a linear form such as

$$\boldsymbol{f}_i\left(\boldsymbol{\xi}\right) = \boldsymbol{F}_i \boldsymbol{\xi} + \boldsymbol{b}_i \tag{6.6.4}$$

using $\boldsymbol{F}_i \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b}_i \in \mathbb{R}^d$. In the case of triangular or rectangular finite elements, they are specifically shown in Eq. (6.4.16) and Eq. (6.4.25). A mapping such as this, combining a linear mapping and a translation with a fixed element, is called an affine mapping. Here, the finite element division when $\boldsymbol{f}_i$ is an affine mapping is called an affine-equivalent finite element division.

The finite element division seen from Sections 6.2 to 6.4 had assumed an affine-equivalent finite element division sequence. With isoparametric finite elements, this relationship does not generally hold, but even if $\boldsymbol{f}_i : \Xi \to \Omega_i$ is a mapping combining a non-linear mapping and a translation with a fixed element, if the Jacobi matrix $\boldsymbol{f}_{i\boldsymbol{\xi}^\top} = \left(\boldsymbol{\nabla}_\xi \boldsymbol{f}_i^\top\right)^\top$ and Jacobian $\omega_i\left(\boldsymbol{\xi}\right) = \det \boldsymbol{f}_{i\boldsymbol{\xi}^\top}$ of $\boldsymbol{f}_i$ shown later have upper limits and lower limits given by positive real numbers, then a similar argument holds.

When using a regular affine-equivalent finite element division sequence defined in this way, in a Poisson problem for example, there is the need to calculate

$$a_i\left(\varphi_{i(\alpha)}, \varphi_{i(\beta)}\right)$$

$$\begin{aligned}
&= \int_{\Omega_i} \boldsymbol{\nabla}_x \varphi_{i(\alpha)} \cdot \boldsymbol{\nabla}_x \varphi_{i(\beta)} \, \mathrm{d}x \\
&= \int_{\Xi} \left\{ \left(\boldsymbol{\nabla}_\xi \boldsymbol{f}_i^\top \right)^{-1} \boldsymbol{\nabla}_\xi \hat{\varphi}_{(\alpha)} \right\} \cdot \left\{ \left(\boldsymbol{\nabla}_\xi \boldsymbol{f}_i^\top \right)^{-1} \boldsymbol{\nabla}_\xi \hat{\varphi}_{(\beta)} \right\} \omega_i \, \mathrm{d}\xi \\
&= \int_{\Xi} \left\{ \left(\boldsymbol{F}_i^\top \right)^{-1} \boldsymbol{\nabla}_\xi \hat{\varphi}_{(\alpha)} \right\} \cdot \left\{ \left(\boldsymbol{F}_i^\top \right)^{-1} \boldsymbol{\nabla}_\xi \hat{\varphi}_{(\beta)} \right\} \omega_i \, \mathrm{d}\xi \qquad (6.6.5)
\end{aligned}$$

with respect to an arbitrary $\alpha, \beta \in \mathcal{N}_i$, where $\boldsymbol{F}_i$ is a matrix used in Eq. (6.6.4) and $\omega_i = \det \boldsymbol{F}_i$. The following results can be obtained with respect to $\boldsymbol{F}_i$ and ω_i of Eq. (6.6.5).

Lemma 6.6.1 (Jacobian of an affine mapping) Let $\mathcal{T}_h = \{\Omega_i\}_{i \in \mathcal{E}}$ be a regular affine-equivalent finite element division sequence with normal domain Ξ as Eq. (6.6.3) and affine mapping $\boldsymbol{f}_i$ as Eq. (6.6.4). Let $\operatorname{diam} \Omega_i = h_i$. In this case,

$$\omega_i = \det \boldsymbol{F}_i \le h_i^d, \quad \omega_i^{-1} = \det (\boldsymbol{F}_i)^{-1} \le c_1 h_i^{-d} \qquad (6.6.6)$$

holds. Moreover, with respect to $\boldsymbol{F}_i = (a_{ijk})_{jk} \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$ and $\boldsymbol{F}_i^{-1} = \left(a_{ijk}^{-1} \right)_{jk} \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$,

$$|a_{ijk}| \le h_i, \quad \left| a_{ijk}^{-1} \right| \le c_2 h_i^{-1} \qquad (6.6.7)$$

is established. Here, c_1 and c_2 is positive constants depending on σ in Eq. (6.6.1) and d. □

Proof If it is a finite element of an affine-equivalent finite element division sequence, ω_i is a constant. When Ξ is Eq. (6.6.3), $0 < |\Xi| \le 1$. $\mathcal{T}_h$ is regular, hence

$$\frac{1}{c_1} h_i^d \le \omega_i = \frac{\int_\Xi \omega_i \, \mathrm{d}\xi}{\int_\Xi \mathrm{d}\xi} = \frac{|\Omega_i|}{|\Xi|} \le h_i^d.$$

In other words, Eq. (6.6.6) holds. Moreover, if $\boldsymbol{f}_i(\boldsymbol{\xi})$ of Eq. (6.6.4) is written as $(f_{ij}(\boldsymbol{\xi}))_j$,

$$|a_{ijk}| = \left| \frac{\partial f_{ij}}{\partial \xi_k} \right| \le h_i$$

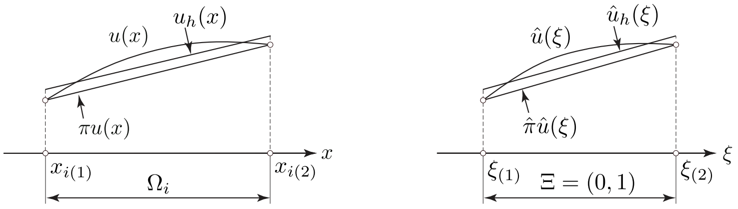
is obtained from $\operatorname{diam} \Omega_i = h_i$. Furthermore, from

$$\omega_i^{-1} = \det \left(\frac{\partial f_{ij}^{-1}}{\partial x_k} \right)_{jk} \le \frac{c_1}{h_i^d},$$

the following is obtained:

$$|a_{ijk}^{-1}| = \left| \frac{\partial f_{ij}^{-1}}{\partial x_k} \right| \le \frac{c_2}{h_i}.$$

In other words, Eq. (6.6.7) is established. □



(a) Interpolation function πu on Ω_i. (b) Interpolation function $\hat{\pi}\hat{u}$ on Ξ.

Fig. 6.39: Interpolation function and finite element solution.

6.6.3 Interpolation Error Estimation

In Sect. 6.6.2, attention was given to the relationship between the normal element and the finite element. Here, let us focus on the approximation ability of the approximate function. In Theorem 6.1.13, the finite element solution error $\|u - u_h\|_U$ measured with a Hilbert space containing the exact solution was seen to be limited to $\inf_{v_h \in U_h} \|u - v_h\|_U$ at the homogeneous basic boundary condition ($u_{\mathrm{D}} = 0$). Here, in order to evaluate its approximation ability, we shall think about creating an approximate function for which it is easy to evaluate error, but the error may be greater than for the finite element solution. If an error of such an approximate function is evaluated, from the fact that the error of the finite element solution is smaller (Theorem 6.1.13), the error of the finite element solution should be evaluated using that error. This is shown in Sect. 6.6.4.

In preparation, in this section, let us think about approximate functions which can be easily evaluated for error. Such an approximate function is assumed to be a function which is an element of U_h and agrees with the exact solution at nodes. Such an approximate function is called an interpolation function. Figure 6.39 (a) shows the relationship between the exact solution u, the finite element solution u_h and the interpolation function πu when basis functions are given by linear functions with respect to a one-dimensional problem. Figure 6.39 (b) shows the functions defined on the normal elements. Here, π and $\hat{\pi}$ are called interpolation operators and are defined as below. Write the function space of exact solutions on Ω_i as $U(\Omega_i)$. Moreover, the function space (linear space) of an interpolation function set by the basis functions $\varphi_{i(\alpha)}$ ($\alpha \in \mathcal{N}_i$) is written as $W(\Omega_i) = \operatorname{span}\left(\varphi_{i(\alpha)}\right)_{\alpha \in \mathcal{N}_i}$ (Definition 4.2.6). On the other hand, the function spaces of the exact solution and interpolation functions defined on Ξ are respectively written as $U(\Xi)$ and $W(\Xi) = \operatorname{span}\left(\hat{\varphi}_{(\alpha)}\right)_{\alpha \in \mathcal{N}_i}$. Here, the operators $\pi : U(\Omega_i) \to W(\Omega_i)$ and $\hat{\pi} : U(\Xi) \to W(\Xi)$ are defined by

$$\pi u(\boldsymbol{x}) = \sum_{\alpha \in \mathcal{N}_i} u\left(\boldsymbol{x}_{i(\alpha)}\right) \varphi_{i(\alpha)}(\boldsymbol{x}), \tag{6.6.8}$$

$$\hat{\pi}\hat{u}(\boldsymbol{\xi}) = \sum_{\alpha \in \mathcal{N}_i} \hat{u}\left(\boldsymbol{\xi}_{(\alpha)}\right) \hat{\varphi}_{(\alpha)}(\boldsymbol{\xi}). \tag{6.6.9}$$

Such an error $\|u - \pi u\|_U$, that is $\|u - \pi u\|_{H^1(\Omega;\mathbb{R})}$, in an interpolation function is called an interpolation error. In order to evaluate this, first let us define the set of all k-th order polynomials and look at results relating to their general approximation abilities. For $k \in \{0, 1, \ldots\}$, write the set of all k-th order polynomials (complete k-th order polynomials) defined on a bounded domain $\Omega \subset \mathbb{R}^d$ as

$$\mathcal{P}_k(\Omega) = \left\{ \sum_{|\boldsymbol{\beta}| \le k} c_{\boldsymbol{\beta}} x_1^{\beta_1} \cdots x_d^{\beta_d} \;\middle|\; c_{\boldsymbol{\beta}} \in \mathbb{R},\ \boldsymbol{\beta} \in \{0, 1, \ldots, k\}^d,\ \boldsymbol{x} \in \Omega \right\},$$

where $\boldsymbol{\beta}$ is multi-indexed. In this case, the following results can be obtained (cf. [1, Theorem14.1, p. 120], [6, Lemma 2.8, p. 60]).

Theorem 6.6.2 (Approximation ability of the k-th order polynomials)

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with piecewise smooth boundary, $p \in [1, \infty]$ and $k \in \{0, 1, \ldots\}$. In this case,

$$\inf_{\phi \in \mathcal{P}_k(\Omega)} \|v - \phi\|_{W^{k+1,p}(\Omega;\mathbb{R})} \le c\, |v|_{W^{k+1,p}(\Omega;\mathbb{R})} \tag{6.6.10}$$

holds with respect to an arbitrary $v \in W^{k+1,p}(\Omega; \mathbb{R})$. Here, c is a positive constant dependent on Ω and k. □

Based on Theorem 6.6.2, the following results can be obtained with respect to the interpolation functions on the domain Ω_i of finite elements $i \in \mathcal{E}$ (cf. [1, Theorem 16.1, p. 126]).

Theorem 6.6.3 (Interpolation error on a finite element) Let $\{\mathcal{T}_h\}_{h \to 0}$ be a regular finite element division sequence with respect to $d \in \{1, 2, 3\}$-dimensional bounded domain $\Omega \subset \mathbb{R}^d$ and $\mathcal{T}_h = \{\Omega_i\}_{i \in \mathcal{E}}$ its element. With respect to $\alpha \in \mathcal{N}_i$, let $\hat{\varphi}_{(\alpha)}$ be basis functions defined on normal coordinates on Ξ and $W(\Xi) = \operatorname{span}\left(\hat{\varphi}_{(\alpha)}\right)_{\alpha \in \mathcal{N}_i}$ the function space of an interpolation function. $p \in [1, \infty]$ and $k, l \in \{0, 1, \ldots\}$ are assumed under the conditions:

$$k + 1 > \frac{d}{p}, \quad k + 1 \ge l \tag{6.6.11}$$

to satisfy

$$\mathcal{P}_k(\Xi) \subset W(\Xi) \subset W^{l,p}(\Xi; \mathbb{R}). \tag{6.6.12}$$

Let π be an interpolation operator of Eq. (6.6.8). In this case, with respect to an arbitrary $v \in W^{k+1,p}(\Omega; \mathbb{R})$, there exists a positive constant c which does not depend on the diameter h_i of Ω_i and

$$|v - \pi v|_{W^{l,p}(\Omega_i;\mathbb{R})} \le c h_i^{k+1-l} |v|_{W^{k+1,p}(\Omega_i;\mathbb{R})}$$

holds. □

Using the results from Theorem 6.6.3, the following result is obtained with respect to the interpolation error on the domain Ω (cf. [1, Theorem16.2, p. 128], [6, Theorem 2.8, p. 62] where $p = 2$ is assumed).

Corollary 6.6.4 (Interpolation error on a domain) Under the assumption of Theorem 6.6.3, let $l \in \{1, 2, \ldots\}$. The basis functions $\boldsymbol{\phi} = (\phi_j)_{j\in\mathcal{N}}$ defined on Ω is taken to be continuous. Let π be an interpolation operator of Eq. (6.6.8). At this point, with respect to an arbitrary $v \in W^{k+1,p}(\Omega;\mathbb{R})$, there exists a positive constant c which does not depend on the maximum diameter h and

$$|v - \pi v|_{W^{l,p}(\Omega;\mathbb{R})} \leq ch^{k+1-l} |v|_{W^{k+1,p}(\Omega;\mathbb{R})}$$

holds. □

Estimating the error using the order with respect to the diameter of a finite element such as in Theorem 6.6.3 or Corollary 6.6.4 is called the order estimation of error.

6.6.4 Error Estimation of Finite Element Solution

The error evaluation of an interpolation function made of basis functions by the finite element method is given by the results when the exact solution in Corollary 6.6.4 is taken to be $v \in W^{k+1,p}(\Omega;\mathbb{R})$. On the other hand, Theorem 6.1.13 shows the results when measuring the basic error $u - u_h$ with $U = H^1(\Omega;\mathbb{R})$ norm of approximate solution by the Galerkin method (finite element solution). Here, the error of the finite element solution can be obtained by using the results when $p = 2$ and $l = 1$ are set in Corollary 6.6.4.

Error evaluation when measuring with $H^1(\Omega;\mathbb{R})$ norm is as follows (cf. [1, Theorem 18.1, p. 138], [6, Theorem 2.9, p. 64]).

Theorem 6.6.5 (Error evaluation of FE solution due to H^1 norm) Let $\{\mathcal{T}_h\}_{h\to 0}$ be a regular finite element division sequence with respect to a $d \in \{1, 2, 3\}$-dimensional polyhedron $\Omega \subset \mathbb{R}^d$ and $\mathcal{T}_h = \{\Omega_i\}_{i\in\mathcal{E}}$ its element. $p = 2$, $l = 1$ and $k \in \{0, 1, \ldots\}$ are to satisfy Eq. (6.6.11) and Eq. (6.6.12). Basis functions $\boldsymbol{\phi} = (\phi_j)_{j\in\mathcal{N}}$ defined on Ω are taken to be continuous. Let the set of approximate solutions in the finite element method be

$$U_h = \left\{ v_h(\bar{\boldsymbol{v}}) = \bar{\boldsymbol{v}} \cdot \boldsymbol{\phi} \;\middle|\; \bar{\boldsymbol{v}} = (\bar{v}_j)_{j\in\mathcal{N}} \in \mathbb{R}^{|\mathcal{N}|} \right\}. \tag{6.6.13}$$

Let $\{u_h\}_{h\to 0}$ ($u_h \in U_h$) be a sequence of finite element solutions with respect to Problem 6.1.6 when the homogeneous fundamental boundary condition ($u_{\mathrm{D}} = 0$). Here, if the exact solution is $u \in U \cap H^{k+1}(\Omega;\mathbb{R})$, there exists a positive constant c which does not depend on h and

$$\|u - u_h\|_{H^1(\Omega;\mathbb{R})} \leq ch^k |u|_{H^{k+1}(\Omega;\mathbb{R})}$$

holds. □

Furthermore, the results when the error $u-u_h$ is measured with an $L^2(\Omega;\mathbb{R})$ norm can be obtained in the following way using a method known as the Aubin–Nitsche trick ([1, Theorem 19.2, p. 142], [6, Theorem 2.11, p. 66], where it is assumed that Ω is a two-dimensional polygonal convex domain).

Theorem 6.6.6 (Error evaluation of FE solution due to L^2 norm) Let $\{\mathcal{T}_h\}_{h\to 0}$ be a regular finite element division sequence with respect to a $d \in \{1,2,3\}$-dimensional polyhedral bounded domain Ω and $\mathcal{T}_h = \{\Omega_i\}_{i\in\mathcal{E}}$ be its element. Suppose Eq. (6.6.12) is satisfied under $p=2$, $l=1$, $d\le 3$ and $k\ge 1$. Basis functions $\boldsymbol{\phi} = (\phi_j)_{j\in\mathcal{N}}$ defined on Ω will be continuous. The set of approximate solutions in the finite element method U_h is Eq. (6.6.13). Let $\{u_h\}_{h\to 0}$ $(u_h \in U_h)$ be a sequence of finite element solutions with respect to Problem 6.1.6 at the homogeneous fundamental boundary condition $(u_{\mathrm{D}}=0)$. Here, if the exact solution is $u \in U \cap H^{k+1}(\Omega;\mathbb{R})$, there exists a positive constant c which does not depend on h and

$$\|u-u_h\|_{L^2(\Omega;\mathbb{R})} \le c h^{k+1} |u|_{H^{k+1}(\Omega;\mathbb{R})}$$

holds. □

Let us conduct an error evaluation of finite element solutions using the above results. Results seen in Section 5.3 can be used for the regularity of exact solutions. First, if the smoothness (regularity) of the exact solution can be determined depending on the smoothness of the known functions, the following result is obtained.

Exercise 6.6.7 (Error evaluation of FE solution) In Problem 6.1.6, the open angle at the boundary between the Dirichlet boundary and Neumann boundary is $\alpha < \pi/2$ and other boundaries are taken to be smooth. Let $b \in L^2(\Omega;\mathbb{R})$ and $p_{\mathrm{N}} = 0$. In this case, show the order evaluation of the error with respect to the finite element solution. □

Answer When $b \in L^2(\Omega;\mathbb{R})$, from $-\Delta u = b$, with respect to the exact solution u, $|u|_{H^2(\Omega;\mathbb{R})} \le c_0 \|b\|_{L^2(\Omega;\mathbb{R})}$ holds. Therefore, when $k=1$ in Theorem 6.6.5 and Theorem 6.6.6,

$$\|u-u_h\|_{H^1(\Omega;\mathbb{R})} \le c_1 h |u|_{H^2(\Omega;\mathbb{R})},$$
$$\|u-u_h\|_{L^2(\Omega;\mathbb{R})} \le c_1 h^2 |u|_{H^2(\Omega;\mathbb{R})}$$

are obtained with respect to the finite element solution. □

Next, let us think about a two-dimensional domain with a non-smooth boundary.

Exercise 6.6.8 (Error evaluation for non-smooth boundary) In Problem 6.1.6, Ω is taken to be a two-dimensional domain with corner point $\boldsymbol{x}_0$ such as in Fig. 5.3.1. Consider the error evaluation of the finite element solution around $\boldsymbol{x}_0$. □

Answer A singularity appears when Γ_1 and Γ_2 have the same boundary with an opening angle of $\alpha > \pi$ (concave angle) and the opening angle at a mixed boundary is $\alpha > \pi/2$. For example, if there is a crack at boundaries of the same type ($\alpha = 2\pi$) or a straight line at a mixed boundary ($\alpha = \pi$), from Eq. (5.3.10),

$$u \in H^{3/2-\epsilon}\left(B\left(\boldsymbol{x}_0, r_0\right) \cap \Omega; \mathbb{R}\right)$$

holds with respect to $\epsilon > 0$ r_0-around the corner point. On the other hand, the order estimation of error with respect to a finite element solution must satisfy $k+1 = 2 > d/p$ with respect to $d \in \{2, 3\}$ and $p = 2$ due to Eq. (6.6.11). In other words, it is not applicable unless the exact solution is in $H^2(\Omega; \mathbb{R})$. Hence, order evaluation of the error is not possible for the finite element solution around the corner points. □

Based on Exercise 6.6.8, order estimation of error could not be made with respect to the finite element solution around singular points. However, the following result can be obtained with respect to convergence to the exact solution due to the fact that $U \cap H^{k+1}(\Omega; \mathbb{R})$ is dense at U (cf. [2, Theorem 5.4, p. 100]).

Theorem 6.6.9 (Convergence of finite element solution using H^1 norm) The same notation as Theorem 6.6.5 is used. Let $\{u_h\}_{h\to 0}$ ($u_h \in U_h$) be a sequence of finite element solutions with respect to Problem 6.1.6 at the homogeneous fundamental boundary condition ($u_{\mathrm{D}} = 0$). Here,

$$\lim_{h \to 0} \|u - u_h\|_{H^1(\Omega;\mathbb{R})} = 0$$

holds with respect to the exact solution $u \in U$. □

Furthermore, with respect to the finite element solution around singularity points, finite elements for expressing the exact solution have been thought of. For example, there are methods developed to add to basis functions which can approximate a series expansion with respect to r around singular points as seen in Section 5.3 (cf. [4, Chap. 8, p. 257], [7]).

6.7 Summary

In Chap. 6, the numerical solutions with respect to boundary value problems in partial differential equations were looked at in terms of the finite element method with the Galerkin method as a leading principle and error evaluation of their numerical solutions. Key points are as follows.

(1) The Galerkin method constructs an approximate function via a linear combination of basis functions multiplied by undetermined multipliers, and by substituting these approximate functions into the weak form, changes a boundary value problem of a partial differential equation to a simultaneous linear equation relating to undetermined multipliers (Sect. 6.1).

(2) The finite element method is a Galerkin method. Here, the finite element method divides the domain into sets of simple shaped domains, and constructs approximate functions using basis functions of low-order polynomials in each domain and continuous at the boundaries of the split domains (Sect. 6.2, Sect. 6.3, Sect. 6.4).

(3) The isoparametric finite element method is a method for evaluating integrals on finite elements in a normal domain by mapping finite element domains to a normal domain. Here, the mappings from finite element domains to the normal domain are taken to use the same basis functions as that used for the approximate functions with respect to the solution. For the numerical integration of a rectangle on a normal domain, Gaussian quadrature is used (Sect. 6.5).

(4) The error norms of approximate solutions from the finite element method can be suppressed with powers of the sizes of the finite elements (Sect. 6.6).

6.8 Practice Problems

6.1 Let $u : (0,1) \to \mathbb{R}$ be the solution of the first-type boundary value problem of a one-dimensional second-order differential equation:

$$-\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} + u = 1 \quad \text{in } (0,1), \quad u(0) = u(1) = 0.$$

Obtain the approximate solution u_h using the Galerkin method. Here, use the same basis functions as Exercise 6.1.5.

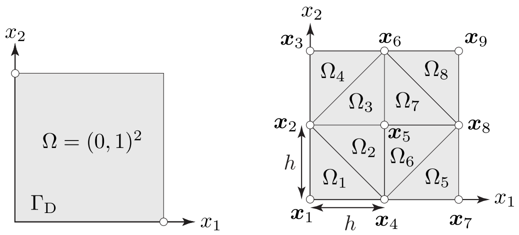
6.2 Obtain the simultaneous linear equations when solving the boundary value problem in Practice **6.1** by the finite element method using the first-order basis functions. Here, let the finite element number be $m = 4$.

6.3 When the three nodes $\boldsymbol{x}_{i(1)}$, $\boldsymbol{x}_{i(2)}$ and $\boldsymbol{x}_{i(3)}$ of the triangular finite element $i \in \mathcal{E}$ are chosen to be in the anti-clockwise direction, show that the γ defined by Eq. (6.3.8) is equal to twice the area $|\Omega_i|$ of the triangular finite element domain Ω_i.

6.4 With respect to a domain $\Omega = (0,1)^2$ such as in Fig. 6.40 (a) and boundaries $\Gamma_{\mathrm{D}} = \{\boldsymbol{x} \in \partial\Omega \mid x_1 = 0,\ x_2 = 0\}$ and $\Gamma_{\mathrm{N}} = \partial\Omega \setminus \bar{\Gamma}_{\mathrm{D}}$, obtain $u : \Omega \to \mathbb{R}$ which satisfies

$$-\Delta u = 1 \quad \text{in } \Omega, \quad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Gamma_{\mathrm{N}}, \quad u = 0 \quad \text{on } \Gamma_{\mathrm{D}}$$

via the finite element method using finite element division such as that shown in Fig. 6.40 (b).



(a) Ω and Γ_{D} (b) Finite element division

Fig. 6.40: Domain and finite element division for Exercise 6.3.2.

6.5 In a two-dimensional Poisson problem (Problem 6.1.7 with $d = 2$), let us suppose that a rectangular first-order element of Fig. 6.26 is used. In this case, obtain the elements of the coefficient matrix and the elements of the known term vector. Here, $b = b_0$ is a constant function and $p = 0$.

6.6 When a plane stress ($\sigma_{33} = \sigma_{13} = \sigma_{23} = 0$) is assumed in a two-dimensional linear elastic problem, show the calculation method of the element coefficient matrix of a 4-node isoparametric finite element in the following order.

- Let $\bar{\boldsymbol{u}}_i = (u_{11}, u_{12}, u_{13}, u_{14}, u_{21}, u_{22}, u_{23}, u_{24})^\top$ be node displacements of a finite element $i \in \mathcal{E}$. Let $\boldsymbol{E}(\boldsymbol{u}(\boldsymbol{\xi})) = (\varepsilon_{jl}(\boldsymbol{\xi}))_{jl}$ be a strain tensor and $\boldsymbol{\varepsilon}(\boldsymbol{\xi}) = (\varepsilon_{11}, \varepsilon_{22}, 2\varepsilon_{12})^\top$ be its vector expression. Here, show the calculation method of displacement–strain matrix $\boldsymbol{B}(\boldsymbol{\xi})$ which becomes

$$\boldsymbol{\varepsilon}(\boldsymbol{\xi}) = \boldsymbol{B}(\boldsymbol{\xi})\,\bar{\boldsymbol{u}}_i.$$

- Let $\boldsymbol{S}(\boldsymbol{u}(\boldsymbol{\xi})) = (\sigma_{jl}(\boldsymbol{\xi}))_{jl}$ be a stress tensor and $\boldsymbol{\sigma}(\boldsymbol{\xi}) = (\sigma_{11}, \sigma_{22}, \sigma_{12})^\top$ its vector expression. When plane stress ($\sigma_{13} = \sigma_{23} = \sigma_{33} = 0$) is assumed, the constitutive law can be given by

$$\boldsymbol{\sigma}(\boldsymbol{\xi}) = \boldsymbol{D}\boldsymbol{\varepsilon}(\boldsymbol{\xi}), \quad \boldsymbol{D} = \frac{e_{\mathrm{Y}}}{1-\nu_{\mathrm{P}}^2}\begin{pmatrix} 1 & \nu_{\mathrm{P}} & 0 \\ \nu_{\mathrm{P}} & 1 & 0 \\ 0 & 0 & (1-\nu_{\mathrm{P}})/2 \end{pmatrix}$$

using Young's modulus e_{Y} and Poisson ratio ν_{P}. Here, show the calculation method of element coefficient matrix $\bar{\boldsymbol{K}}_i$.

References

[1] Ciarlet, P. G. *Finite Element Methods.* Handbook of Numerical Analysis, P.G. Ciarlet, J.L. Lions, general editors. Elsevier, Amsterdam; Tokyo: North-Holl, 1991.

[2] Kikuchi, F. *Mathematics of Finite Element Method: Mathematical Basics and Error Analysis (in Japanese).* Baifukan, Tokyo, 1994.

[3] Kikuchi, F. *Overview of Finite Element Method, New and Revised Edition (in Japanese).* Saiensu-sha, Tokyo, 1999.

[4] Strang, G. and Fix, G. J. *An Analysis of the Finite Element Method.* Prentice-Hall, Englewood Cliffs, N.J., 1973.

[5] Tabata, M. *Numerical Solutions of Differential Equations II (in Japanese).* Iwanami Kouza Applied Mathematics. Iwanami Shoten, Tokyo, 1994.

[6] Tabata, M. *Numerical Analysis of Partial Differential Equations (in Japanese).* Iwanami Kouza Applied Mathematics. Iwanami Shoten, Tokyo, 2010.

[7] Tabata, M., Fujii, H., and Miyoshi, T. Finite element mthod using singular function (in Japanese). *bit*, 5:1035–1040, 1973.